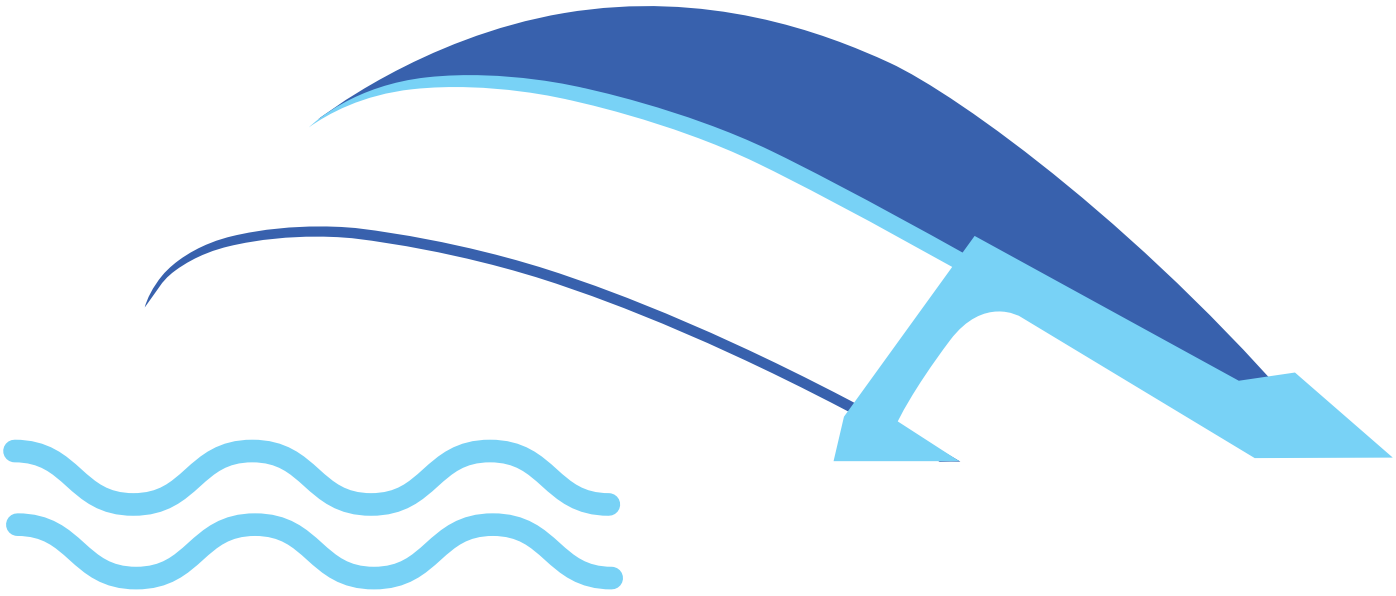
Proceedings from the 9th European Congress of Methodology



RESEARCHDESIGNBIGDATASTATISTICSRIGOUR
MEASUREMENTTRANSPARENCYREPLICATION

Encouraging Advance in Methodology
European Association of Methodology

PROCEEDINGS FROM THE 9TH EUROPEAN CONGRESS OF METHODOLOGY

PROCEEDINGS FROM THE 9TH EUROPEAN CONGRESS OF METHODOLOGY

Ana Hernández & Inés Tomás
(coords.)

UNIVERSITAT DE VALÈNCIA

DOI: http://dx.doi.org/10.7203/PUV-OA-438-5

ISBN: 978-84-9133-438-5

Digital edition

Contents

SYMPOSIA

Proceedings from the 9th European Congress of Methodology

Preface

This book collects some of the papers (posters, oral presentations and symposia) presented at the 9th European Congress of Methodology. The European Congress of Methodology is the biennial congress of the European Association of Methodology (EAM), a society founded in 2004, which brings together a large number of researchers from all over the world to exchange ideas on developing new methods and applying new methodologies in empirical research.

Due to the global Covid-19 pandemic, the 9th European Congress of Methodology, which was initially planned for July 2020, had to be postponed to July 2021. However, to ensure that our community did not miss out on the extraordinary input of the proposals that had been accepted, the organizing committee extended the option of sharing accepted proposals on the congress website. This was the so-called Virtual Phase of EAM2020-2021. A number of papers written by those who decided to participate in this virtual phase were made available from August 2020 at https://congresos.adeituv.es/EAM2020/paginas/pagina_577_1.en.html. Some of these papers, specifically seven oral presentations and four symposia, are included in this volume.

In July 2021 (21st to 23rd), due to the ongoing Covid-19 pandemic, the congress was held in hybrid format, mixing in-person and on-line presentations. Most were live presentations, although some of them were pre-recorded. A total of 13 oral presentations, three posters and three symposia presented in July 2021 are included in this book. The abstracts of the four excellent keynotes and seven state-of-the-art talks are included too.

We would like to thank all the participants for their contributions to the EAM2020-2021 program and for their contributions to this Proceedings Book. We would also like to express our heartfelt gratitude and appreciation to the EAM2020-2021 organizing committee, the sponsors and the students who worked voluntarily during the event.

Ana Hernández (Congress Chair) and Inés Tomás (Congress Deputy Chair)

ORGANIZING COMMITTEE

All members of the Organizing Committee belong to the Department of Methodology of Behavioral Sciences at the University of Valencia (Spain).

Chair

Ana Hernández (Associate Professor)

Deputy Chair

Inés Tomás (Associate Professor)

General Secretariat

Begoña Espejo (Associate Professor)
Laura Galiana (Associate Professor)
Laura Badenes (Assistant Professor)

Treasury Management

Juan C. Ruiz (Associate Professor, Head of the Department of Methodology of Behavioral Sciences
M. Castillo Fuentes (Assistant Professor)

Committee Members

M. Dolores Sancerni (Associate Professor)
Carmen Dasí (Associate Professor)
Amparo Oliver (Full Professor)
José M. Tomás (Full Professor)
María F. Rodrigo (Associate Professor)
Joan García-Perales (Lecturer)
Irene Checa (Lecturer)
David Peris- Del Campo (Lecturer)
Irene Fernández (PhD Student)
Sara Martínez (PhD Student)
María Fernández-López (PhD Student)

SCIENTIFIC COMMITTEE

Prof. Dr. Axel Mayer, RWTH Aachen University (Germany)
Associate Professor, Methodology of Behavioral Sciences

Prof. Dr. Albert Sesé, Universitat de les Illes Balears (Spain)
Full Professor, Psychology Department

Prof. Dr. Mirjam Moerbeek, Utrecht University (The Netherlands)
Associate Professor, Department of Methodology and Statistics

Prof. Dr. Ana Hernández, Universitat de València (Spain)
Associate Professor, Methodology of Behavioral Sciences

Prof. Dr. Johannes Hartig, German Institute for International Educational Research (Germany)
Full Professor, Educational Quality and Evaluation

Prof. Dr. Noémi Schuurman, Tilburg University (Netherlands)
Associate Professor, Tilburg School of Social and Behavioral Sciences

Prof. Dr. Steffi Pohl, Freie Universität Berlin (Germany)
Assistant Professor, Department of Education and Psychology

Prof. Dr. Salvador Chacón, Universidad de Sevilla (Spain)
Full Professor, Experimental Psychology Department

Prof. Dr. María-José Blanca, Universidad de Málaga (Spain)
Full Professor, Psychobiology and Methodology of Behavioral Sciences

Prof. Dr. Rolf Steyer, Friedrich-Schiller-Universität Jena (Germany)
Full Professor, Institute of Psychology

Prof. Dr. Joop Hox, Utrecht University (The Netherlands)
Full professor, Faculty of Social Sciences

Prof. Dr. José Muñiz, Nebrija University (Spain)
Full Professor, Psychology Department

Prof. Dr. Tamás Rudas, Eötvös Loránd Universityer (Germany)
Full Professor, Academic Director of TARKI and Head of the Statistics Department at the Faculty of Social Sciences

Prof. Dr. Mª Dolores Hidalgo, Universidad de Murcia (Spain)
Full Professor, Department of Basic Psychology and Methodology

Prof. Dr. Inés Tomás, Universitat de València (Spain)
Associate Professor, Methodology of Behavioral Sciences

Prof. Dr. Amparo Oliver, Universitat de València (Spain)
Full Professor, Methodology of Behavioral Sciences

Prof. Dr. José Manuel Tomás, Universitat de València (Spain)
Full Professor, Methodology of Behavioral Sciences

KEYNOTE SPEAKERS

ABTRACTS

Revisiting psychometrics at twenty-first century: Improving psychological science and its implementation

Albert Sesé

University of the Balearic Islands
Spain

Most of the psychological variables used in basic and applied research are latent and a strongly consensual theoretical definition is needed in order to obtain reliable and valid evidence of the measures of such constructs. Nothing new under the sun, but the problem is that the patterns of development and use of psychometric instruments in psychological research show important shortcomings, especially related to the definition of constructs, to the quality of their operationalization, and to the application of a magical thought that believes that the theoretical weakness of a measure will be mitigated by the statistical sophistication of the validation procedures used. These topics may be behind the lack of reproducibility in Psychological Science stated by several recent meta-analytic studies. The problem is so serious that some authorized voices have declared that Psychology is in crisis. Needless to say that psychological measurement occupies a preponderant place in this critical scenario. And the problem is not only related to improving the quality of the generated psychometric scientific evidence but also to increasing the efficiency of the implementation of these measures in real contexts. The gap between academic procedures and professional ones must be dramatically reduced in order to improve psychological intervention. Moreover, the emergence of new technologies such as virtual reality and access to massive data via smartphones can contribute to the improvement of measurement in Psychology. The objective of this keynote is to analyze and discuss some of the problems and challenges of 21st century Psychometrics, while always trying to stay true to the foundations of Psychological Measurement.

Items in the digital era: Quo vadis?

José Muñiz

University of Oviedo
Spain

Items constitute the basic units, the building bricks, with which tests are constructed. Therefore, they are essential if the test is to have the necessary psychometric quality. The emergence of new information technologies and the internet has had a major impact on psychological evaluation in general, and on the development of items in particular. The presentation reviews the changes that are currently taking place in the construction and analysis of items, and comments on some future perspectives and challenges. Firstly, the problem of item classification is analyzed, proposing a new generative taxonomy that enables not only consistent classification of existing items, but also serves as a guide to generate new item modalities. Secondly, the fundamental principles for the suitable development of items are discussed, with special emphasis on the novel types of items that arise in the digital era, and automatic item generation. Items for the assessment of non-cognitive variables are reviewed, with special mention of the ubiquitous Likert-type format and the modelling of forced-choice response formats. The increase in online evaluations poses the problem of item security, and some of the viable alternatives are analyzed. Increasing international and globalized evaluation raises the problem of intercultural equivalence of items, and the need to avoid differential item functioning, to ensure fair, equitable evaluation. Finally, some future perspectives and challenges are discussed, such as the cognitive processing of items, the validation of new digital formats, the use of ecological momentary assessment, network analysis, and the psychometric treatment of omic data.

Probabilistic causality: Beyond Rubin and Pearl

Rolf Steyer

Friedrich Schiller University Jena
Germany

I introduce the theory of causal effects in which individual, conditional, and average total causal effects are defined exclusively in terms of probability theory. This theory avoids the deterministic assumption that the outcome is a fixed number given person and treatment, which is the starting point in Donald Rubin's definition of individual and average effects. Instead it is only assumed that the conditional distribution of the outcome variable is fixed given person and treatment. It also avoids Judea Pearl's misleading do-operator, showing that causal effects are defined without referring to an intervention, that is, without the necessity of "doing" something, and also without knowing the causal relations between all the variables involved. The theory also provides several causality conditions under which conditional expectations such as E(Y|X) and E(Y|X, Z) describe (conditional) causal dependencies of the outcome variable Y from the treatment variable X given the (possibly multivariate) covariate Z. Focusing on E(Y|X, Z), important causality conditions are unbiasedness and strong ignorability, which are of theoretical interest, but empirically untestable. Other causality conditions are conditional independence of X and all potential confounders given Z, mean-independence of Y from all potential confounders given X and Z, and unconfoundedness of E(Y|X, Z). These causality conditions are empirically testable and imply strong ignorability and unbiasedness. Hence, they can be used in selecting the (possibly multivariate) covariate Z, for which the corresponding causality condition for E(Y|X, Z) is assumed to hold. Surprisingly, not only Z-conditional independence of X and potential confounders imply strong ignorability, so does mean-independence of Y from potential confounders given X and Z. Because strong ignorability suffices for a valid propensity score analysis, this condition allows supplementing the strategy to condition on the covariates that determine the treatment probability by the alternative strategy to condition on the covariates determining the conditional expectation of the outcome variable Y.

Do methodological tutorials have an impact?

Lisa L. Harlow

University of Rhode Island
USA

Science is undergoing close scrutiny with renewed interest in encouraging more open and reproducible practices. One of the main goals of behavioral science is to understand and explain human behavior in coherent and credible ways. Within this focus, there is a need to learn about and apply rigorous quantitative methods that go beyond an over-emphasis on dichotomous decision-making. Whereas statistical journals provide a forum for presenting and applying innovative methods, they may be directed at a narrow, methodological readership and miss reaching a wider applied audience that could help in moving science forward. Methodological tutorials are explored to see if they can offer a channel that researchers can turn to in order to help in producing and illuminating reliable and worthwhile findings. The journal Psychological Methods serves as a case study for assessing the impact of manuscripts published in response to a General Call for Tutorials that was initiated in September 2014. The number and nature of citations for these kinds of instructive articles are compared to those for other manuscripts published in this journal during a similar time frame. Discussion will include identifying factors that appear to contribute to an effective and accessible article, and other journals will also be briefly studied to consider whether and to what extent methodological tutorials or similar teachers' corner papers have an impact.

STATE-OF-THE-ART TALKS

ABSTRACTS

Faking behavior in high-stakes assessments: Can it be modelled? Can it be prevented?

Anna Brown

University of Kent
United Kingdom

Asking people to self-report is by far the most popular method for measuring personality, attitudes and other traits where objective measurement is difficult. However, self-reported data are commonly affected by conscious and unconscious distortions. Examples include individual styles in using rating options, an unconscious tendency to present oneself in a positive light, or conscious manipulation of responses to manage impressions in high-stakes assessments. The extent to which respondents engage in such behaviors varies, and if not controlled, these distortions are threat to validity.

This talk focuses on impression management (aka faking) as the most challenging distortion to understand and model. I will provide a brief overview of the evolution of views on the problem, and present key approaches to statistical control of faking, and their respective shortcomings. I will then present my recent proposal to model faking as a Grade of Membership (F-GoM) in two archetypical profiles – 'real' and 'ideal', whereby an individual's profile is a mixture of responses - some are reported as retrieved ('real') and some are edited before reporting to present an 'ideal' image. This approach is based on the Retrieve-Edit-Select by Böckenholt (2014) and has some real strength in understanding individual differences in faking behavior.

Alternatives to statistical control include prevention, and there have been advances in this area too. Forced-choice response formats have been used as a bias prevention method, and with the advent of appropriate measurement models for ipsative data (e.g. Brown & Maydeu-Olivares, 2011; Stark, Chernyshenko & Drasgow, 2005) we can attempt to evaluate the effects of faking on this format too. I will conclude with an outlook for research in this area.

Disentangling aspects of test performance and consequences for reporting

Steffi Pohl

Freie Universität Berlin
Germany

Results of assessments, such as PISA and PIAAC, are not only an indicator of competencies, but are also impacted by their effort and their test-taking behavior. Examinees differ with respect to the pace at which they choose to work on the items, nonresponse behavior, and guessing. The currently reported levels of competencies do not reflect competence levels alone but are instead a mix of the test-taking behavior and competence level of the examinees. This results in unfair comparisons across persons and countries that differ in test-taking behavior. In order to understand the performance of the examinees and to take and evaluate appropriate policy measures, we suggest disentangling and separately reporting the different aspects that drive performance.

I make use of process data from computer-based assessment and use them to gain information on the examinees' test-taking behavior. In my work I bring together research on missing values and guessing with approaches for modeling timing data. I propose different models for different test-taking behavior. The proposed models enable a) modeling different kinds of test-taking behavior and b) a deeper understanding of examinees' performance. As test takers use different test-taking strategies, which are reflected in different timing data, response patterns and occurrence of missing values, instead of reporting just one competence score, I suggest reporting a profile of different aspects that describe the performance of the test takers. I will discuss the implications of the proposed approach as compared to current practice for reporting on competence levels in large-scale assessments.

Small sample solutions for SEM

Yves Rosseel

Ghent University
Belgium

Structural equation modeling (SEM) is a widely used statistical technique for studying relationships among multivariate data. Unfortunately, when the sample size is small, several problems may arise: non-convergence, bias, and non-admissible solutions (e.g., negative variances). A popular solution often suggested in the literature is to switch to a Bayesian approach. However, in this presentation, I follow the frequentist framework and present two solutions that may fix many of the current problems. A first solution is merely a computational trick. Instead of using unconstrained optimization (using, for example, quasi-Newton methods), one could impose simple lower and upper bounds on a selection of model parameters during optimization. By using well chosen bounds that are just outside the admissible parameter space, we can stabilize regular ML estimation in (very) small samples.

A second solution is the so-called structural-after-measurement (SAM) approach. In this approach, estimation proceeds in several steps. In a first step, only parameters related to the measurement part of the model are estimated. In a second step, parameters (only) related to the structural part are estimated. Several implementations of this old idea will be presented. A distinction will be made between local and global SAM, and it will be suggested that various alternative estimators (including non-iterative estimators) could be used for the different model parts. It turns out that this approach is not only effective in small samples, but it is also robust in many types of model misspecification. Many existing alternatives (factor score regression with Croon corrections, sum scores with fixed reliabilities, model-implied instrumental variables estimation, Fuller's method...) turn out to be special cases of this general framework. Finally, I will briefly demonstrate how these solutions can be used in the R package lavaan.

Bayesian dynamic borrowing for single and multilevel models

David Kaplan

University of Wisconsin – Madison
USA

The central feature of Bayesian statistics that distinguishes it from conventional frequentist statistics is the ability to formally incorporate prior information into statistical analyses. Prior information is specified in terms of prior probability distributions which encode the information that researchers might have regarding what is reasonable to believe about the parameters of their model. However, the elicitation of substantive prior information is difficult. Typically, researchers utilizing Bayesian methods rely on software default settings that presume non-informative prior information. Nevertheless, in education research, long-standing large-scale educational assessments such as the Programme for International Student Assessment (PISA) could be used to develop informative prior information to be incorporated into Bayesian modeling for policy-relevant research. The purpose of this talk is to share recent work on a novel extension of *Bayesian dynamic borrowing* (a method originally developed for case-control studies) to single and multilevel regression models with applications for large-scale educational assessments. An attractive feature of Bayesian dynamic borrowing is that the method allows a researcher to account for the fact that not all historical data, even from the same survey program, are exchangeable. As such, prior information can be systematically adjusted to reflect the analyst's degree-of-confidence in the exchangeability of sources of prior data and current data. We present a detailed simulation study and case study, comparing our extension of Bayesian dynamic borrowing to conventional pooling and to power priors. Our results demonstrate the advantages of Bayesian dynamic borrowing, particularly in cases where data sets are relatively heterogenous. We also present a Shiny App that will provide researchers with a tool to incorporate Bayesian dynamic borrowing and power priors in single and multilevel settings.

Meta-analysis: Its role in the reproducibility of psychological research

julio Sánchez-Meca

University of Murcia
Spain

Psychological research (and other related disciplines) is suffering a confidence crisis due to the difficulties in replicating and reproducing original psychological findings. The excessive 'researcher degrees of freedom' have led to a wide range of questionable research practices (e.g., p-hacking, HARKing, reporting bias, publication bias), whose main consequence is the reporting of biased findings.

In recent years, several international initiatives have begun to try to replicate original results by involving independent research teams in collaborative large-scale replication studies (Open Science Foundation, Many Labs Project, Registered Replication Reports). However, there is no consensus on which criteria should be applied to determine whether a set of replication studies actually replicates the original finding or not. In this talk, the advantages and limitations of different criteria applied in these large-scale replication studies are discussed, as well as the critical role that meta-analytic thinking has in these studies promoted by the Open Science framework.

Situational judgment tests work, but how?

Marise Born

Erasmus University Rotterdam
The Netherlands

Situational Judgment Tests (SJTs) are a method in which respondents are asked to react to work- or study-related situation descriptions. The SJT method is more than 100 years old and has most often been used for personnel and academic selection. SJTs have numerous appearances, varying from written descriptions to virtual reality situations, with response options in the form of a rating format or open- ended responses. They have been developed for many mostly non-cognitive constructs, such as leadership and integrity, but also for behavioristic prediction of future job or study performance without considering construct validity. One feature they have in common is that applicants like them, most probably because of being absorbed in realistic but imaginary situations. SJTs are also known for having other positive features, such as a good predictive validity and less susceptibility to faking and bias. This state-of-the-art presentation will focus on disentangling the "how" behind the working of SJTs by discussing the effects of SJTs' building blocks: what do we know about the effects of situations, response formats, instruction types, and scoring methods of SJTs on this method's effectiveness? To this end, I will discuss a series of studies conducted with my colleagues, which among other things have focused on the so-called implicit trait policy (ITP), recognizing how *not* to respond, instructing to judge what *others* would do, and faking. By combining our findings with research published by others on the workings of SJTs, and by comparing SJTs with an equivalent method, namely the Assessment Center (AC) method, I will draw several conclusions about the mechanisms of the SJT method.

Alternative approaches to longitudinal panel data analysis

Albert Satorra

Universitat Pompeu Fabra
Spain

Proper modeling of longitudinal data enables controlling for unobserved confounders, just as multiple indicators (the factor model) help to assess the relationship between latent (unobservable) variables. Different disciplines, however, have developed alternative approaches to longitudinal panel data. We see a big contrast between the practice in econometrics, dominated by mixed- effects regression, and psychometrics and behavioral science methods based on simultaneous sets of equations, SEM models. We will review and compare the different approaches to longitudinal panel data assessing their comparative advantages and the relative robustness to standard assumptions (e.g., the robustness of the full information ML approach to non-normality). Methods for missing data will also be assessed in the comparison. This review's urgency arises when we see that standard widely used software (e.g., Stata) allows the analysis of the same longitudinal model using the same software platform but with an alternative model and computational types of machinery. We will discuss examples of applications by way of illustration.

ORAL PRESENTATIONS & POSTERS

Comparison of the effects of country-level income inequality and individual perceived income inequality on self-rated health: a multilevel analysis

Toktam Paykani[1], Hassan Rafiey[2], Homeira Sajjadi[3]

[1]*Social Development and Health Promotion Research Center, Gonabad University of Medical Sciences, Gonabad, Iran*

[2]*Department of Social Welfare, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.*

[3] *Social Welfare Management Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.*

Abstract

Although the relationship between individuals' health and income inequality has been tested by many researchers, it is still a controversial issue. We advanced this research area by simultaneously assessing the effects of country-level income inequality and individual perceived income inequality on self-rated health (SRH). A series of mixed-effects models was conducted. All data at individual level were obtained from the International Social Survey Programme (ISSP), module on Social Networks (Version 2.0.0, 2019-8-19) which covered 44,492 individuals nested in 30 countries. There was a social gradient in SRH. We did not find a negative relationship between country-level inequality and SRH. The study results also did not support the threshold effects hypothesis which posits a threshold of income inequality over which the adverse effects on health increase. In contrast, people who perceived the society as being more unequal experienced poorer health. However, the effect size was small. This study showed that SRH was less affected by the distribution of income in that society. Instead, higher perceived inequality and lower socioeconomic position increased the odds for people in reporting poor health.

Keywords: Income inequality; multilevel modeling; perceived inequality; self-rated health; social determinants of health; threshold effects.

E-mail: paykani.t@gmu.ac.ir

1. Introduction

The idea that individual income correlates with individual health is widely accepted. Nevertheless, the relationship between income inequality and health is still a non-conclusive relation of support (Jen et al., 2009). In order to explore findings for and against the income inequality-health association, Wilkinson and Pickett (2006) reviewed studies on income inequality and population health. They concluded that studies in larger areas, where income inequality is an indicator of social stratification or hierarchy, were more supportive of the aforementioned relationship (Kondo et al., 2012; Wilkinson & Pickett, 2006). A number of researchers have suggested that the adverse health impact of income inequality may be evident beyond a certain threshold of inequality. It has been demonstrated that the income inequality-health association may be stronger in societies with higher levels of average income inequality (Kondo et al., 2009; Kondo et al., 2012; Subramanian & Kawachi, 2004). It has also been suggested that income inequality exerted negative effects on outcomes with an inverse social gradient (Wilkinson & Pickett, 2008).

In this study we advanced this research area by simultaneously assessing the effects of country-level income inequality and individual perceived income inequality on self-rated health. Previous studies have reported that one's perception of income inequality, resulting from a subjective comparison between one's income with that of others, decreases well-being (Cheung & Lucas, 2016; Kondo, Kawachi, Subramanian, Takeda, & Yamagata, 2008; Wilkinson & Pickett, 2019). The main goals of our study were hence;

1) To assess the effects of individuals' socioeconomic position on self-rated health after taking account of demographic factors. This examined whether or not there is a socioeconomic gradient in SRH.
2) To estimate the effect of individual perceived income inequality and country-level income inequality on SRH, after controlling for the effects of potential confounders, including individual socioeconomic status and country wealth. This study differs from previous research as it assessed the contribution of perceived inequality to explain differential health status.
3) To test whether inequality is harmful for SRH beyond a certain threshold.

2. Methods

2.1. Data

All data at individual level were obtained from the International Social Survey Programme (ISSP) Social Networks survey (2017) (Version 2.0.0, 2019-8-19 final release). After excluding the respondents with missing values in one of our study variables, the sample size in our study included 40,163 adults aged 19 years or older.

2.2. Dependent variable

Our target research outcome was SRH (Idler & Benyamini, 1997; Jylhä, 2009). The state of health in ISSP-Social Network survey consists of a single item (In general, would you say your health is… Excellent, Very good, Good, Fair, or Poor) (ISSP, 2017). The original four-point

response scale was dichotomized: “excellent”, “good” and “very good” responses were recoded as zero, and “poor” and “fair” responses were recoded as one.

2.3. Independent variables

2.3.1. Socioeconomic status

Education level and household income were used as indicators of socioeconomic status in this study (ISSP, 2017).

2.3.2. Perceived income inequality

Respondents were asked about their views on income inequality in their society: “Differences in income in [COUNTRY] are too large”. The answers ranged from “Strongly agree” to “Strongly disagree” (ISSP, 2017).

2.3.3. Country wealth

The World Bank provides the country’s nominal income data on GDP per capita PPP (constant 2011 international $) on an annual basis for 232 countries. GDP per capita is an indicator for economic development, macroeconomic performance, and country wealth.

2.3.4. Country income inequality

Finding a valid measure for income inequality was a challenge (Jen et al., 2009). The Gini coefficient is the most common indicator of income inequality in multilevel studies on health disparities (Kondo et al., 2009; Subramanian & Kawachi, 2004). Data on the Gini coefficient were obtained from the World Bank indicators. For each country, the score on the Gini index was obtained from the country survey year.

2.4. Statistical analysis

Because of the multilevel nature of the income inequality-health hypothesis (Subramanian & Kawachi, 2004), we considered a two-level model for binary responses. Data analysis was carried out using glmer (lme4 library, Bates et al., 2015). Overall, four models were built. Model 0, called the null model, was an intercept-only model to estimate whether there was a significant variation between countries in self-rated health. Model 1 included all individual level variables in the fixed part of the model 0. This model contained socioeconomic status indicators and assessed whether there was a social gradient in SRH while controlling for individual demographic characteristics. In addition, Model 1 incorporated the effect of individual perceived income inequality to estimate how much perceived income inequality affected individual SRH. Model 2 was built on Model 1 by adding two contextual variables, including national wealth and country income inequality (Gini index), to assess to what extent country-level variables explained SRH. In Model 3, an interaction term between Gini and a dummy term for the Gini threshold allowed us to test income inequality-health hypothesis for all countries, as well as countries separately in term of countries above or below the inequality threshold (>=30) (Kondo et al., 2012).

3. Results

The descriptive statistics for all study variables are summarized in Table 1.

Table 1. Sample characteristics

Outcome measure	N (%), Mean (SD)
Good self-rated health (ref) Poor and fair self-rated health	29,730 (74%) 10,433 (26%)
Respondent-level covariates (n=40,163)	
Age-centered (−30.2–53.8)	0 (16.9)
Female	21,350 (53.2%)
Male (ref)	18,813 (47,8%)
Married/Couple	23,037 (57.4%)
Single/Divorced/Separated/Widowed (ref)	17,126 (42.6%)
Education Lower Middle (ref) Upper	 15,098 (37.6%) 14,302 (35,6%) 10,763 (26.8%)
Household income 1 (Lowest level) 2 3 (ref) 4 5 (Highest level) Don't know	 4,615 (11.5%) 9,002 (22.4%) 12,679 (31.6%) 9.255 (23%) 3,686 (9.2%) 936 (2.3%)
Perceived income inequality High Low (ref)	 32,449 (80.8%) 7,714 (19.2%)
Country-level covariates (n=30)	
Ln GDP per capita, PPP, constant 2011 international $)	9.88 (1.48)
Gini (Range: 25.4–63)	36.97 (8.9)

Table 2 summarizes the results of multilevel analysis as odds ratios and 95% confidence intervals.

Table 2. Fixed effect multilevel logistic models of poor SRH

	Null model	**Model 1**	**Model 2**	**Model 3**
Constant	.3 (.23.37)***	.24 (.18–.32)***	.137 (.02–.92)*	.012 (.00–1.346)
Individual predictors				
Age (centered around 49)		1.04 (1.03–1.04)***	1.04 (1.03–1.04)***	1.04 (1.038–1.042) ***

	Null model	Model 1	Model 2	Model 3
Female		1.11 (1.06–1.17)***	1.11 (1.06–.1.17)***	1.116 (1.061–1.175) ***
Married		0.78 (.74–.82)***	.78 (.74–.82)***	.778 (.738–.821)***
Education				
Lower		1.28 (1.2–1.37)***	1.28 (1.2–1.37)**	1.288 (1.208–1.372)***
Higher		.7 (.65–.76)***	.7 (.65–.76)**	.708 (.658–.762)***
Household income				
1 (Lowest)		2.29 (2.1–2.49)***	2.29 (2.1–2.49)***	2.294 (2.107–2.497)***
2		1.6 (1.49–1.71)***	1.6 (1.49–1.71)***	1.602 (1.498–1.712) ***
4		.67 (.62–.72)***	.67 (.62–.72)***	.67 (.622–.723)***
5 (Highest)		.51 (.45–.58)***	.51 (.45–.58)***	.515 (.457–.581)***
Don't Know		1.08 (.29–3.9)	1.09 (.29–4.07)	1.49(.35–6.5)
High perceived income inequality		1.078 (1.006–1.15)*	1.07 (1.006–1.15)*	0.927 (0.864 – 0.993)
Country predictors				
Ln GDP per capita, PPP			1.05 (.87 – 1.28)	1.068 (.886 – 1.288)
Gini (centered around 37)			1.006 (.97–1.03)	0.79 (.523–1.193)
Gini x inequality threshold				1.27 (0.842–1.925)
Random part				
Between-country	.434 (.26–.74)	.419 (.25–.7)	.414 (.25–.69)	.396 (.237–.662)
Log likelihood	−21306.075	−18955.18	−18954.98	−18954.32
AIC	42616	37936	37939	37942
N (Countries)	30	30	30	30
N (Observation)	40,163	40,163	40,163	40,163

* P<0.05 **p <0.01 *** p < 0.001

4. Conclusions

As seen, country-level income inequality was not significantly associated with poor SRH.

Researchers have suggested a threshold effect of income inequality on population health as a potential factor that could explain the heterogeneity between studies (Kondo et al., 2009;

Kondo et al., 2012; Subramanian & Kawachi, 2004). However, our results did not support the implications of the inequality threshold effect hypothesis.

Another finding of this study was related to socioeconomic disparities in self-rated health. Although Wilkinson argued that income inequality exerted negative effects on outcomes with an inverse social gradient (Wilkinson & Pickett, 2008; Wilkinson & Pickett, 2010), our findings showed a social gradient in SRH. Yet, we did not find a negative relationship between country-level income inequality and population health. This is consistent with previous studies (Gerdtham & Johannesson, 2004; Jen et al., 2009; Kondo et al., 2008).

A possible explanation for these results is provided by critics who argued that differences in SRH between countries are not predictors of objective health as measured by death rates which is the most reliable measure of population health. The core of the problem is that although the measure of SRH is a good predictor of mortality and morbidity within countries (Idler & Benyamini, 1997; Jylhä, 2009), it breaks down entirely when making comparisons between countries (Barford, Dorling, & Pickett, 2010).

There was a weak negative association between perceived income inequality and SRH. People who perceived society as being more unequal, experienced poorer health. The effect size was small. However, even a slightly adverse effect of inequality on health, especially in later life, can result in a financial burden for the population (Vauclair et al., 2015). This finding might be consistent with a psychosocial explanation of the effect of income inequality on health raised by Picket and Wilkinson. These two researchers argued that increasing the level of stress in individuals' experience in their domestic life caused by undesirable socioeconomic conditions can lead to lasting psychological and emotional damage (Wilkinson & Pickett, 2010, 2019).

A limitation in this study was that we measured countries' income inequality closest to the time when our outcome, i.e. self-rated health, was measured. However, income inequality might not have an instantaneous effect on perceived health (Kondo et al., 2012). Since the ISSP data used here had a cross-sectional rather than a longitudinal design, we could not examine the potential lag effects of income inequality on health.

References

Barford, A., Dorling, D., & Pickett, K. (2010). *Re-evaluating self-evaluation. A commentary on Jen, Jones, and Johnston* (68: 4, 2009). Social science & medicine, 70(4), 496–497.

Cheung, F., & Lucas, R. E. (2016). *Income inequality is associated with stronger social comparison effects: The effect of relative income on life satisfaction*. Journal of personality and social psychology, 110(2), 332.

Gerdtham, U.-G., & Johannesson, M. (2004). *Absolute income, relative income, income inequality, and mortality*. Journal of Human Resources, 39(1), 228–247.

Idler, E. L., & Benyamini, Y. (1997). *Self-rated health and mortality: a review of twenty-seven community studies*. Journal of health and social behavior, 21–37.

ISSP. (2017). *Social Networks and Social Resources*. http://www.issp.org.

Jen, M. H., Jones, K., & Johnston, R. (2009). *Global variations in health: evaluating Wilkinson's income inequality hypothesis using the World Values Survey*. Social science & medicine, 68(4), 643–653.

Jylhä, M. (2009). *What is self-rated health and why does it predict mortality? Towards a unified conceptual model*. Social science & medicine, 69(3), 307–316.

Kondo, N., Kawachi, I., Subramanian, S., Takeda, Y., & Yamagata, Z. (2008). *Do social comparisons explain the association between income inequality and health?: Relative deprivation and perceived health among male and female Japanese individuals*. Social science & medicine, 67(6), 982–987.

Kondo, N., Sembajwe, G., Kawachi, I., van Dam, R. M., Subramanian, S., & Yamagata, Z. (2009). *Income inequality, mortality, and self rated health: meta-analysis of multilevel studies*. Bmj, 339, b4471.

Kondo, N., van Dam, R. M., Sembajwe, G., Subramanian, S., Kawachi, I., & Yamagata, Z. (2012). *Income inequality and health: the role of population size, inequality threshold, period effects and lag effects*. J Epidemiol Community Health, 66(6), e11–e11.

Subramanian, S. V., & Kawachi, I. (2004). *Income inequality and health: what have we learned so far?* Epidemiologic reviews, 26(1), 78–91.

Vauclair, C.-M., Marques, S., Lima, M. L., Abrams, D., Swift, H., & Bratt, C. (2015). *Perceived age discrimination as a mediator of the association between income inequality and older people's self-rated health in the European region*. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 70(6), 901–912.

Wilkinson, R., & Pickett, K. (2006). *Income inequality and population health: a review and explanation of the evidence*. Social science & medicine, 62(7), 1768–1784.

Wilkinson, R., & Pickett, K. (2008). *Income inequality and socioeconomic gradients in mortality*. American Journal of Public Health, 98(4), 699–704.

Wilkinson, R., & Pickett, K. (2010). *The spirit level: Why equality is better for everyone*. Penguin UK.

Wilkinson, R., & Pickett, K. (2019). *The inner level: How more equal societies reduce stress, restore sanity and improve everyone's well-being*: Penguin Press.

The use of None-of-the-above in statistics concept measurement

Susana Sanz[1], Carmen García[1], Ricardo Olmos[1]

[1] *Department of Social Psychology and Methodology, Faculty of Psychology, Autonomous University of Madrid, Spain*

Abstract

Purpose: Nowadays, a large amount of educational assessment is made using multiple choice items, as they evaluate large groups of students quickly and accurately (Haladyna et al., 2019). However, writing good items involves considerable effort and, sometimes, lecturers use general options which are easy to create, such as None-of-the-above (NOTA). These options may be inadvisable, because they can lead to poorer psychometric properties, though the research done in this sense is not conclusive. Therefore, the aim of this study was to develop a Spanish statistic concepts inventory (SCI) to test the effects of NOTA in university assessment. *Method/Design*: we used the SCI (Stone et al., 2004) to assess statistical concepts. Then, we adapted it to Psychology students, resulting in a 30-item test. We created three forms: one with three specific options, and two where the NOTA option was also included, balancing its use as the correct option and as one of the distractors. We applied it to a sample of 449 Psychology students and performed tests to check whether our groups were equivalent. *Results*: the invariance analysis carried out with the anchor test seemed to demonstrate that the groups were equivalent. *Conclusions*: as a result of ascertaining that the groups are equivalent, we can test whether the use of NOTA involves differences in performance.

Keywords: None-of-the-above, multiple choice, educational assessment, psychometric properties, test development, item writing guidelines.

E-mails: susana.sanzv@uam.es; carmen.garcia@uam.es; ricardo.olmos@uv.es

1. Introduction

Measuring students' knowledge in a particular subject is not easy but it is very important in academia. However, it involves time and effort when lecturers want to ensure this is done correctly. For this reason, multiple choice items (MC) are popular nowadays in measuring knowledge, as they enable the evaluation of a large number of students in an accurate, quick way (Haladyna, et al., 2019). However, creating items with good psychometric quality is not easy as these must be clear and must avoid inducing errors because of the way they are written to ensure that the ultimate goal is solely assessing students' knowledge.

The main trailblazers of standardizing item writing recommendations were Haladyna and Downing (1989a; 1989b), who compiled the set of recommendations proposed in manuals to create items and selected those which were most popular among authors, finally collecting 22 recommendations. Ten out of these 22 initial recommendations were specifically focused on writing alternatives to the correct answer, which are called *distractors* (Haladyna & Rodríguez, 2013). The importance of writing distractors lies in their complexity, as it requires a lot of time if the items are to be well written. If these distractors are not plausible, the students being evaluated will easily discard them even if they do not know the correct option. Thus, the proportion of correct answers in each item will increase and the items will not be able to distinguish properly between people who have comprehensive knowledge of the subject contents and those who do not. Writing these items involves considerable effort, which is why teachers sometimes prefer alternatives which are quicker to write, such as *All-of-the-above* (AOTA) and *None-of-the-above* (NOTA), since they are very easy to generate (Frey et al., 2005). However, the use of these options is discouraged, despite the fact that, as we have mentioned before, research on the relevance of writing recommendations is limited and the empirical evidence collected offers inconclusive results (e.g. Tarrant et al., 2009).

Therefore, we consider that looking for cost reduction strategies in item writing that do not detract from their quality is a relevant research area. In particular, we are interested in the case of NOTA options, since the results obtained up until now have been contradictory. When making their first review, Haladyna and Downing (1989b) stated that items with NOTA were more difficult and less discriminatory. Test scores were also less reliable, and criterion validity was negatively affected. They concluded that they saw no advantages in this alternative when considering their recommendations. However, years later, they contradicted themselves and changed this recommendation, pointing out that NOTA can be used with caution, since it increases the difficulty, but does not affect discrimination. Even so, they stated that the ideal scenario continues to be writing distractors that provide information to the item, that is, they are clearly focused on the particular question (Haladyna & Rodriguez, 2013). Since then, some authors have tried to put together more empirical evidence. The general results seem to be that items with NOTA as the correct answer are more difficult than items with NOTA as a distractor, or without NOTA, and that discrimination is not affected (e.g. Boland et al., 2010; DiBattista et al., 2014; Martínez et al., 2009; Pachai et al., 2015). However, contradictory results continue to be found. Accordingly, our aim was to find out how NOTA actually works, because if it is acceptable in some circumstances, it can save teachers time and effort in creating some elements which they can invest in other items, resulting in more valid test results and fairer assessment.

Therefore, the aim of this study was to create an instrument to capture the differences between the use of NOTA as a correct answer or a distractor, compared to when it was not used. We decided to use a concept inventory, which has traditionally been used to address misconceptions in some fields, especially in physics and chemistry. The first inventory concept was developed

by Halloun and Hestenes (1985) to help students understand Newtonian physics concepts, and they demonstrated that it was very useful to change the previous beliefs that students had about difficult concepts. For this reason, and knowing that statistic misconceptions are usual and hard to change, we thought that this kind of instrument could be suitable for our purpose.

2. Method

2.1. Participants

The participants were 449 Psychology student volunteers, who were rewarded with extra credits for their participation. The task had three forms: Form A was answered by 146 participants, Form B by 156 participants, and Form C by 147 participants. We carried out some control checks and retained 435 students as a result (140 for Form A, 152 for Form B and 147 for Form C). The mean age of the participants was 20.35 years old. Sixty-four participants were male, and 352 were female (17 did not answer).

2.2. Materials and tasks

To create our test, we used the Statistic Concepts Inventory (SCI; Stone et al.,2004) which was originally developed to assess Engineering students' understanding of statistical concepts. The inventory included items referred to probability, descriptive statistics, inferential statistics and graphical displays. We adapted it to our Psychology students, resulting in a 30-item test to measure descriptive and inferential statistics. Ten items had three specific options, which were the same for all participants. This was used as an anchor test. Then, we balanced the other 20 items to finally obtain three versions of each one: one with three specific options, one with NOTA as the correct answer and one with NOTA as one of the distractors. With these items, we created three forms for our test: one with three specific options for each item, and two with the 10 anchor items and 20 with a NOTA option, which was used as the correct option or as one of the distractors on these two different forms.

We also collected some personal (age, gender…) and academic information, such as the grades in each statistics subject in the degree.

2.3. Design and procedure

The variables that we considered most relevant to prove that the tool was useful and that the groups were comparable were difficulty, the proportion of choosing the correct answer, the item discrimination as the item-rest correlation, and reliability. At this stage, we only took the 10 anchor items into account. In order to fully check whether the three different groups had the same efficiency, we performed an invariance analysis with the anchor test scores.

The analyses were conducted with version 3.6.3 of R software (2020) and version 8 of Mplus (2017).

3. Results

3.1. Classical Test Theory

First, we wanted to check whether there were any differences between the groups regarding the 10 anchor items, which had three specific response options for each one. This helped us to see whether the results were comparable between groups. Accordingly, the first step was to

check the score reliability in the three anchor tests. Table 1 shows the Cronbach's α for the three forms, and the confidence interval for them.

Table 1. Cronbach's α for the three forms and confidence intervals (95%)

Form	Cronbach's α (95% CI)
A	.387 (.224–.528)
B	.509 (.384–.618)
C	.421 (.269–.553)

When we compared the three alphas (Feldt et al, 1987), we did not find significant differences, so we retained the null hypothesis of equality ($\chi^2_{2;\ .05}$=1.602, p=0.449).

We continued to verify whether there were any differences in difficulty between forms in the anchor test. Table 2 shows the different values in difficulty in the three forms.

Table 2. Difficulty in anchor items in the three forms

	Item 1	Item 5	Item 8	Item 13	Item 16	Item 19	Item 24	Item 26	Item 29	Item 33
Form A	.743	.357	.193	.393	.450	.671	.107	.650	.686	.243
Form B	.724	.303	.204	.414	.434	.632	.145	.664	.684	.401
Form C	.797	.315	.238	.427	.385	.769	.140	.573	.678	.259

We performed a between-subject ANOVA to contrast the differences between groups. At first, we found differences in item 19 (forms B and C; α = .05; $F_{1.46,\ 91.64}$ = 3.452; p = 0.033) and item 33 (forms A and B, and B and C; α = .05; $F_{2.26,\ 89.69}$ = 5.432; p = 0.005). However, when we applied Bonferroni multiple-comparison correction these differences disappeared.

Third, we studied whether there were differences in discrimination between forms, by the item-rest correlation. Table 3 shows these results.

Table 3. Item-rest correlations in anchor items in the three forms

	Item 1	Item 5	Item 8	Item 13	Item 16	Item 19	Item 24	Item 26	Item 29	Item 33
Form A	.119	−.033	.107	.022	.271	.228	.211	.220	.269	.106
Form B	.119	.189	.127	.138	.324	.145	.262	.298	.366	.189
Form C	.195	−.017	.130	-.036	.196	.219	.144	.359	.305	.194

We checked the differences for the item-rest correlations (Fisher, 1925; Zou, 2007), and we did not find any differences between the forms.

3.2. Invariance analysis

To obtain more information about the groups' equivalence, we finally performed a measurement invariance analysis. First, we checked whether the unidimensional model fitted the data well in each of our three groups, together and separately. Table 4 shows the results of this analysis.

Table 4. Unidimensional model fit by forms and together

Form	χ^2	gl	*p*	RMSEA	CFI	TLI
A	36.265	35	0.410	.016	.964	.954
B	34.299	35	0.502	>.001	1.000	1.016
C	33.054	35	0.562	>.001	1.000	1.063
Together	35.769	35	0.432	.007	.995	.993

As we can see, fit statistics for the one-factor solution were consistent with good model fit. In all the groups, no remarkable points of strain were noted in either solution, as reflected by the small modification indexes. We continued to check the different types of invariance. Table 5 shows these results.

Table 5. Invariance for the 10 anchor items

	χ^2	gl	$\Delta\chi^2$	Δgl	*p*	RMSEA	CFI	TLI
Configural invariance	103.578	105	–	–	–	>.001	1.000	1.000
Metric invariance	117.717	123	15.128	18	0.653	>.001	1.000	1.000
Scalar invariance	145.677	141	30.554	18	0.032	.015	.964	.966
Test the equality of factor variances	144.011	143	0.730	2	0.694	.007	.992	.993
Test the equality of latent means	143.784	145	1.045	2	0.593	>.001	1.000	1.000

As we can see, the only problematic source of invariance was scalar invariance, as there were differences in item 33, which we have already seen in the CTT analyses. Taking into account that our sample size was controversial in terms of analysis stability, because it was around the recommended limit (Brown, 2006), we decided to apply a MIMIC model (Jöreskog & Goldberger, 1975; Muthén, 1989) in which the latent factor and the items were regressed onto the form (i.e., the group was taken as a covariate). Table 6 shows the results.

Table 6. MIMIC results

	χ^2	gl	*p*	RMSEA	CFI	TLI
MIMIC model	45.279	44	0.418	.008	.991	.989

As we can see, the MIMIC model provided a good fit to the data; the inclusion of the form covariate did not alter the factor structure or produce salient areas of strain in the solution (e.g., all modification indices <10.0). Consistent with the previous results, the results of the MIMIC model showed that the indicators were invariant for the three forms. There was no evidence of differential item functioning; that is, the item behaved equally as an indicator in the three forms.

4. Conclusions

The main goal of this work was to demonstrate that the three groups were equivalent through the anchor items of our concept inventory, so that we could conclude whether our test was

reliable and the results about NOTA were also going to be reliable. We decided to use a concept inventory because we were concerned with statistics misconceptions. Including NOTA as an option also meant that these concepts had to be clear. In addition, the version without NOTA may be useful in the future to help students change these misconceptions, as this is the first time that this has been applied in statistic concepts aimed at Psychology students in Spain.

The results that we found for the anchor test show that the score reliability was the same in the three forms. It seems low, but the anchor test had 10 items, so we considered that the complete test would be reliable. In terms of difficulty, we consider that, even when the items were more difficult than expected, they had a good range, and there were no differences between groups. The discrimination in general was low, but we consider it to be good taking into account that the items measured diverse contents, as we wanted the anchor test to sample all the relevant contents, so we decided to retain the 10 items.

Furthermore, taking into account the invariance and MIMIC model analyses and the results that we have obtained, we conclude that measurement invariance is obtained. The only problematic item was item 33, but as there are multiple comparisons, when the type I error rate was corrected (Bonferroni) the differences disappeared. We decided to retain item 33, because of this and in order to keep the content.

In conclusion, we are now sure that our groups are comparable, so all the differences that we find between the forms can be derived from this, and not from the students.

References

Boland, R.J., Lester, N.A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry, 34*(4), 310–316.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford publications.

DiBattista, D., Sinnige-Egger, J., & Fortuna, G. (2014). The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education, 82*(2), 168–183. doi:10.1080/00220973.2013.795127

Downing, S.M. (2005). The effects of violating standard item writing principles on test and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10,* 133–143, DOI 10.1007/s10459-004-4019-5

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93–103.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society, 22*(5), 700–725.

Frey, B.B., Petersen, S., Edwards, L. M., Pedrotti, J.T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*(4), 357–364. doi:10.1016/j.tate.2005.01.008

Haladyna, T.M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37–50.

Haladyna, T.M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51–78.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Haladyna, T.M., Rodriguez, M.C., & Stevens, C. (2019) Are Multiple-choice Items Too Fat?. *Applied Measurement in Education, 32*(4), 350–364, https://doi.org/10.1080/08957347.2019.1660348

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *journal of the American Statistical Association*, *70*(351a), 631–639.

Halloun, I. & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, *53*(11), 1043–1055.

Martínez, R.J., Moreno, R., Martín, I., & Trigo, M.E. (2008). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psichotema, 21*(2), 326–330.

Muthén. & Muthén. (2017). *MPLUS* (Version 7). [Computer Sofware]. https://www.statmodel.com/

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585.

Pachai, M.V., DiBattista, D., & Kim, J.A. (2015). A systematic assesment of "None of the above" on multiple choice test in a first year psychology classroom. *The Canadian journal for the scholarship of teaching and learning, 6*(3), 2–14. http://dx.doi.org/10.5206/cjsotl-rcacea.2015.3.2

R Core Team (2020). *R: A language and environment for statistical computing* (Version 3.6) [Computer Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Stone, A., Allen, K., Rhoads, T.R., Murphy, T.J., Shehab, R.L., & Saha, C. (2004). The statistic concept inventory: a pilot study. *Frontiers in Education Conference.*

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education, 9*, 40. doi:10.1186/1472-6920-9-40

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological methods*, *12*(4), 399–413. https://psycnet.apa.org/doi/10.1037/1082-989X.12.4.399

Cognitive diagnostic computerized adaptive testing in R using the `cdcatR` package

Miguel A. Sorrel[1], Pablo Nájera[1], Francisco J. Abad[1]

[1]*Department of Social Psychology and Methodology, Autonomous University of Madrid, Spain*

Abstract

Cognitive diagnosis models (CDMs) are confirmatory latent class models with important implications for educators and other professionals. Specifically, CDMs provide fine-grained information about skills and cognitive processes. Numerous studies have been published aimed at generating developments that allow the application of computerized adaptive testing based on these models (CD-CAT). Empirical adaptive applications are, however, still scarce. To facilitate research and the emergence of empirical applications in this area, we have developed the R `cdcatR` package. The purpose of this document is to illustrate the different functions included in this package. The illustration includes demonstrations on the CD-CAT item bank and data generation, model selection on the basis of relative fit information, and CD-CAT performance evaluation in terms of accuracy, item exposure, and test length. In conclusion, an R package is made available to researchers and practitioners that allows the application of computerized adaptive tests based on CDMs. This is expected to facilitate the development of empirical applications in this area.

Keywords: Computerized adaptive testing; cognitive diagnosis modeling; R statistical programming.

Funding: This study has been supported by The Ministry of Science, Innovation and Universities of Spain (Grant PSI2017-85022-P) and Psychometric Models and Applications Chair (Institute of Knowledge Engineering (IIC) and Autonomous University of Madrid).

E-mails: miguel.sorrel@uam.es; pablo.najera@uam.es; francisco.j.abad@uam.es

1. Introduction

Cognitive diagnosis models (CDMs) are confirmatory latent class models that can be used to classify examinees in a set of discrete latent attributes. These models emerged and became popular in the educational setting (Haertel, 1989), although they have now spread to other areas such as clinical psychology (Templin & Henson, 2006). A lot of work has been done in recent years on the developments needed to apply these models in computerized adaptive testing (CD-CAT). Empirical applications are, however, still scarce. To facilitate research and the emergence of empirical applications in this area, we have developed the `cdcatR` package (Sorrel et al., 2020) for R (R Core Team, 2020). The purpose of this document is to showcase the functions included in this package. To this end, the statistical foundations will briefly be presented, and two illustrations will be described.

1.1. Cognitive Diagnosis Models

Let *K* denote the number of attributes being measured by the test items. The main output of CDMs consists of an attribute vector $\alpha_i = (\alpha_{i1},\ldots,\alpha_{iK})$. For most CDMs, $\alpha_k \in \{0,1\}$, that is, α_{ik} indicates whether or not the *i*th examinee masters attribute *k*. The number of latent classes that can be formed is given by 2^K. Each of the *J* items in the test typically measures a subset of *K* denoted by K_j^*. Each of the possible combinations of the K_j^* attributes is called a latent group α_{lj}^*. General CDMs, like the generalized deterministic inputs, noisy "*and*" gate (G-DINA, de la Torre, 2011) model, are saturated models, meaning each latent group has its own success probability (or, depending on the context, its probability to endorse the item). This is represented in Figure 1 for two items measuring one and three attributes, respectively.

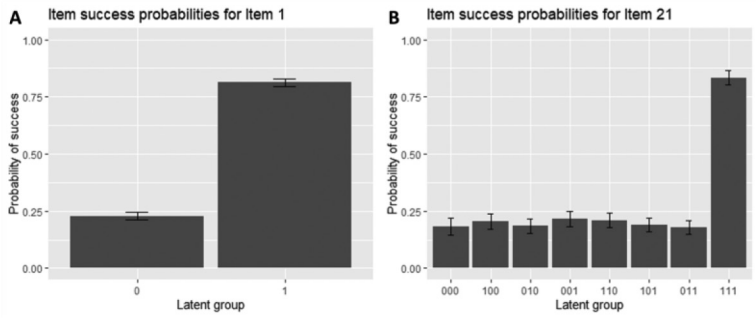


Figure 1. Item response functions for two items under the G-DINA model

In the general case, the probability of success (i.e. the item response function, IRF) is given by (de la Torre, 2011)

$$P\left(\alpha_{lj}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}, \qquad (1)$$

where δ_{j0} is the intercept of item *j*; δ_{jk} is the main effect due to α_k; $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$; and $\delta_{12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. There might be occasions where several groups have a similar probability of success. Specifically, the IRF represented in Panel B of Figure 1 is fairly similar to that of the DINA model (Haertel, 1989),

a non-compensatory model where only examinees mastering the K_j^* attributes have a higher probability of success equal to $\delta_{j0} + \delta_{12...K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$. The rest of the examinees are included in the same latent group with the probability of success δ_{j0}. This makes this model easier to estimate. Relative fit statistics can be used to explore whether these constraints can be imposed without a significant loss of fit. Among other statistics, the two-step likelihood ratio test (2LR; Sorrel et al, 2017) has been found to have an acceptable performance. For computing the 2LR statistic, a G-DINA model calibration is used to approximate the number of examinees and correct responses in each latent group ($\boldsymbol{I}_j = \{I_{\alpha_{lj}^*}\}$ and $\boldsymbol{R}_j = \{R_{\alpha_{lj}^*}\}$, respectively). These values are then used to obtain an approximation to the DINA model item parameters ($\psi_j = \{P\left(\alpha_{lj}^*\right)^{DINA}\}$). For item *j*, 2LR is computed as

$$2LR_j = 2\left[\log L\left(\boldsymbol{P}_j|\boldsymbol{R}_j, \boldsymbol{I}_j\right) - \log L\left(\psi_j|\boldsymbol{R}_j, \boldsymbol{I}_j\right)\right], \tag{2}$$

where $\boldsymbol{P}_j = \{P\left(\alpha_{lj}^*\right)^{G-DINA}\}$ includes the G-DINA item parameters. $2LR_j$ is χ^2-distributed with degrees of freedom equal to the difference in the number of item parameters.

1.2. Application to Computerized Adaptive Testing

Several item-selection rules (ISRs) have been proposed for CD-CATs. In 2099, Chen introduced the posterior weighted Kullback-Leibler (PWKL) index:

$$PWKL_j\left(\hat{\alpha}_i^{(t)}\right) = \sum_{l=1}^{2^K}\left[\sum_{y=0}^{1}\log\left(\frac{P\left(y_j = y\,|\,\hat{\alpha}_i^{(t)}\right)}{P\left(y_j = y\,|\,\alpha_l\right)}\right)P\left(y_j = y\,|\,\hat{\alpha}_i^{(t)}\right)\pi_i^{(t)}\left(\alpha_l\right)\right], \tag{3}$$

where $\hat{\alpha}_i^{(t)}$ is the punctual estimator for the *i*th examinee at step *t* of the CAT application and $\pi_i^{(t)}\left(\alpha_l\right)$ is the posterior distribution of the latent classes. In this case, as in those that follow, the chosen item is the one that maximizes the index. Later, Kaplan et al. (2015) introduced the modified PWKL (MPWKL), an improved version of this index that considers the entire posterior distribution instead of a single estimator. The authors also introduced another index based on the general discrimination index (GDI) previously proposed for the G-DINA model:

$$GDI_j = \sum_{l=1}^{2^{K_j^*}}\pi_i^{(t)}\left(\alpha_{lj}^*\right)\left[P\left(\alpha_{lj}^*\right) - \overline{P}_j\right]^2, \tag{4}$$

where $\overline{P}_j$ is the mean probability of success across the latent groups. GDI and MPWKL performed similarly in terms of accuracy, with the GDI computation proving remarkably faster.

Indices based on the concepts of entropy and divergence (e.g. Minchen & de la Torre, 2016; Xu et al., 2016) have also been tested. The works of Yigit et al. (2019) and Wang et al. (2020) have explored the relationships between several of these indices, finding that they are closely related. One of these indexes is the Jensen-Shannon divergence (JSD) index. This index is computed as (Yigit et al., 2019)

$$JSD_j = S\left(\gamma_j \times \pi^{(t)'}\right) - \sum_{l=1}^{2^K} \pi_l^{(t)} S\left(\gamma_{jl}\right), \tag{5}$$

where γ_j denotes a 2×2^K matrix where the l column represents the probability of success for item j $P\left(\alpha_l\right)$ and its complement $1 - P\left(\alpha_l\right)$ and $S(\cdot)$ the Shannon entropy function.

All the indices discussed so far require the estimation of a statistical parametric model. Alternatively, Chang et al. (2019) made a non-parametric selection (NPS) proposal. The procedure starts by administering K items according to the Q-optimal criterion (Xu et al., 2016). Then, the examinee's attribute vector is estimated using the non-parametric classification method (NPC; Chiu & Douglas, 2013). The NPC uses the Hamming distance to compute the discrepancy between the examinee's response pattern and the ideal responses associated to each attribute profile according to either a deterministic conjunctive or disjunctive rule. Items are randomly selected from those that can differentiate between the ideal response of the current estimate of the attribute vector and the ideal response of the second most likely attribute vector until a test length criterion is reached.

The rest of the components in an adaptive application such as the stopping rule and trait level estimator have also been studied in the previous literature. The most popular estimator is the maximum a posterior (MAP) estimator. The MAP estimator for examinee i is defined as:

$$MAP\left(\alpha_i\right) = \arg\max_{\alpha_l} [P(\alpha_l \mid \boldsymbol{Y}_i)], \tag{6}$$

where $P(\alpha_i \mid Y_i) = L(Y_i \mid \alpha_i)\, \pi(\alpha_i) / \sum_{c=1}^{2^K} L(Y_i \mid \alpha_c)\, \pi(\alpha_c)$ is the posterior distribution and $L(Y_i \mid \alpha_i) = \sum_{j=1}^{j} P(\alpha_i)^{Y_{ij}} [1 - P(\alpha_i)]^{1-Y_{ij}}$ is the likelihood function. Fixed-precision applications are implemented by specifying a specifying a minimum value for the posterior probability of the assigned latent class.

2. Method

Through the various functions, `cdcatR` works with the most common cognitive diagnosis models. It adaptively applies banks of items calibrated according to these models, evaluates the applications in terms of accuracy and use of the item bank, and simulates item banks by manipulating their length, complexity of the Q-matrix, and quality of the items for the different models mentioned. Below are details of two illustrations to show the contents of the package. The R code is available at https://osf.io/vru4e/.

2.1. Illustration 1: CD-CAT for a G-DINA Calibrated Item Bank

The main function of the package is `cdcat()`, which allows for parametric and non-parametric CAT applications. To demonstrate this function, a dataset inspired by the empirical application by Liu et al. (2013) on English assessment was simulated. There were 8 attributes measured by 352 items with a relatively simple structure (330 items were one-attribute items). The authors estimated the DINA model using a sample of 38,600 students. To represent this, we generated an item bank of similar quality items and used the true parameters in a CAT administration to 500 examinees. The item bank and dataset were generated using the `gen.itembank()`

and `gen.data()` functions, respectively. Item parameters resembled the distribution of the DINA parameters reported in Liu et al. (2013). The ISR was GDI. The stopping rule was a fixed-precision rule where all $P(\alpha_l|\boldsymbol{Y}_i)$ had to be greater than 0.90. It was further established that the maximum number of items to be applied per examinee could not exceed 80. The results were described in terms of latent class at the vector and individual attribute levels and CAT length. This information is provided by the `cdcat.summary()` function. The detailed results for two examinees were explored using the `att.plot()` function.

2.2. Illustration 2: Item Selection Rules Comparison

The `cdcat()` function includes an `itemSelect` argument to specify the ISR. Version 1.0.2 of the package includes the ISRs mentioned in the introduction. Prior research indicates that GDI, JSD, and MPWKL are expected to perform very similarly in terms of accuracy, but GDI would be much faster in terms of computation time. To replicate these results, these ISRs were compared in terms of pattern recovery, test overlap, and computation time. On this occasion, the dataset was composed of 500 examinees and 180 items. Items were of medium quality and the number of attributes was 5. The CDM used in the data generation was the DINA model. In an applied context, there is no true model, so it would be incorrect to estimate the DINA model without first checking its fit to the data. The package includes the `LR.2step()` function to compare the fit of different models to that of the saturated model (i.e., G-DINA). This was a preliminary step in this illustration. Subsequently, the CAT was based on the estimates for the selected model. The application stopped after the administration of 15 items (i.e., fixed-length stopping rule).

3. Results

3.1. Illustration 1: CD-CAT for a G-DINA Calibrated Item Bank

Accuracy results are shown in Table 1. The first row of the table indicates that all examinees were correctly classified, at least, in 7 of the 8 attributes. Furthermore, and because the stopping rule dictated that the CAT should end once the posterior distribution for the punctual estimator was higher than 0.90, the proportion of completely correctly classified attribute patterns was approximately 0.90 (i.e., 0.91). The attribute that recovered the worst was attribute 5, but even in this case there were 97% of correct classifications. The distribution for the CAT length across the 500 examinees is represented in Figure 2. The average number of items administered was 16.45, with 30 being the maximum. Figure 3 represents the detailed results for two examinees in two of the eight attributes. Examine A required 22 items to be administered. The estimated latent class was $\alpha_A = (0,1,1,1,1,1,0,0)$. Examinee B required only 13 items to be estimated and was assigned to latent class $\alpha_A = (0,1,1,0,1,1,1,0)$.

Table 1. Accuracy results

#	1	2	3	4	5	6	7	8
#correct out of 8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91
By attribute	0.99	1.00	0.98	0.99	0.97	0.99	0.98	1.00

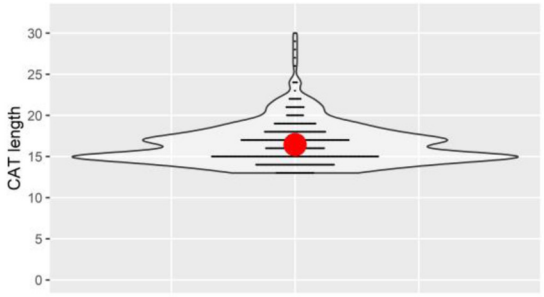


Figure 2. CAT length distribution

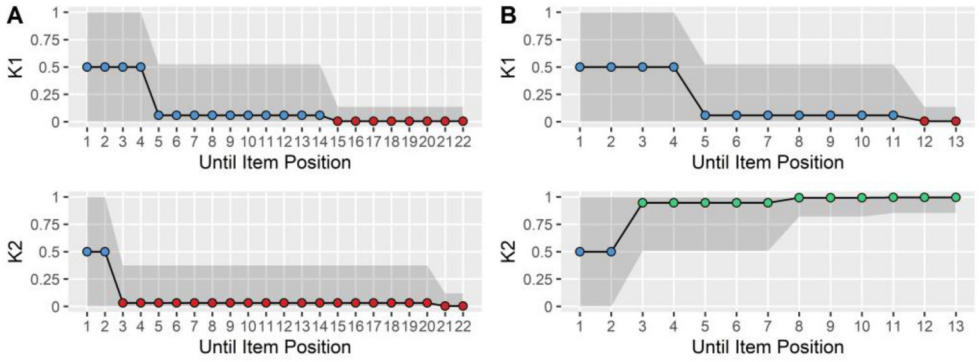


Figure 3. Plot monitoring the attribute mastery status of attributes 1 and 8 for two examinees

3.2. Illustration 2: Item Selection Rules Comparison

In this example, the 2LR statistics always detected the true, generating DINA model. Thus, the CAT applications were based on the DINA model calibration. Figure 4 compares the pattern recovery and item exposure results of the different ISRs. As can be seen from the figure, GDI, JSD, and MPWKL performed in a very similar way in terms of pattern recovery. The performance of NPS and PWKL was comparatively worse. This reflects the strength of global parametric ISRs compared to rules that rely on a punctual estimator of the latent trait (i.e., PWKL) or do not rely on item parameters (i.e., NPS). All the ISRs performed much better than the random selection rule. It should be noted that parametric ISRs had high (higher than 0.50) exposure rates for some items. In high-stakes situations, it would therefore be advisable to

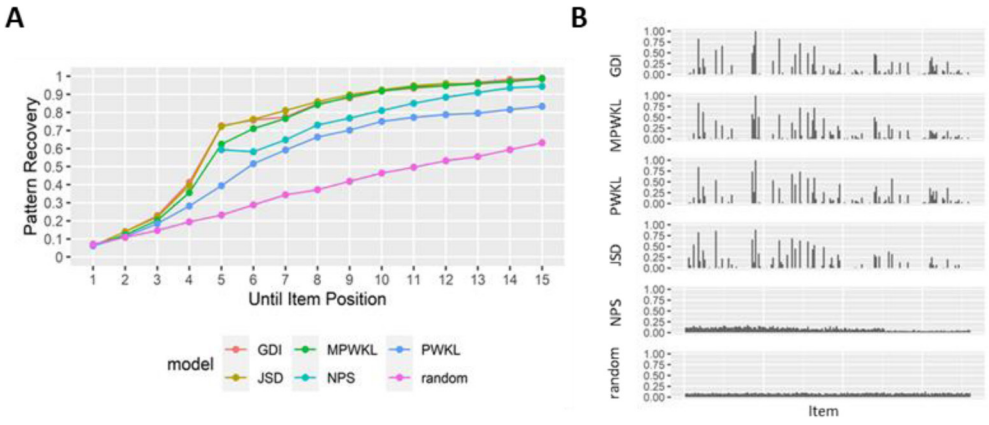


Figure 4. ISRs results in terms of pattern recovery (Panel A) and item exposure rate (Panel B)

implement exposure control methods. The pattern of item usage depicted in the figure shows that these ISRs generally administered a similar subset of items. Regarding computation time, MPWKL was the slowest, and GDI was the fastest ISR: GDI, JSD, MPWKL, PWKL, and NPS required 0.96, 14.24, 36.29, 1.60, and 4.76 ms per item, respectively.

4. Conclusions

This document has described the statistical basis of the R package `cdcatR`. In addition, two illustrations that replicate results from previous literature using the package functions have been described. We hope that this will motivate the development of new empirical applications using this framework. It should be noted that all the results shown here were based on unrestricted CATs. An area that remains to be covered is the implementation of exposure control methods and content (e.g., Wang et al., 2011). Future developments of the package will try to fill this gap.

References

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632.

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*(2), 225–250.

De la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4):301–321.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied psychological measurement*, *39*(3), 167–188.

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, *30*(2), 152–172.

Minchen, N., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence*. Paper presented at the International Meeting of the Psychometric Society, Asheville, NC, United States

R Core Team (2020). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *13*(S1), 39.

Sorrel, M. A., Nájera, P., Abad, F. J. (2020). cdcatR: Cognitive Diagnostic Computerized Adaptive Testing. R package version 1.0.2. https://CRAN.R-project.org/package=cdcatR

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287.

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*(3), 255–273.

Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A Note on the Relationship of the Shannon Entropy Procedure and the Jensen–Shannon Divergence in Cognitive Diagnostic Computerized Adaptive Testing. *SAGE Open*, https://doi.org/ 10.1177/2158244019899046

Xu, G., Wang, C., Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. British Journal of Mathematical and Statistical Psychology, *69*(3), 291–315.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, *43*(5), 388–401.

The identification of the difficulty factor using variance estimates

Karl Schweizer[1], Christine DiStefano[2], Stefan Troche[3]

[1]*Faculty of Psychology and Sports Sciences, Goethe University Frankfurt, Germany,*

[2]*Department of Educational Studies, University of South Carolina, USA,*

[3]*Department of Psychology, University of Bern, Switzerland*

Abstract

A major characteristic of a difficulty factor is defined as possessing factor loadings which reflect the difficulties of the items. The utility of this characteristic for identifying a difficulty factor is called into question as other method factors, such as the item-position and speed factors, may show similar patterns of factor loadings. Our search for another characteristic that distinguishes a difficulty factor from other method factors concentrated on the factor variance. Theoretical analyses suggest that a difficulty factor may show a rather large factor variance if it is allowed to correlate with the main factor (and the correlation is negative). We investigated reasoning data that might give rise to a difficulty factor by Confirmatory Factor Analysis (CFA) models with and without factor correlations. In uncorrelated factors, the factor variance of the difficulty factor was insignificant but reached about half the size of the main factor variance when allowed to correlate. This correlation was negative, as was expected. We conclude that a rather large factor variance in combination with a negative correlation involving the main factor contributes to identifying a difficulty factor.

Keywords: difficulty factor; method effect; confirmatory factor analysis.

E-mail: k.schweizer@psych.uni-frankfurt.de

1. Introduction

When a difficulty factor was reported for the first time, it was characterized as a factor with factor loadings reflecting the difficulty levels of the items and with a lack of specific meaning (Guilford, 1941). This report stimulated a large number of studies that aimed to clarify the conditions leading to a difficulty factor and the underlying source that drives a difficulty factor (Hattie, 1985). In recent years, reports on difficulty factors have become rare. As far as observations of a difficulty factor in confirmatory factor analysis are concerned, researchers can employ other methods for dealing with the systematic variation which a difficulty factor otherwise captures. Such a method allows item uniquenesses to correlate with each other.

In the meanwhile, further characteristics of a difficulty factor have been identified: (1) there are items with very high difficulty levels among the set of investigated items (Bandalos & Gerstner, 2016), and (2) observations of a difficulty factor are restricted to investigations of binary data (Floyd & Widaman, 1995). These characteristics have implications that may further characterize a difficulty factor. The variances of binary variables are quite small and depend on the probability levels of the items. If there are very high difficulty levels, the variances are especially small. This suggests that the amount of systematic variation that is captured by a difficulty factor is rather small.

Another characteristic of a difficulty factor appears to be its lack of a unique underlying substantive dimension. Spearman's explanatory approach of factor analysis suggests that a source of responding to items creates systematic variation that is captured by a factor. This approach suggests that, in the case of several factors capturing systematic variation, there should be several sources of systematic variation. However, as has been demonstrated by McDonald and Ahlawat (1974), a difficulty factor may be observed in the absence of a source that is unique to this factor. There is the possibility that, in this case, a difficulty factor is driven by the same latent source as the main factor. This means that a main factor and a difficulty factor originate from the same underlying source. These factors are likely to show a high correlation between each other.

In confirmatory factor analysis, this relationship is represented within the covariance matrix model that serves parameter estimation and fit investigation (Jöreskog, 1970). Two types of relationships are possible: the first is a positive relationship and the other a negative relationship. In the case of a positive relationship, there is decomposition of the systematic variation of data into parts associated with main and difficulty factors and also a part which the factors have in common. All of these variation components are positive parts. In contrast, in the case of a negative relationship, there are positive parts captured by main and difficulty factors but the part which the factors have in common is negative. An obvious difference between these two options is that, in the case of a negative relationship, a larger range of possible (positive and negative) parameter estimates can be considered for reproducing the empirical covariances than in the case of a positive relationship.

To illustrate this point, we provide the following example regarding two items: assume the covariance of the two items is 0.5. In the case of the first option it can, for example, be reproduced by contributions of 0.2 (main factor), 0.1 (difficulty factor) and 0.2 (common to both factors). No single number can be larger than 0.5. In contrast, in the case of the second option, 0.5 is, for example, reproduceable by contributions of 0.6 (main factor), 0.2 (difficulty factor) and −0.3 (common to both factors). In this case, there is no upper limit of 0.5 for contributions. This means that the second option offers a better precondition for finding parameters that reproduce the empirical covariance matrix well.

In sum, although the variance of a difficulty factor can be expected to be rather small, in combination with a negative relationship between main and corresponding difficulty factors, it can become quite large.

1.1. Analytic strategy

A two-factor confirmatory factor analysis (CFA) model is required for investigating data if one underlying dimension of systematic variation is expected and very high difficulty levels of some items give rise to the expectation of a difficulty factor. Such a model can be designed as a bifactor model (Reise, 2012) or an extended version of the congeneric model (Brown, 2015). This model describes the $p \times 1$ vector of manifest variables, $\mathbf{x}$, as the sum of three components: the product of a $p \times 1$ vector of factor loadings on the main factor, $\boldsymbol{\lambda}_{main\_factor}$, times the main factor, ξ_{main}, the product of the $p \times 1$ vector of factor loadings on the difficulty factor, $\boldsymbol{\lambda}_{difficulty\_factor}$, times the difficulty factor, $\xi_{difficulty}$, and the $p \times 1$ vector of error variables, $\boldsymbol{\delta}$:

$$\mathbf{x} = \boldsymbol{\lambda}_{main\_factor} \times \xi_{main} + \boldsymbol{\lambda}_{difficulty\_factor} \times \xi_{difficulty} + \boldsymbol{\delta}.$$

The estimation of factor variances can be accomplished by the variance parameter of the covariance matrix model (Jöreskog, 1970). Variance parameters regarding the main factor and difficulty factor, ϕ_{main} and $\phi_{difficulty\text{-}factor}$, are included in the $q \times q$ matrix of variances and covariances of factors, $\boldsymbol{\Phi}$, of the model-implied covariance matrix, $\boldsymbol{\Sigma}$. The sizes of variance parameters depend on the other parameters of $\boldsymbol{\Sigma}$ (Schweizer et al., 2019) meaning that they may not be compared with each other unless they are scaled appropriately. The scaling occurs in the following steps: (1) estimation of factor loadings ($\boldsymbol{\lambda}$) with variance parameters fixed to one ($\phi = 1$); (2) transformation of $\boldsymbol{\lambda}$ into $\boldsymbol{\lambda}^*$ ($= c\boldsymbol{\lambda}$) by means of scalar c ($c > 0$) so that the following equation holds,

$$1 = \text{trace}(\boldsymbol{\lambda}^*\boldsymbol{\lambda}^{*\prime});$$

(3) replacement of the parameters of $\boldsymbol{\lambda}$ by constraints according to $\boldsymbol{\lambda}^*$; (4) estimation of φ (which is free for estimation) (Schweizer & Troche, 2019).

Determining the overall variances explained by sets of factors offers an alternative way of learning about the variances of individual factors. Using this alternative way, it is possible to estimate the contributions of individual factors when taking overlaps with other factors into consideration. We define the overall variance, υ, as a function of $\boldsymbol{\Phi}$ that is pre- and post-multiplied with the unity vector:

$$\upsilon = \mathbf{1}' \boldsymbol{\Phi} \mathbf{1}.$$

In order to find out about the genuine variance of a factor when taking overlaps with other factors into account, the overall variance for the set of factors except the factor of interest ($\upsilon_{all\_but\_factor\_of\_interest}$) needs be compared with the overall variance for the set of factors (υ_{all}).

1.2. Objective

A major aim of the research reported in the following sections was to explore the effect of the correlation between main and difficulty factors on the variance estimates and to examine this re-

lationship in a real dataset. Furthermore, research investigated whether the estimated variances would show overly large sizes in the case of a negative relationship.

2. Method

2.1. Participants

Data were collected in a sample of 287 university students. Data collection was part of a comprehensive student assessment that included the application of a number of cognitive scales.

2.2. Material

Eighteen items taken from Raven's APM (Raven, Raven, & Court, 1997) had to be completed within a time span of 20 minutes.

2.3. Statistical analyses

The data were investigated using the software LISREL (Jöreskog & Sörbom, 2006). Models including one factor (reasoning factor), two factors (reasoning and difficulty factors) and three factors (reasoning, difficulty and position-effect factors) were employed. The position-effect factor was additionally considered for reaching a good model fit. The focus of the investigation was the estimates of scaled variance parameters and of overall variances based on scaled variance parameters.

3. Results

3.1. Scaled variance estimates for models without and with correlations between factors

The scaled factor variances estimated for the one-factor, two-factor and three-factor models are shown in Table 1. The first row provides the results for uncorrelated factors. The largest variance estimate for the reasoning factor was observed in the one-factor model. As there was no other factor accounting for systematic variation, this could be considered as the upper limit. Neither of the two estimates for the difficulty factor was significant. The second row provides the results for correlated factors. Negative covariances of difficulty and main factors of −0.59 (two-factor model: $t = 2.66$, $p < .05$) and of -0.55 (three-factor model: $t = 2.66$, $p < .05$) were observed. As a consequence, the sizes of the reasoning factor variances increased by about 100 percent while the size of the difficulty factor variances rose to about 50 percent of the estimate for the corresponding reasoning factor.

Table 1. Scaled factor variances for reasoning, difficulty and position-effect factors of one-factor, two-factor and three-factor models without and with correlations between factors.

Relationship between factors	One-factor model	Two-factor model		Three-factor model		
	R	R	D	R	D	PE
No correlation	0.52*	0.46*	0.09	0.44*	0.09	0.07*
Correlation(s)		0.80*	0.36*	0.91*	0.40*	0.15*

N.B. R abbreviates reasoning factor, D difficulty factor, PE position-effect factor, * $p < .05$.

3.2. Overall variances for models without and with correlations between factors

The overall variances for complete models (see definition of υ; second equation on previous page) are shown in Table 2. The variances reported in the first row of Table 2 were obtained by models without correlations between factors. An increase is seen from the first to the second models and from second to third models. However, after the removal of insignificant parameters (see numbers in brackets), there was no further increase. The variances reported for the models with correlations also yielded increases, but by a smaller

Table 2. Overall variance estimates for one-factor, two-factor and three-factor models without and with correlations between factors.

Relationship between factors	One-factor model	Two-factor model	Three-factor model
No correlation	0.52	0.55 (0.46)[1]	0.60 (0.51) [1]
Correlation(s)		0.52	0.53 (0.58) [1]

[1] Size after removal of insignificant contributions.

amount. The increase from the one-factor model to the two-factor model was 0.007 and from the two-factor model to the three-factor model 0.01. After the removal of insignificant contributions, the latter was 0.06. In this case, the large size of the increase was mainly due to the position-effect factor that was added.

4. Conclusions

A difficulty factor contributes to explaining data if it is allowed to correlate with the main factor of the model. A negative correlation leads to an overestimation of the systematic variation for which a difficulty factor accounts.

References

Bandalos, D. L., & Gerstner, J. J. (2016). Using factor analysis in test construction. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction* (pp. 26–51). Germany: Hogrefe Publishing.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299. http://doi.org/10.1037/1040-3590.7.3.286

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6*(2), 67–77. http://doi.org/10.1007/BF02292175

Hattie, J. (1985): Methodological review: assessing unidimensionality of tests and items. Applied *Psychological Measurement, 9*(2), 139–164. http://dx.doi.org/10.1177/014662168500900204

Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika, 57(2)*, 239–257. http://dx.doi.org/10.2307/2334833

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*(1), 82–99. http://dx.doi.org/10.1111/j.2044-8317.1974.tb00530.x

Raven, J. C., Raven, J., & Court, J. H. (1997). *Raven's progressive matrices and vocabulary scales*. Edinburgh: J.C. Raven Ltd.

Schweizer, K., & Troche, S. (2019). The EV scaling method for variances of latent variables. *Methodology, 15*(4), 175–184. http://doi.org/10.1027/1614-2241/a000179

Schweizer, K., Troche, S., & DiStefano, C. (2019). Scaling the variance of a latent variable while assuring constancy of the model. *Frontiers in Psychology* (Section Quantitative Psychology and Measurement), *10*, ArtID 887. http://doi.org/10.3389/ fpsyg.2019.00887

A cutoff-free method for Q-matrix validation

Pablo Nájera[1], Miguel A. Sorrel[1], Jimmy de la Torre[2], Francisco J. Abad[1]

[1]*Department of Social Psychology and Methodology, Autonomous University of Madrid, Spain,*

[2]*Faculty of Education, The University of Hong Kong, Hong Kong*

Abstract

In cognitive diagnosis modeling (CDM), the Q-matrix identifies the subset of attributes measured by each item. Q-matrix misspecifications negatively impact classification accuracy. Among the several empirical Q-matrix validation methods that have been proposed to address this problem, the GDI method has received the most attention. However, it requires the use of a cutoff point, which might be suboptimal. The Hull method presented here aims to find an optimal fit-parsimony balance without relying on a cutoff point. Furthermore, it can be used either with a measure of item discrimination (PVAF) or a coefficient of determination (pseudo-R^2). Results from a comprehensive simulation study showed that the Hull method obtained great overall accuracy, correctly recovering more than 95% of the Q-matrices. The PVAF consistently obtained slightly better results than the pseudo-R^2. The poor overall performance of the GDI method was due to the condition of high number of attributes. The absence of a cutoff point makes the Hull method a flexible solution to the Q-matrix specification problem in different applied settings.

Keywords: Cognitive diagnosis modeling; diagnostic classification models; diagnostic accuracy; construct validity; Q-matrix.

Funding: This study has been supported by Ministerio de Ciencia, Innovación y Universidades, Spain (Grant PSI2017-85022-P), the European Social Fund, and Cátedra de Modelos y Aplicaciones Psicométricas (Instituto de Ingeniería del Conocimiento and Autonomous University of Madrid).

E-mails: pablo.najera@uam.es; miguel.sorrel@uam.es; j.delatorre@hku.hk; francisco.j.abad@uam.es

1. Introduction

Cognitive diagnosis models (CDMs) are multidimensional latent variable models. In contrast to item response theory models, CDMs' latent variables are discrete and receive the name of *attributes*. Usually employed in the context of educational assessment, the main purpose of CDMs is to classify examinees in latent classes or *attribute profiles*, which reflect the mastery or non-mastery of each attribute. For the usual case of K dichotomous attributes, there are $L = 2^K$ different attribute profiles.

The Q-matrix (Tatsuoka, 1983) is a required input for CDMs. It determines the relationships between the J items and the K attributes. In an item's q-vector (q_j), each q-entry (q_{jk}) can adopt a value of 1 or 0, depending on whether item j measures attribute k or not, respectively. The Q-matrix can be seen as the structural model in a confirmatory factor analysis, and thus it is usually constructed by domain experts. However, the subjective nature of the Q-matrix specification process makes it susceptible to mistakes. Q-matrix misspecifications can greatly disrupt item parameter estimation, which negatively affect the subsequent attribute profile classification (Rupp & Templin, 2008). To address this problem, several empirical Q-matrix validation methods have been proposed in the last years.

The *general discrimination index* method (GDI method; de la Torre & Chiu, 2016) has received the most attention due to some desirable features: good performance, applicability to general CDMs, and easy accessibility due to their inclusion in the `GDINA` package (Ma & de la Torre, 2020). However, it relies on a cutoff point for retaining the suggested q-vectors, which might be suboptimal under different data conditions. Nájera et al. (2019) proposed a formula to predict the optimal cutoff point as a function of the sample size, the test length, and the item quality. However, the formula was developed under a fixed number of attributes ($K = 5$). Even though it obtained accurate predictions, it might seem unrealistic to include all potentially relevant factors in a predictive formula. Thus, the performance of the GDI method under some realistic conditions, such as a high number of attributes (Sessoms & Henson, 2018), remains uncertain.

The purpose of this research is twofold. First, it aims to propose a cutoff-free empirical Q-matrix validation method that can achieve an optimal fit-parsimony balance. Second, it aims to compare the performance of the proposed method with that of the GDI method under a wide range of realistic conditions.

1.1. The Hull method for Q-matrix validation

The Hull method was first developed by Lorenzo-Seva et al. (2011) as a factor retention method. The method aims to find the number of factors that provide an optimal balance between fit and parsimony. To achieve this, a two-dimensional graph (*hull plot*) is created by representing the number of parameters in the *x-axis* and a model fit index (e.g., CFI, RMSEA) in the *y-axis*. Then, different solutions, from 0 to K factors, are depicted in the hull plot forming a monotonically increasing curve. After removing the solutions that fall below the segment connecting any two other solutions (i.e., forming a *convex hull*), the most pronounced elbow in the curve is found by using the *st* index (Ceulemans & Kiers, 2006):

$$st_k = \frac{(f_k - f_{k-1}) / (np_k - np_{k-1})}{(f_{k+1} - f_k) / (np_{k+1} - np_k)}, \#(2)$$

where f_k and np_k denote the model fit index and number of parameters associated with the solution with k factors, respectively. The solution that maximizes the *st* index is retained. Note that

the extreme solutions of the curve (i.e., with 0 and K factors) are not candidates for election, since either the previous or posterior solution is not available.

We propose to apply the Hull method at item level to retain the q-vector that leads to the optimal fit-parsimony balance. The biggest difference with respect to Lorenzo-Seva et al. (2011)'s proposal is the election of the fit index. Two indices are considered in the present study: the *proportion of variance accounted for* (PVAF; de la Torre & Chiu, 2016) and McFadden's pseudo-R^2 (McFadden, 1974). The former is an item discrimination index, while the latter is a coefficient of determination. Both are dependent on the item and q-vector specification. The PVAF relies on the weighted variance of the probabilities of success in the different latent groups (given item j and a q-vector specification). Since a q-vector with specified K^* attributes will always have a higher variance than a nested q-vector, the variance of each q-vector is divided by the variance of the fully-specified q-vector (i.e., the one with all specified attributes). This ratio, which is enclosed between 0 and 1, is the PVAF index. On the other hand, the pseudo-R^2 is an index often used in logistic regression models that measures the proportion of variance accounting for the observed responses. It is computed as

$$R^2 = 1 - \frac{\log\left(L_M\right)}{\log\left(L_0\right)}, \#(3)$$

where L_M denotes the likelihood of the model being tested, and L_0 denotes the likelihood of the null model. The likelihood of the model is computed after estimating the probabilities of success of the examinees' in item j given a q-vector specification. The names Hull_P and Hull_R are used to refer to the Hull variants with the PVAF or pseudo-R^2, respectively.

Once a fit index has been chosen and computed for item j and the different possible q-vectors, the candidate q-vectors for item j are determined. The q-vector with the highest PVAF (or pseudo-R^2) among the q-vectors with the specified K^* attributes is a candidate q-vector. Thus, there are K candidate q-vectors for each item. These candidate q-vectors are expected to be nested (de la Torre & Chiu, 2016), which means that any attributes included in a candidate q-vector with the specified K^* attributes will also be included in the candidate q-vectors with more than the specified K^* attributes. For a general CDM, the number of parameters associated to a q-vector with specified K^* attributes is equal to 2^{K^*}.

An illustration of the Hull method with the PVAF is provided in Figure 1. The algorithm of the Hull method for Q-matrix validation is as follows:

1. Select the candidate q-vectors for item j according to the PVAF or pseudo-R^2.
2. Create a hull plot (Lorenzo-Seva et al., 2011) by representing the number of parameters in the *x-axis* and the fit index (PVAF or pseudo-R^2) in the *y-axis*.
3. Set the *origin* of the hull plot at $np_0 = f_0 = 0$, so that the q-vector with one attribute specified is suitable for election.
4. Remove the q-vectors that are not part of the convex hull (i.e., the most possible upper curve).
    a. If only the origin and the fully specified q-vectors remain, then retain the fully specified q-vector.
    b. If two or more q-vectors remain, go to step 5.
5. Compute the *st* index for the remaining q-vectors. Retain the one that maximizes it.

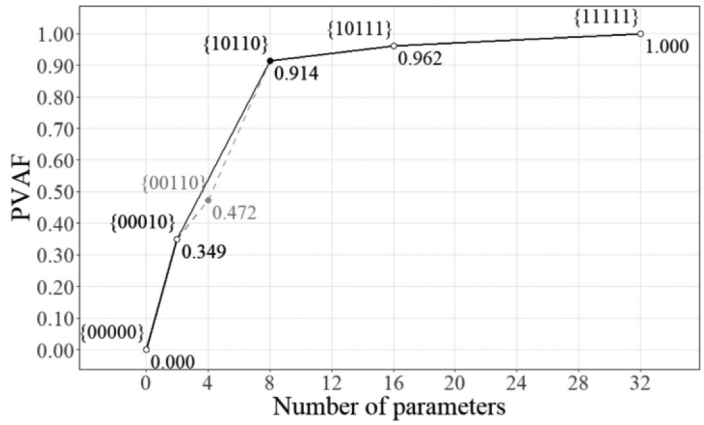


Figure 1. Illustration of a hull plot for item j with the PVAF on the y-axis. The convex hull is represented with the black line. The grey line shows the removed q-vector. In this example, st_1 = 1.85, st_3 = 15.69, and st_4 = 2.53 for {00010}, {10110}, and {10111}, respectively. {10110} is retained as the suggested q-vector.

2. Method

2.1. Design

A simulation study was conducted to test the performance of the three validation methods (GDI, $Hull_P$, and $Hull_R$). Five factors were manipulated: Q-matrix misspecification rate (*QM*; 0, .15, .30), number of attributes (*K*; 4, 8), average item quality (*IQ*; .4, .6, .8), sample size (*N*; 500, 1000), and ratio of number of items to attribute (*JK*; 4, 8). These factor levels were chosen in pursuit of the representativeness of the applied settings (Nájera et al., 2019). A high number of attributes (i.e., *K* = 8), which has not been previously explored in Q-matrix validation studies, is a common scenario in applied studies (Sessoms & Henson, 2018).

The empirical cutoff point was used for the GDI method (Nájera et al., 2019). In addition, both the GDI and Hull methods were implemented iteratively (Nájera et al., 2020), re-estimating the CDM after each Q-matrix modification. Specifically, the *test-attribute* iterative implementation was used. The whole Q-matrix was modified in each iteration by changing the smallest possible number of attributes in each q-vector.

2.2. Data generation

The G-DINA model (de la Torre, 2011) was used to generate examinees' responses. Attribute patterns were generated following a uniform distribution. Item quality (*IQ*) was defined as the difference in success probabilities between the latent group that masters all the required attributes and the one that masters none of them (see Nájera et al., 2019). The sum of the parameters associated to an attribute was constrained to be higher than .15 to ensure relevant effects for all attributes.

The true Q-matrices generation process followed the recommendations for identifiability by Xu and Shang (2018). They were randomly generated with the following constraints: each Q-matrix contained at least two identity matrices; the Q-matrix had 50% of one-attribute items, and 25% of two-attribute and three-attribute items; the maximum correlation between attributes in the Q-matrices was set to 0.3 to avoid overlapping. On the other hand, Q-matrix misspecifications were randomly introduced with two constraints: all items had to be measured by at least one attribute, and the first identity matrix was always retained.

One hundred datasets were generated for each simulation condition. All simulations and analyses were conducted in R software (R Core Team, 2019), using the `GDINA` package and self-developed functions.

2.3. Dependent variables

The main dependent variable was the Q-matrix recovery rate (*QRR*), which is the proportion of q-entries correctly specified after the Q-matrix validation. The true positive rate (*TPR*) and the true negative rate (*TNR*) were also computed. The *TPR* can be understood as the specificity: the proportion of correctly specified q-entries among the originally correctly specified ones. The *TNR* can be understood as the sensitivity: the proportion of correctly specified q-entries among the originally misspecified ones. Two classification accuracy measures were also examined. The proportion of correctly classified attributes (*PCA*) and the proportion of correctly classified attribute profiles (*PCP*) were computed. Finally, the convergence rate (*CR*) was also recorded.

3. Results

Medians instead of means are provided for *QRR*, *TNR*, *TPR*, *PCA*, and *PCP* due to the presence of asymmetry. The overall results for the three validation methods are shown in Table 1. The Hull method obtained the best overall performance in all dependent variables, with a convergence rate close to 1. The Hull_p variant performed consistently better than the Hull_R, with a *QRR* and *TNR* $\geq$ .961, and a *TPR* = .842. The high Q-matrix recovery of both variants resulted in high overall accuracy in classifying attributes (*PCA* = .882) and attribute profiles (*PCP* $\geq$ .558). On the other hand, the GDI method obtained poor overall performance (*QRR* = .703), with low specificity (*TPR* = .756) and sensitivity (*TNR* = .403). This resulted in low attribute profile classification accuracy (*PCP* = .314). Furthermore, the GDI method obtained a very low convergence rate (*CR* = .393).

Table 1. Overall results

Method	*QRR*	*TPR*	*TNR*	*PCA*	*PCP*	*CR*
GDI	.703	.756	.403	.806	.314	.393
Hull_p	**.961**	**.970**	**.842**	**.882**	**.566**	.995
Hull_R	.955	.969	.814	**.882**	.558	**.996**

Note. Best result by dependent variable is highlighted in bold.

A separate ANOVA for each method was conducted to determine the factors with the greatest effect on their *QRR*. The most relevant factors according to eta-partial-squared were *QM*, *IQ*, and *K* (see Table 2). The Hull_p variant obtained the best *QRR* under all factor levels, except for *IQ* = .8, where the Hull_R variant performed slightly better. The better performance of Hull_p over Hull_R was more prominent under the most unfavorable conditions (i.e., *QM* = .30, *IQ* = .4, and *K* = 8). Another notable result was the effect of *K* on the GDI method. While it obtained very good results with *K* = 4, its performance dramatically decreased with *K* = 8. An additional analysis revealed that the GDI method's convergence rate with *K* = 8 was as low as .015.

Table 2. *QRR* across the different factor levels

	QM			***IQ***			***K***	
Method	0	.15	.30	.4	.6	.8	4	8
GDI	.906	.742	.641	.688	.719	.703	.938	.340
$Hull_P$	**.977**	**.945**	**.902**	**.867**	**.961**	.973	**.953**	**.965**
$Hull_R$	**.977**	.944	.891	.859	.953	**.974**	**.953**	.963

Note. Best result by factor level is highlighted in bold.

4. Conclusions

In cognitive diagnosis modeling, a correct Q-matrix specification is a necessary condition for an accurate attribute profile classification (Rupp & Templin, 2008). This is why many empirical Q-matrix validation methods have been proposed in recent years. Among the many options, the GDI method (de la Torre & Chiu, 2016) has received the most attention. Despite its advantages, it relies on a cutoff point to retain the suggested q-vectors, which might be suboptimal under certain data conditions. To address this, the Hull method for Q-matrix validation is proposed here as a more flexible, cutoff-free alternative that can be used in a variety of applied settings. The Hull method aims to obtain the optimal balance in fit and parsimony while retaining the suggested q-vectors. Furthermore, it can be used with different fit indices. The PVAF (de la Torre & Chiu, 2016) and the pseudo-R^2 (McFadden, 1974) were used in the present study.

Results from the simulation study showed great overall performance of the Hull method, especially the $Hull_P$ variant (*QRR* = .961). It achieved the standard goals of specificity (*TPR* > .95) and sensitivity (*TNR* > .80), while having an almost perfect convergence rate. Because of this, the $Hull_P$ can be recommended as the empirical Q-matrix validation method to be used in many applied settings. On the contrary, the GDI method showed bad overall results. Even though it obtained very good performance with $K = 4$, which is in line with the results found by Nájera et al. (2019), the lack of convergence achieved under $K = 8$ resulted in an extremely low Q-matrix recovery rate.

This research is not without its limitations, and future studies should be conducted to improve our knowledge in Q-matrix validation. For instance, empirical validation methods (and the Hull method is not an exception) assume that the number of attributes specified in the Q-matrix is correct. Dimensionality estimation methods, which have been widely studied in the factor analysis framework, have not been systematically examined in CDMs. Moreover, some relevant factors, such as the complexity of the Q-matrix, remain unexplored. Further research is required to extend the applicability of the Hull method to an even wider range of conditions. Finally, it is important to emphasize that empirical Q-matrix validation methods are not supposed to replace the contribution of domain experts in the Q-matrix construction process, but to complement it. The suitability of the modifications suggested by the validation methods should therefore be revised by the experts. Using both theoretical foundations and empirical support is the best way to ensure an interpretable and valid measurement instrument.

References

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, *59*, 133–150.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 528–529.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*, 340–364.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*, 1–26.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Economics* (pp. 105–142). Academic Press.

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement*, *79*, 727–753.

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2020). Improving robustness in Q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement*. DOI: 10.1177/0146621620909904

R Core Team (2019). R (Version 3.6) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*, 1–17.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconception based on item response theory. *Journal of Education Statistic*, *20*, 345–354.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*, 1284–1295.

Using information criteria to determine the number of factors in maximum-likelihood exploratory factor analysis

Eric Klopp[1]

[1]*Department of Education, Saarland University, Germany*

Abstract

In exploratory factor analysis, determining the number of factors is an essential task for which a variety of criteria exists. Determining the number of factors has recently been discussed in the context of model selection, in which the use of information criteria is common. Studies investigated the use of AIC and the BIC, but until now, there is no detailed investigation of how information criteria perform in determining the number of factors. We investigate the ability of AIC, BIC, SBIC, and HBIC to determine the number of factors in maximum-likelihood EFA. We use various sample sizes and factor structures in Monte Carlo simulations with different population factor loadings and we also use factor structures derived from real examples like the Holzinger Swineford and Big Five datasets as population models. The results show that the information criteria were apt to recover the correct number of factors. The results also show that the three information criteria differ in their performance.

Keywords: Exploratory factor analysis, number of factors, information criteria, model selection, Monte Carlo simulation.

E-mail: e.klopp@mx.uni-saarland.de

1. Introduction

Exploratory factor analysis (EFA) is widely used in psychology and other social sciences. Determining the number of factors is a central problem in applying EFA for which plenty of methods exist (e.g., Fabrigar & Wegener, 2012). Recently, Preacher, Zhang, Kim, and Mels (2013) discussed the choice of the number of factors in the context of model selection. Information criteria like AIC or BIC are common in selecting the number of latent variables in a model (e.g., Bollen, Harden, Ray, & Zavisca, 2014). Until now, studies have investigated the use of AIC (e.g., Akaike, 1987, Ichikawa, 1988) and the BIC (e.g., Golino & Epskamp, 2017). However, these studies considered only a limited range of EFA models or targeted another method and used IC only for comparison. This is astounding because the BIC tends to select the true model if the true model is among the candidate models (Hertzog, 2019; Vrieze, 2012) which is the case in choosing the correct number of factors in a series of models with an increasing number of factors. Additionally, there are more variants of the BIC like the sample-size adjusted BIC (SBIC; Scolve, 1987) and Houghton's BIC (HBIC; Haughton, 1988) that should be given consideration.

1.1 The EFA model, maximum-likelihood estimation, and information criteria

In this paper, we use the common EFA model, in which m latent factors explain the correlational pattern of a set of p normally distributed manifest variables. Due to rotational indeterminacy and under the usual assumptions, we can, without loss of generality, assume that the factors are uncorrelated (Bartholomew, Knott, & Moustaki, 2011, p. 48), and the EFA model is

$$\boldsymbol{R}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \tag{1}$$

in which $\boldsymbol{\Lambda}$ is a $p \times m$ loading matrix, $\boldsymbol{\Theta}$ is a $p \times p$ diagonal matrix of the manifest residual variances, and $\boldsymbol{R}(\boldsymbol{\theta})$ is the $p \times p$ correlation matrix of the manifest variables. The ML estimation for a given sample size N provides a test statistic T_{ML}. Under the null hypothesis, i.e., the model has the correct number of factors, the test statistic is χ^2-distributed with df degrees of freedom. The number of degrees of freedom and the expectation is given by

$$df = p(p+1)/2 - \left(pm + p - m\left(m-1\right)/2\right), \text{ and} \tag{2}$$

$$E\left(T_{ML}\right) = df. \tag{3}$$

Equation (2) shows that the number of degrees of freedom is an upper limit for the number of factors in a model. Only as many factors can be extracted as there are positive degrees of freedom. The calculation of the information criteria draws on the test statistic and the degrees of freedom of a model. In addition to this, equation (3) means that for a correctly specified model, the test statistic should equal the number of degrees of freedom on average, or, for a single model, should be close to the number of degrees of freedom with a high probability. The calculation of the information criteria draws on the test statistic and the degrees of freedom. According to Bollen et al. (2014), the information criteria, which we consider in this paper, are given by:

$$\text{AIC} = T_{ML} - 2df \tag{4}$$

$$\text{BIC} = T_{ML} - \log Ndf \tag{5}$$

$$\text{SBIC} = T_{ML} - \log \frac{N+2}{24} df \tag{6}$$

$$\text{HBIC} = T_{ML} - \log \frac{N}{2\pi} df \tag{7}$$

In a model selection task, the information criteria provide a rank order of the competing models. The model with the smallest value of an information criterion is selected, and the expectation is that this model is the most likely one to replicate (Kline, 2016, p. 287).

1.2 Determining the number of factors – an issue of correct model specification

The usual procedure to determine the number of factors is to specify a series of candidate models with an increasing number of factors to extract. If the EFA model holds, then the true model is among the candidate models. As mentioned above, the number of degrees of freedom of a model is an upper bound for the possible number of factors. For the following consideration, we look at a population model according to equation (1) for which the true number of factors is m_{dgp}. Let m_s be the number of factors that are contained in a specified model so that it is possible to distinguish three cases:

1. Case $m_s < m_{dgp}$:
   In this case, the model is misspecified because there are not enough factors, and the fewer the number of factors, the larger the misspecification. Due to the misspecification, the test statistic follows a non-central -distribution (Browne, 1984) with the non-centrality parameter being directly related to the size of misspecification, i.e., the fewer the number of factors, the larger the test statistic. Therefore, we expect that the test statistic is on average greater than the number of degrees of freedom.
2. Case $m_s = m_{dgp}$:
   In the second case, the model is correctly specified. We expect the test statistic to equal the number of degrees of freedom on average.
3. Case $m_s > m_{dgp}$:
   In the last case, the model is overfitted. Therefore, the test statistic is no longer -distributed (Hayashi, Benter, & Yuan, 2007). The larger the number of factors, the smaller the test statistic becomes (Bartholomew et al., 2011, p. 58). We expect the test statistic to be smaller on average than its expected value.

In combination with equations (4) to (7), it is evident that the information criteria should be at a minimum in all three cases if the number of specified factors equals the number of factors in the dgp, which is consistent with the statistical theory behind the information criteria.

1.3 Goal of the current study

The current study aims to provide systematic simulation studies using artificial and real-world data generating processes of how AIC, BIC and the variants of BIC perform in determining

the number of factors. Additionally, the goal is to examine the performance of the information criteria using artificially created factor structures and real-world factor structures taken from the psychological domain.

2. Method

To study the aptness of the information criteria to determine the correct number of factors in EFA, we conducted a series of Monte Carlo studies. In the first series of studies, we used artificial dgps in which we varied the number of factors (1, 2, or 3 factors), the number of main loadings (4 or 8) and the size of the main loadings (small: .30-.45, medium: .45-.65, large: .50-.80) while the cross-loadings were randomly drawn from the interval [.00,.20]. The conditions with small loadings represented a factor structure with generally low communalities, while these conditions with medium loadings represented a factor structure with medium communalities. The conditions with large loadings represented a factor structure with large communalities. Due to rotational indeterminacy, we assumed uncorrelated factors in the dgps.

In the second series of studies, we used a 3-factor structure resulting from the Holzinger Swineford (HS) data set and a 5-factor structure resulting from the BFI data set (Revelle, 2020). To construct a dgp, we factor analyzed each data set assuming orthogonal factors and used the analysis results as dgps for simulation.

We used R (R Core Team, 2020) and the MASS package (Venables & Ripley, 2002) for all simulations. For each dgp, we simulated 1000 correlation matrices for sample sizes 100, 200, 300, 400, 500, 750, and 1000 that were factor analyzed with the base R factanal function. We calculated the number of correct decisions (hit rate), the bias in the number of factors, and the mean of the information criteria over the replications as measures. The hit rate is the number of cases in which the information criteria select the correct number of factors over the 1000 replications. The bias is the mean of the difference between the true number of factors and the number of factors selected by an information criterion over the number of replications.

3. Results

In the following, we provide selected results for the simulation results due to space limitations. Detailed results are available upon request. For the artificial dgps, we present the hit rate for the population models with three factors for all loading conditions and samples sizes in Table 1.

Table 1. Hit rates for the 3-factor dgps for all information criteria.

Loading	**small**				**medium**				**large**			
N	AIC	BIC	SBIC	HBIC	AIC	BIC	SBIC	HBIC	AIC	BIC	SBIC	HBIC
100	0.061	0.000	0.261	0.004	0.475	0.001	0.429	0.162	0.776	0.895	0.362	0.968
200	0.110	0.000	0.071	0.000	0.793	0.020	0.819	0.422	0.802	1.000	0.849	0.996
300	0.164	0.000	0.030	0.000	0.842	0.167	0.941	0.733	0.816	1.000	0.954	0.999
400	0.249	0.000	0.020	0.000	0.814	0.442	0.976	0.918	0.799	1.000	0.976	1.000
500	0.319	0.000	0.023	0.000	0.802	0.763	0.992	0.983	0.795	1.000	0.988	0.999
750	0.548	0.000	0.033	0.001	0.837	0.990	0.999	1.000	0.806	1.000	0.999	1.000
1000	0.666	0.000	0.067	0.002	0.785	1.000	0.999	1.000	0.764	1.000	1.000	1.000

The results in Table 1 show that in the conditions with small loadings, AIC outperforms BIC and its variants, except for a very small sample size of N =100 in which the SBIC has at least a marginal hit rate. For the condition with large loadings, BIC and SBIC outperform AIC. Concerning BIC variants, BIC and the HBIC outperform the SBIC. However, in this condition, the SBIC does not perform well for a very small sample size.

Table 2. Hit rates and bias for the HS dgp for all information criteria.

	Hit rate				Bias			
N	AIC	BIC	SBIC	HBIC	AIC	BIC	SBIC	HBIC
100	0.845	0.707	0.588	0.928	−0.145	0.294	−0.456	0.006
200	0.850	0.994	0.894	0.994	−0.152	0.006	−0.108	−0.006
300	0.861	1.000	0.958	0.998	−0.141	0.000	−0.042	−0.002
400	0.849	1.000	0.978	1.000	−0.156	0.000	−0.022	0.000
500	0.845	1.000	0.984	0.998	−0.160	0.000	−0.017	−0.002
750	0.851	1.000	0.995	1.000	−0.159	0.000	−0.005	0.000
1000	0.844	1.000	0.999	1.000	−0.162	0.000	−0.001	0.000

Table 2 shows the hit rates and the bias for the HS dgp. The HS dgp corresponds to a dgp with rather large communalities. The results for the hit rate are similar to the results for the artificial dgp in the condition with large factor loadings. Concerning the bias, the results show that AIC and SBIC generally select too few factors regardless of the sample size, whereas BIC tends to pick too many factors for very small sample sizes. The bias of the HBIC is, in general, relatively low regardless of the sample size.

Table 3. Hit rates and bias for the BFI dgp for all information criteria.

	Hit rate				Bias			
N	AIC	BIC	SBIC	HBIC	AIC	BIC	SBIC	HBIC
100	0.777	0.031	0.112	0.708	−0.189	1.989	−2.221	0.302
200	0.812	0.693	0.900	0.992	−0.229	0.335	−0.113	0.008
300	0.825	0.989	0.989	1.000	−0.185	0.011	−0.011	0.000
400	0.814	1.000	0.998	1.000	−0.213	0.000	−0.002	0.000
500	0.829	1.000	0.999	1.000	−0.189	0.000	−0.001	0.000
750	0.816	1.000	1.000	1.000	−0.197	0.000	0.000	0.000
1000	0.825	1.000	1.000	1.000	−0.189	0.000	0.000	0.000

Table 3 shows the hit rates and the bias for the BFI dgp. This dgp corresponds again to a condition with medium to large communalities. For a very low sample size, AIC and HBIC outperform BIC and HBIC. For a sample size of N = 200, the SBIC and HBIC have a high hit rate, closely followed by AIC. Beginning at a sample size of N = 300, all BIC variants outperform AIC. Again, AIC tends to select too few factors regardless of the sample size. The SBIC

also underestimates the number of factors. However, the bias rapidly decreases with a growing sample size. Table 4 shows the mean of the information criteria for various sample sizes and the number of extracted factors. Consistent with our theoretical expectations, the information criteria show the minimum mean at the true number of factors. The only exception is the SBIC for a sample size of $N = 100$ where the mean minimum mean indicates three instead of five factors. This result reflects the hit rates and the bias where the SBIC shows a very low hit rate and a large negative bias.

4. Conclusions

The results show that, in general, the information criteria are apt to recover the correct number of factors. BIC and its variants mostly outperform AIC. A possible reason is that AIC does not correct for parsimony because the term penalty term $2df$ "is a technical correction for the small sample bias in estimating the mean expected log-likelihood" for a model with a certain number of free parameters (Mulaik, 2009, p. 350). Thus, the general recommendation favors BIC and its variants. However, the results indicated that AIC should be preferred when both communalities and sample size are low.

Nevertheless, against a broad application of the SBIC and HBIC speaks the purpose of this criteria. The SBIC was developed to find clusters in multivariate samples (Scolve, 1987), and the HBIC was developed for samples from exponential distribution (Houghton, 1988). It is unclear how these two different backgrounds relate to EFA settings. Thus, if sample size and communalities are reasonable, the BIC should be applied.

A limitation is that in applications, the communality of the indicators is not known. However, it would be possible to determine the squared multiple correlations of the indicators that is a lower bound of the communality (Fabrigar & Wegener, 2012, p. 43; Ramsay & Gibson, 2006). In conjunction with the sample size, the squared multiple correlations provide a picture of the researcher's situation. Consequently, the researcher can decide which information criteria should be preferred.

A further limitation is the assumption that $\boldsymbol{\Theta}$ is a diagonal matrix. In applications, this assumption is usually violated and there is some covariation between the measurement residuals. However, the correlated residuals are empirically indistinguishable from factors (a phenomenon known as bloated specifics) that can potentially distort the number of factors indicated by an information criterion or any other criterion.

Table 4. Means of the information criteria for various sample sizes and various numbers of extracted factors.

N	Number of factors	*df*	AIC	BIC	SBIC	HBIC
100	3	228	−88.666	−682.645	37.436	−263.609
100	4	206	−135.164	−671.829	−21.229	−293.226
100	5	185	−160.431	−642.387	−58.111	−302.380
100	6	165	−155.341	−585.194	−64.083	−281.945
100	7	146	−146.323	−526.678	−65.574	−258.348
300	3	228	139.588	−704.875	18.207	−285.839
300	4	206	−37.482	−800.461	−147.151	−421.858

N	Number of factors	*df*	AIC	BIC	SBIC	HBIC
300	5	185	−177.943	−863.143	−276.432	−523.136
300	6	165	−172.371	−783.496	−260.213	−480.246
300	7	146	−162.534	−703.287	−240.261	−434.957
500	3	228	380.171	−580.760	142.927	−161.724
500	4	206	77.230	−790.979	−137.122	−412.377
500	5	185	−180.241	−959.943	−372.742	−619.936
500	6	165	−174.760	−870.170	−346.450	−566.920
500	7	146	−164.949	−780.282	−316.869	−511.952
750	3	228	689.216	−364.160	359.829	54.876
750	4	206	228.548	−723.187	−69.056	−344.584
750	5	185	−183.005	−1037.718	−450.271	−697.711
750	6	165	−177.768	−940.080	−416.141	−636.830
750	7	146	−167.753	−842.283	−378.676	−573.953

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3), 317–332.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Oxford: Wiley-Blackwell.

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1), 1–19.

Browne, M. W. (1984). Asymptotically distribution-free methods for analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford: Oxford Univ. Press.

Golino, H. F., & Epskamp, S. (2017, 06). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE*, 12(6), 1–26.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16, 342–355.

Hayashi, K., Bentler, P. M., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 505–526.

Ichikawa, M. (1988). Empirical assessments of AIC procedure for model selection in factor analysis. *Behaviormetrika*, 15(24), 33–40.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth ed.). New York: The Guilford Press.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton: Chapman & Hall/CRC.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48(1), 28–56.

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ramsey, P. H. & Gibson, W. A. (2006). Improved communality estimation in factor analysis. *Journal of Statistical Computation and Simulation*, 76(2), 93–101.

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research* [Computer software manual]. Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych (R package version 2.0.9)

Scolve, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333–343.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). New York: Springer.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2), 228–243.

An integrated model of competences related to academic performance: a mixed methods approach

Juan F. Luesia[1], Milagrosa Sánchez-Martín[1], Isabel Benítez[2]

[1]*Department of Psychology, Universidad Loyola Andalucía, Spain,*

[2]*Department of Methodology for Behavioral Science, University of Granada, Spain*

Abstract

Purpose: The evaluation of academic performance is commonly addressed by assessing cognitive skills. Furthermore, conclusions of these studies are based on quantitative results, which makes the interpretation of the causes originating the outputs difficult. The aim of this study is to apply a mixed methods framework where different sources of information are integrated in order to draw up a comprehensive model about competences in students related to their academic performance. *Method*: two sources of information were used: First, previous studies were examined through a systematic review. Then, experts' answers about the variables perceived as relevant were extracted from the narratives obtained by conducting different focus groups. *Results*: A convergence model was obtained by integrating results from the two sources. A total of 43 competences were obtained through two sources of information. All the competences were grouped as cognitive and non-cognitive and both were included in a comprehensive model. The relative importance of each variable was considered. *Conclusions***:** A mixed methods approach seems useful to develop a broad model of competencies. Future steps will focus on adding components to the model in order to reflect all the variables affecting academic performance.

Keywords: Mixed methods, prediction, academic performance, systematic review, focus groups.

E-mails: jfluesia@uloyola.es; msanchez@uloyola.es; ibenitez@ugr.es

1. Introduction

Traditionally, within the university context, the prediction of academic performance has been addressed through the assessment of the "intelligence" construct (Donnon et al., 2007). Under this approach, many protocols have been developed such as the Scholastic Assessment Test (College Board, 2017), the Medical College Admissions Test (Association of American Medical Colleges, 2018) and the Graduate Record Examinations (Educational Testing Service, 2015).

Other approaches have focused on analyzing the influence of non-cognitive competences on academic performance such as personality traits, learning strategies and interpersonal skills, (Albanese et al., 2003; Richardson et al., 2012; West and Sadoski, 2011; Zhou et al., 2016; Zimmermann et al., 2017); attitudinal variables such as intrinsic / extrinsic motivation, empathy and the quality of interpersonal relationships (Carrothers et al., 2000; Petrides and Furnham, 2000).

Previous studies reviewing the non-cognitive competences associated with academic performance have grouped the competences by using different classifications. For instance, Richardson et al. (2012) divided the 43 competences found in a review into 5 groups: personality traits, motivation factors, self-regulatory learning strategies, students' approach to learning and psychosocial and contextual influences.

The diversity of approaches when studying the variables associated with academic performance is also linked to the lack of protocols for assessing the competences influencing academic performance. Identifying which academic competences are relevant at the start of a degree would help to improve students' academic success (Ferguson et al., 2002). Identifying the academic competences of undergraduate students which are relevant before starting university would help improve their academic journey in order to improve their academic success (Sommerfeld, 2011),

Moreover, the inclusion of the intended variables in a single protocol could improve the predictive validity of models related to academic performance. That is, in addition to intelligence (cognitive skills) and previous qualifications, the evaluation of people´s characteristics and their context could be useful for developing a comprehensive protocol for evaluating competences (Furnham et al., 2002). In this regard, mixed methods approaches provide a useful framework to obtain a more comprehensive model to gather the intended competences and to interpret the results. This approach would allow the researcher to better understand complex issues and develop a more complete understanding of the topic (Creswell and Plano Clark, 2011).

For this reason, the aim of this work is to obtain a model of competences based on a mixed methods approach through two sources of information: (1) a systematic review of studies analyzing variables related to the academic performance of new university students; (2) experiences of teachers and professionals about academic competences connected to academic success collected through focus groups. Both sources of information are integrated in order to analyze the overlap between them and propose a classification of the core competences determining the academic performance of students.

2. Method

2.1. Systematic review

A systematic literature review was conducted based on the preferred reporting items for systematic reviews and meta-analyses (PRISMA). In this case, PRISMA for abstract 12-item checklist was followed (Beller et al., 2013).

The inclusion and exclusion criteria were defined as shown in Table 1:

Table 1. Inclusion and exclusion criteria.

Criteria	Inclusion criteria	Exclusion criteria
Population	College students or undergraduate students.	Others: clinical population, childhood, non-university adults and elderly persons.
Assessment: competences	Competence/s is/are assessed studies in an assessment protocol.	There are no measures of competences or there are none included in a protocol.
Period of assessment	Before starting university.	During or after graduation.
Assessment: academic performance	The relationship of competences with college academic performance is analyzed.	Predictive validity of competence/s is/are not analyzed.
Design	Correlational non-intervention designs.	Other designs.
Language	English or Spanish.	Other languages.

The literature search was performed through the following electronic bibliographic databases: Web of Science, Scopus, ERIC, PsycINFO and PsycTEST. Grey literature was also examined. The following search terms and derivatives were used and combined using Boolean operators: *undergraduate student*, *prediction*, *competence*, *admission assessment* and *academic performance.*

The study selection was conducted by two independent researchers in two different phases: first, a screening of titles and abstracts was analyzed for their eligibility; and second, full text articles were reviewed for final inclusion. Risk of bias was addressed by the Newcastle-Ottawa Scale (NOS) for non-randomized studies (Wells et al., 2009).

2.2. Focus groups with experts

Relevant information was collected through the focus group methodology. 30 university professionals were recruited in four groups (deanery, course coordinators, degree coordinators and college services).

Interviews were recorded and later transcribed in July 2019. An analysis was carried out for identifying academic competences which were considered relevant in academic performance. These competences were coded and analyzed according to different characteristics (frequency, consensus, discrepancy, number of arguments...) in a coding template developed in Excel (Onwuegbuzie et al., 2011; Rabiee, 2004).

3. Results

3.1. Systematic review

A total of 2,681 articles were identified after removing duplicates. Finally, 22 articles were included in the study according to the review conducted by two independent researchers.

A total of 20 different competences were identified, both cognitive and non-cognitive. Table 2 illustrates the competences and specifies whether these competences significantly predicted subsequent academic performance. We used the model proposed by Richardson et al. (2012) for grouping the competences.

3.2. Focus groups with experts

The focus group narratives were transcribed and coded by identifying the competences described by the experts. A list of 38 competences relevant to university academic performance were obtained.

The experts participating in the different focus groups described 23 specific competences: *Social skills, social commitment, humility, gratitude, honesty, otherness, service vocation, respect, proactivity, community participation, university role, reasoning, maturity, resilience, autonomy, self-efficacy, self-confidence, academic orientation, interest in other cultures, self-criticism, patience, creativity and teamwork*.

Table 2. Relevant competences obtained by the systematic review, using Richardson categories.

Cognitive abilities[c]	**n**[a]	**%**[b]	**Personality traits**	**n**[a]	**%**[b]	**Motivation factors**	**n**[a]	**%**[b]
Mathematical ability	9	77.8	Emotional intelligence	6	16.7	Motivation	7	28.6
Verbal ability	13	61.5	Conscientiousness	2	50.0	Locus of control	1	100.0
Spelling	13	61.5	Leadership[d]	2	100.0	Effort[d]	1	100.0
			Procrastination	1	100.0			
			Agreeableness	1	100.0			
			Extraversion	1	100.0			
Learning strategies	**n**[a]	**%**[b]	**Psychosocial influences**	**n**[a]	**%**[b]			
Critical thinking	4	50.0	Adaptation	3	33.3			
Time management	2	50.0	Stress/Anxiety	3	33.3			
Concentration	4	25.5	Communication skills[d]	3	66.7			
Perseverance	3	33.3						
Organization	4	25.0						

[a] Number of articles used to assess the competence.

[b] Percentage of articles in which the competence evaluated correlates significantly with academic performance.

[c] Category not identified by the author.

[d] Competences not initially identified by the author.

3.3. Integration of results

The next step consisted of analyzing the convergence between the competences extracted by the systematic review and the focus groups. A total of 15 competences appeared in both sources of information: *spelling, verbal ability, mathematical ability, conscientiousness, motivation, critical thinking, communication skills, effort, leadership, adaptation, concentration, perseverance, time management, organization and emotional intelligence.*

4. Conclusions

The aim of the study was to propose a convergence model of competences related to academic performance. To do this, a systematic review and focus groups were developed.

A total of 43 competences were obtained from the two sources of information. Among them, 15 competences were common to both sources whereas 5 competences appeared only in the systematic review and 23 were described by experts.

As expected, including experts' experiences as a source of information helped us to interpret relevant competences that had already been previously obtained and to build a broader model of academic competences.

The non-cognitive competences classification by Richardson et al. (2012) was useful for defining the number of non-cognitive competences that could be relevant to academic performance. Relevant non-cognitive competences were grouped by using four categories: personality traits (6 competences), motivational factors (3), self-regulatory learning strategies (5) and contextual influences (3).

These results provide a model of competences which can be used to design an assessment protocol composed of evaluation instruments focused on measuring all the relevant competences. Future steps will focus on improving the model by incorporating specific indicators of the importance (weight) of each competence into the model. In addition, additional sources of information will be included to optimize the utility of the model.

References

Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., & Farrell, P. M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine*, *78*(3), 313–321. https://doi.org/10.1097/00001888-200303000-00016

Association of American Medical Colleges. (2018). *The MCAT Essentials for Testing Year 2018*. http://aamcorange.global.ssl.fastly.net/production/media/filer_public/b9/c3/b9c382ef-5746-4da1-9265-ae570bb655e1/mcat_essentials_2016_-_final2.pdf.

Beller, E. M., Glasziou, P. P., Altman, D. G., Hopewell, S., Bastian, H., Chalmers, I., Gøtzsche, P. C., Lasserson, T., Tovey, D., & Group, for the P. for A. (2013). PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts. *PLOS Medicine*, *10*(4), e1001419. https://doi.org/10.1371/journal.pmed.1001419

Carrothers, R. M., Gregory, S. W., & Gallagher, T. J. (2000). Measuring emotional intelligence of medical school applicants. *Academic Medicine, 75*(5), 456–463. https://doi.org/10.1097/00001888-200005000-00016

College Board. (2017). *SAT Suite of Assessments Technical Manual.* https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual.pdf

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research (3rd ed.)*. Los Angeles: SAGE Publications.

Donnon, T., Paolucci, E. O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: A meta-analysis of the published research. *Academic Medicine, 82(*1), 100–106. https://doi.org/10.1097/01.ACM.0000249878.25186.b7

Educational Testing Service. (2015). *GRE guide to the use of scores 2015–2016*. http://www.ets.org/s/gre/pdf/gre_guide.pdf

Ferguson, E., James, D., & Madeley, L. (2002). Factors associated with success in medical school: Systematic review of the literature. *British Medical Journal, 324*(7343), 952–957. https://doi.org/10.1136/bmj.324.7343.952

Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2002). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences, 14*(1), 47–64. https://doi.org/10.1016/j.lindif.2003.08.002

Onwuegbuzie, A., Dickinson, W., Leech, N., & Zoran, A. (2011). Un marco cualitativo para la recolección y análisis de datos en la investigación basada en grupos focales. *Paradigmas, 3*(1), 127–157.

Petrides, K. V, & Furnham, A. (2000). On the dimensional structure of emotional intelligence. *Personality and Individual Differences, 29*(2), 313–320. https://doi.org/10.1016/S0191-8869(99)00195-6

Rabiee, F. (2004). Focus-group interview and data analysis. *Proceedings of the Nutrition Society, 63(*4), 655–660. https://doi.org/10.1079/pns2004399

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: *A systematic review and meta-analysis. Psychological Bulletin, 138*(2), 353–387. https://doi.org/10.1037/a0026838

Sommerfeld, A. (2011). Recasting Non-Cognitive Factors in College Readiness as What They Truly Are: Non-Academic Factors. *Journal of College Admission, 213*(1), 18–22.

Wells, G. A, Shea, B., O'Connel, D. et al. (2009). *The Newcastle-Ottawa scale (NOS) for assessing the quailty of nonrandomised studies in meta-analyses. http://www ohri ca/programs/clinical_epidemiology/oxford htm 2009 Feb 1*. 2009.

West, C., & Sadoski, M. (2011). *Do study strategies predict academic performance in medical school? Medical Education, 45(*7), 696–703. https://doi.org/10.1111/j.1365-2923.2011.03929.x

Zhou, Y., Graham, L., & West, C. (2016). The relationship between study strategies and academic performance. *International journal of medical education, 7*(1), 324–332. https://doi.org/10.5116/ijme.57dc.fe0f

Zimmermann, J., von Davier, A., & Heinimann, H. (2017). Adaptive admissions process for effective and fair graduate admission. *International Journal of Educational Management, 31*(4), 540–558. https://doi.org/10.1108/IJEM-06-2015-0080

Development and psychometric properties of a Barriers Questionnaire for Physical Activity (BQPA) in a representative sample of the Spanish adult population: A preliminary study

Sergio Navas-León[1], Ana Tajadura-Jiménez[2], Patricia Rick[1], Luis Morales Márquez[1], Mercedes Borda Mas[3], Aneesha Singh[4], Nadia Bianchi-Berthouze[4], Milagrosa Sánchez Martín[1,*]

[1]*Departament of Psychology, Universidad Loyola Andalucia, Spain,*

[2]*Departament of Informatics, Universidad Carlos III de Madrid, Spain,*

[3]*Departament of Psychology, Universidad de Sevilla, Spain,*

[4]*UCL Interaction Centre, University College London, UK*

Abstract

Objective: This study aimed to develop a BQPA and evaluate its psychometric properties, covering all the relevant barriers for Physical Activity (PA) reported in the literature. Method/Design: A cross-sectional study was performed in 2019 through a dedicated online panel. A sample of 610 participants was selected using stratified random sampling. We tested the factorial structure of the BQPA through an Exploratory Factor Analysis (EFA) with half of the sample and replicated the structure in the other half through Confirmatory Factor Analysis (CFA). Internal consistency was also analyzed. Results: The proposed BQPA consists of 61 items measured by a 5-point Likert scale, which cover three dimensions of barriers: psychological (42), physical (5) and contextual (14). The first-order three-factor model exhibited a good fit [CFI = 0.948; TLI = 0.945; RMSEA = 0.054 (90% CI = 0.049-0.059); WRMR = 1.159]. Cronbach's Alpha values were satisfactory for each factor: "Personal" (22 items; α = 0.93), "External" (10 items; α = 0.82) and "Predisposition to Physical Activity" (8 items; α = 0.90). Conclusions: The developed BQPA shows adequate psychometric properties. It can detect specific barriers for PA and could be useful to design interventions for promoting PA adapted to each person or to specific groups.

Keywords: Barriers; Factor analysis; Physical inactivity; Psychometric properties; Questionnaire

Funding: This study has been supported by grants PSI2016-79004-R (AEI/FEDER UE) and PID2019-105579RB-I00 (AEI / 10.13039/501100011033).

E-mail: (corresponding author) msanchez@uloyola.es

1. Introduction

The benefits of physical activity (PA) are well known. However, physical inactivity is a worldwide problem. Therefore, the promotion of PA is a public health priority (Beighle & Morrow, 2014).

Theories of behavioral change have emerged to help understand PA and how to promote it (Rhodes, 2017). However, they have been questioned in terms of effectiveness. For example, various meta-analyses for theories-based interventions have reported small to moderate effect sizes ($d = 0.20$; Conn et al., 2011, $d = 0.27$, Rhodes et al., 2017). An alternative approach would be to study the components of these theories, in order to understand which variables affect PA and inform theories to refine them and increase their effectiveness (Baranowski et al., 1998). Under these inductive approaches, various authors have studied the barriers for engaging in PA (e.g., lack of motivation; low self-efficacy). The identification of these barriers is vital for effective interventions (Miles, 2007).

To systematize the study of these barriers, various classifications have been proposed (e.g., internal, interpersonal and environmental, Brinthaupt et al., 2010; internal and external, Hsu et al., 2011). However, despite the fact that psychological variables play a key role in PA (Nigg & Geller, 2012), they have not received comprehensive treatment in the current classifications. On the other hand, numerous psychometric instruments (Self-Report on Barriers to Exercising, Niñerola i Maymí et al., 2006; San Diego Health and Exercise Questionnaire, Rauh et al., 1992; Perceived Barriers Questionnaire, O'Neill & Reid, 1991) have been developed but most have focused on specific populations such as ethnic groups (Jackson et al., 2016), adults with chronic diseases (Desveaux et al., 2016; van Adrichem et al., 2016) or specific sociodemographic characteristics (Cary et al., 2016; Cramp & Bray, 2009) limiting the generalization of these results to the general population.

For all these reasons, the purpose of this study was to develop and evaluate the psychometric properties of a Barriers Questionnaire for Physical Activity (BQPA) in a general representative sample of the Spanish adult population.

2. Method

2.1. Recruitment and Participants

A cross-sectional study was conducted in 2019. The data for this analysis were collected through a dedicated online panel. 610 Spanish adults filled out the questionnaires. The sample was stratified with respect to gender and age range (18-24, 25-34, 35-44, 45-54, 55-64, 65-74 and 75 or older). Eligible participants were required: (a) to be ≥18 years; (b) not to have known medical issues for which PA was not recommended; and (c) to provide informed consent. The study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and approved by the local Ethics Committees.

We controlled the existence of outliers administering the Short International Physical Activity Questionnaire (IPAQ-SF) (Hallal & Vectora, 2004), which resulted in 116 subjects being excluded from the sample. The final sample comprised 494 participants, representative of the Spanish population, whose age ranged between 18 and 65 years (M = 40.34; SD = 13.30). 51.2 % were men (M = 43.25; N = 253) and 48.8 % women (M = 37.27; N = 241). Regarding educational level, 0.4% (N = 2) had not completed basic education, 32% of subjects (N = 158) had reached primary/secondary studies, 17.2% higher studies (N = 85), and 50.4% university studies (N = 249). With respect to employment, 27.3 % (N = 134) were working, 36.5%

(N = 180) were unemployed and 32.4% (N = 160) had retired. Finally, most participants (23.1 %, N = 114) had a gross monthly salary between €1101 and €1800, lower (23.7 %; N = 117) or no income (13.8%; N = 68) while 15.6% had an income of €1801 - €2700 or higher (5.8%; N = 25).

2.2 Instruments

Barriers Questionnaire for Physical Activity (BQPA). The BQPA derives from the findings of our original previous literature review (Rick et al., 2020). The proposed dimensions are based on a framework presented by Singh (2016) (psychological, contextual/personal and physical variables). 38 variables were identified as PA barriers/facilitators which were turned into 61 items: 42 items for the psychological dimension (e.g., "I'm not into exercising"), 14 items for the contextual dimension (e.g., "I don't have equipment for physical activity") and 5 items for the physical dimension (e.g., "I feel pain when exercising"). The response format is a 5-point Likert scale, ranging from 0 (not at all) to 4 (a lot).

2.3. Statistical Analysis

Descriptive statistics were used to describe the characteristics of the sample. The adequacy of the sample for this procedure was measured by Kaiser-Meyer-Olkin (KMO) (Kaiser, 1970) and Barlett's sphericity test (Barlett, 1950). For construct validity, the whole sample was randomly split into two sets for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The number of factors extracted was based on three different criteria: a) dimensions on which the BQPA was based; b) factor loadings of items; c) and Parallel Analysis based on Minimum Rank Factor Analysis (PA-MRFA) with a 95% threshold (Lorenzo-Seva, 2011). We conducted the EFA (N = 248) using the Principal Axis Factoring extraction method with Promax rotation, which allowed correlation between factors (Brown, 2015). The factor structure was further investigated using CFA (N = 246) through the Weighted Least Squares Means and Variance (WLSMV) (Muthen, 1993) estimation method, which is appropriate for skewed items (Brown, 2015). Model fit was evaluated with the following goodness-of-fit indices: Comparative Fit Index (CFI) and Tucker and Lewis Index (TLI) between .90 and .95; Root Mean Square Error of Approximation (RMSEA) less than .08; non-significant Test of Approximate Fit of RMSEA; and Weighted Root Mean Square Residual (WRMR) around 1.0 (Hooper et al., 2008; Hu & Bentler, 1999). Cronbach's alpha and coefficient Omega were used to analyzed internal consistency (Cronbach, 1951; McDonald, 1999). FACTOR v.10.05.02 software was used for the PA (Lorenzo-Seva & Ferrando, 2006), MPLUS 8.4 for EFA and CFA (Muthen & Muthen , 2017), and IBM SPSS Statistics version 25.0 for the remaining analyses (IBM, 2017). All statistical procedures adopted a significance level ≤ .05.

3. Results

3.1 Evidence of Construct Validity

The KMO verified the adequacy of the sample for the analysis (KMO = .89), and Bartlett's Test of Sphericity was significant ($p < .01$). Results from Mardia's multivariate normality test (1970) showed non-normality (Mardia's = 46.692; $p < 0.05$). The PA-MRFA with a 95% threshold suggested the retention of 3 factors. Prior to the EFA, items with standard deviation (SD) below <1.00 were removed (items: 16, 39, 44, 56, 59); 56 items remained. Next, we ran a principal axis factoring (PFA) with Promax Rotation. Items with cross-loadings below <0.20

were removed. Taking into account the aforementioned criteria, a three-factor model was more adequate (RMSEA = 0.044) than a two-factor model (RMSEA = 0.050) or four-factor model (RMSEA = 0.042). For the three-factor model, the following items were eliminated one by one: 8, 13, 14, 18, 21, 22, 24, 31, 34, 38, 40, 41, 49, 50, 51, 54 (40 items remained). For the CFA, we tested a first-order three-factor model. This showed an adequate data-fit [CFI = 0.948; TLI = 0.945; RMSEA = 0.054 (90% CI = 0.049-0.059); WRMR = 1.159]. Inter-factor correlations (F1-F2 = 0.78; F2-F3 = 0.64; F1-F3 = 0.69) showed moderate discriminant validity. Standardized factor loadings ranged from .89 to .49 ($p < .001$). Residual variances ranged from .75 to .19 and the item R-square ranged from .80 to .29. No post-hoc modifications were conducted. F1 was mainly made up of Personality, Mental Health, Physical Status and Affect barriers (Personal), F2 by Infrastructure-Daily Life Demands (External) and F3 by Motivation-Behavior (Predisposition to Physical Activity).

3.2 Reliability

Table 1. Reliability through different dimensions. Note: Ω (CR) = composite reliability; α = Cronbach's Alpha; AEV = Average Extracted Variance

Dimensions	Ω (CR)	α	AEV
Personal	0.96	0.93	0.52
External	0.88	0.82	0.43
Predisposition to Physical Activity	0.93	0.90	0.65
Total	0.97	0.95	0.52

Cronbach's Alpha and Omega values were satisfactory. Table 1 summarizes the main results for each dimension.

4. Conclusions

The objective of this study was to develop a BQPA and to evaluate its psychometric properties. The BQPA showed moderate psychometric properties in terms of validity and reliability for the population under study. The BQPA could be useful for interventions promoting PA and here we recommend it for further investigations. However, several limitations exist. First, in terms of reliability, direct measures like the use of speedometers or accelerometers are recommended when measuring PA (Ahmad et al., 2018). In our case, like in other studies (Gobbi et al., 2012), the large sample under study prevented this recommendation. Second, polychoric-based Parallel Analysis is recommended with ordinal data (Dominguez-Lara, 2014). However, we experienced convergence problems that prevented its application (Lorenzo-Seva & Ferrando, 2020). To solve this problem, we used Pearson-based PA, which is recommended under these circumstances (Timmerman & Lorenzo-Seva, 2011). Finally, the use of WLSMV yielded moderate overestimation of the interfactor correlations when the sample was relatively small or moderately non-normal (e.g., N = 200) (Li, 2016), similar to the findings of previous studies (Wegmann et al., 2011). Future research should replicate these findings in broader samples.

References

Ahmad, M. H., Salleh, R., Mohamad Nor, N. S., Baharuddin, A., Rodzlan Hasani, W. S., Omar, A., Jamil, A. T., Appukutty, M., Wan Muda, W. A. M., & Aris, T. (2018). Comparison between self-reported physical activity (IPAQ-SF) and pedometer among overweight and obese women in the MyBFF@home study. *BMC Women's Health*, *18*(Suppl 1). https://doi.org/10.1186/s12905-018-0599-8

Baranowski, T., Anderson, C., & Carmack, C. (1998). Mediating variable framework in physical activity interventions: How are we doing? How might we do better? *American Journal of Preventive Medicine*, *15*(4), 266–297. https://doi.org/10.1016/S0749-3797(98)00080-4

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, 2(1), 109-133. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Beighle, A., & Morrow, J. R. (2014). Promoting Physical Activity: Addressing Barriers and Moving Forward. *Journal of Physical Education, Recreation & Dance*, *85*(7), 23–26. https://doi.org/10.1080/07303084.2014.937190

Brinthaupt, T. M., Kang, M., & Anshel, M. H. (2010). A delivery model for overcoming psycho-behavioral barriers to exercise. *Psychology of Sport and Exercise*, *11*(4), 259–266. https://doi.org/10.1016/j.psychsport.2010.03.003

Brown, T. A. (2015).Confirmatory factor analysis for applied research(2nded.). New York, NY: The Guilford Press.

Cary, M. A., Brittain, D. R., Dinger, M. K., Ford, M. L., Cain, M., & Sharp, T. A. (2016). Barriers to Physical Activity Among Gay Men. *American Journal of Men's Health*, *10*(5), 408–417. https://doi.org/10.1177/1557988315569297

Conn, V. S., Hafdahl, A. R., & Mehr, D. R. (2011). Interventions to increase physical activity among healthy adults: meta-analysis of outcomes. *American Journal of Public Health*, *101*(4), 751–758. https://doi.org/10.2105/AJPH.2010.194381

Cramp, A. G., & Bray, S. R. (2009). A prospective examination of exercise and barrier self-efficacy to engage in leisure-time physical activity during pregnancy. *Annals of Behavioral Medicine*, *37*(3), 325–334. https://doi.org/10.1007/s12160-009-9102-y

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. Psychometrika *16*(3)**:** 297–334. https://doi.org/10.1007/BF02310555

Desveaux, L., Goldstein, R., Mathur, S., & Brooks, D. (2016). Barriers to physical activity following rehabilitation: Perspectives of older adults with chronic disease. *Journal of Aging and Physical Activity*, *24*(2), 223–233. https://doi.org/10.1123/japa.2015-0018

Dominguez-Lara, S. (2014). ¿Matrices Policóricas/Tetracóricas o Matrices Pearson? Un estudio metodológico. *Revista Argentina de Ciencias del Comportamiento (RACC) 6*(1), 39–48). https://doi.org/10.30882/1852.4206.v6.n1.6357

Gobbi, S., Sebastião, E., Papini, C. B., Nakamura, P. M., Valdanha Netto, A., Gobbi, L. T. B., & Kokubun, E. (2012). Physical inactivity and related barriers: A study in a community dwelling of older brazilians. *Journal of Aging Research*, *2012*. https://doi.org/10.1155/2012/685190

Hallal, P.D & Victora, C.G. (2004). Reliability and validity of the International Physical Activity Questionnaire (IPAQ). *Medicine and Science in sports and exercise*. *36*(3), 556. https://doi.org/10.1249/01.mss.0000117161.66394.07

Hsu, H. T., Dodd, M. J., Guo, S. E., Lee, K. A., Hwang, S. L., & Lai, Y. H. (2011). Predictors of exercise frequency in breast cancer survivors in Taiwan. *Journal of Clinical Nursing*, *20*(13–14), 1923–1935. https://doi.org/10.1111/j.1365-2702.2010.03690.x

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, 6, 53–60. Technological University Dublin Library Services. https://arrow.tudublin.ie/buschmanart/2/

Hu, L.T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

IBM Corp. (2017). IBM SPSS Statistics for Windows. Armonk, NY: IBM Corp.

Jackson, H., Yates, B. C., Blanchard, S., Zimmerman, L. M., Hudson, D., & Pozehl, B. (2016). Behavior-Specific Influences for Physical Activity Among African American Women. *Western Journal of Nursing Research*, *38*(8), 992–1011. https://doi.org/10.1177/0193945916640724

Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/10.1007/BF02291817

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Lorenzo-Seva, U. (2011). Horn's parallel analysis for selecting the number of dimensions in correspondence analysis. *Methodology European Journal of Research Methods for the Behavioral and Social Sciences 7*(3):96-102. https://doi.org/10.1027/1614-2241%2Fa000027

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91. https://doi.org/10.3758/bf03192753

Lorenzo-Seva, U., & Ferrando, P. J. (2020). Not Positive Definite Correlation Matrices in Exploratory Item Factor Analysis: Causes, Consequences and a Proposed Solution. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-10. https://doi.org/10.1080/10705511.2020.1735393

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530. https://doi.org/10.1093/biomet/57.3.519

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates.

Miles, L. (2007). Physical activity and health. *Nutrition Bulletin*, *32*(4), 314–363. https://doi.org/10.1111/j.1467-3010.2007.00668.x

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 205–243). Newbury Park, CA: Sage.

Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén. Statmodel. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Niñerola i Maymí, J., Capdevila-Ortís L., & Pintanel-Bassets, M. (2006). Barreras percibidas y actividad física: el autoinforme de barreras para la práctica de ejercicio físico. *Revista de*

Psicología del Deporte, *15*(1), 53-69. Dipòsit Digital de la Universitat de Barcelona: https://ddd.uab.cat/pub/revpsidep/19885636v15n1/19885636v15n1p53.pdf

Nigg, C. R., & Geller, K. S. (2012). Theoretical approaches to physical activity intervention. In The Oxford handbook of exercise psychology (p. 252). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195394313.013.0014

O'Neill, K., & Reid, G. R. E. G. (1991). Perceived barriers to physical activity by older adults. *Canadian journal of public health*, *82*(6), 392–396. https://pubmed.ncbi.nlm.nih.gov/1790502/

Rauh, M. J., Hovell, M. F., Hofstetter, C. R., Sallis, J. F., & Gleghorn, A. (1992). Reliability and validity of self-reported physical activity in Latinos. *International Journal of Epidemiology*, *21*(5), 966–971. https://doi.org/10.1093/ije/21.5.966

Rhodes, R. E. (2017). The Evolving Understanding of Physical Activity Behavior. In *Advances in Motivation Science*, *4*, 171-205). Elsevier Ltd. https://doi.org/10.1016/bs.adms.2016.11.001

Rhodes, R. E., Janssen, I., Bredin, S. S. D., Warburton, D. E. R., & Bauman, A. (2017). Physical activity: Health impact, prevalence, correlates and interventions. *Psychology and Health*, *32*(8), 942–975. https://doi.org/10.1080/08870446.2017.1325486

Rick P, Sánchez-Martín M, Singh A, Navas-León S, Bordá-Más M, Bianchi-Berthouze N, Tajadura-Jiménez A. (2020). Embedding Psychological Factors in Technology Design to Improve Adherence to Physical Activity: Literature Review and Survey. JMIR Preprints. https://preprints.jmir.org/preprint/19663

Singh A. (2016). *Staying active despite pain: Investigating feedback mechanisms to support physical activity in people with chronic musculoskeletal pain*. [Doctoral dissertation, University College London. UCL Discovery. https://discovery.ucl.ac.uk/id/eprint/1532144/

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353

van Adrichem, E. J., van de Zande, S. C., Dekker, R., Verschuuren, E. A., Dijkstra, P. U., & van der Schans, C. P. (2016). Perceived barriers to and facilitators of physical activity in recipients of solid organ transplantation, a qualitative study. *PLoS One*, *11*(9), e0162725. https://doi.org/10.1371/journal.pone.0162725

Wegmann, K. M., Thompson, A. M., & Bowen, N. K. (2011). A confirmatory factor analysis of home environment and home social behavior data from the elementary school success profile for families. *Social Work Research*, *35*(2), 117–127. https://doi.org/10.1093/swr/35.2.117

How to and how not to impute incomplete count data

Kristian Kleinke[1], Jost Reinecke[2]

[1]*Institute of Psychology, University of Siegen, Germany,*

[2]*Faculty of Sociology, University of Bielefeld, Germany*

Abstract

Missing data pose a threat to the validity of statistical inferences, when they are numerous, not missing completely at random, and when they are handled in an inadequate way. Multiple imputation is a state-of-the-art method to handle the missing data problem and produces unbiased inferences, when (distributional) assumptions are at least approximately met. Count data are non-negative integer values, and often skewed. Most MI software does not support count models or supports only basic count models. Van Buuren (2018) therefore recommends the following strategies to impute count data: predictive mean matching (pmm), ordered categorical regression, (zero-inflated) Poisson regression, and (zero-inflated) negative binomial regression. In the present paper, we evaluate these recommendations by means of Monte Carlo simulation. Based on our findings, we discourage the use of proxy strategies with ill-fitting (distributional) assumptions.

Keywords: missing data; multiple imputation; count data.

E-mail: kristian.kleinke@uni-siegen.de

1. Introduction

Count data are non-negative integer values and give the frequency of occurrence of a certain event or behavior within a given timespan, for example the number of days a patient spends in hospital or the count of delinquent behaviors a person has committed in a year. Count data are often not normally distributed but skewed, and usually require special analysis and imputation techniques. Yet, most of the currently available multiple imputation packages are very limited with regard to count data imputation models. A basic Poisson imputation model is for example available in "IVEware" (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). "Ice" for Stata (Royston, 2009) also supports negative binomial regression-based imputation. Zero-inflation or multilevel count models are typically not supported. Proxy strategies to impute various kinds of count data include a) ignoring the fact that count data are (often skewed) non-negative integer values, and using standard procedures, for example, based on OLS regression under the assumption of normal homoscedastic errors, b) applying normalizing transformation, such as a log or square root transformation, followed by normal model multiple imputation and back transformation to the original scale afterwards, or c) treating the data as ordered categorical data and using an ordinal logistic regression imputation model. The transformation-imputation-back transformation approach was implemented in Schafer's "norm" for Windows software from 1999, for example, creating imputations under the assumption of multivariate normality. Back then, norm was one of the very few user-friendly software solutions generally available to applied researchers. Transforming non-normal variables before the imputation to make the normality assumption of the imputation model more plausible was one of the few options applied researchers had at that time to handle missing data in non-normally distributed data. Today, missing data researchers (e.g., von Hippel, 2013) usually discourage the use of transformations to make the normality assumption of an imputation model more plausible. Usually, better alternatives are available. Van Buuren (2018, Chap. 3.7.1) recommends the following strategies to impute count data: a) using a semiparametric k nearest neighbor imputation approach (i.e., predictive mean matching, pmm), b) using ordered categorical regression, c) using a (zero-inflated) Poisson regression model, or d) creating the imputations based on a (zero-inflated) negative binomial regression model. In this paper, we would like to discuss these recommendations in the light of current missing data research and aim to corroborate our points by a Monte Carlo simulation. Firstly, predictive mean matching is usually a good all-round method that can be recommended for many scenarios including count data that are not too severely skewed. Kleinke (2017), for example, who has simulated incomplete Poisson distributed data, has shown that pmm yielded acceptable results when skewness was only mild to moderate, when not too many data had to be imputed, and when the sample size was sufficiently large. However, inferences were biased when count data were rather heavily skewed and the missing data percentage was substantial. We thus might assume that pmm is not a good imputation strategy for zero-inflated count data (i.e. count data with a very large percentage of zero counts), which are typically quite heavily skewed.

Secondly, we are not aware of systematic simulations that have tested how appropriate ordered categorical regression-based imputation is for zero-inflated Poisson or zero-inflated negative binomial data. Generally speaking, (ordered) categorical imputation could be feasible when the number of categories is not too large (otherwise, one might run into estimation problems / empty cell problems): Van Buuren and Groothuis-Oudshoorn (2011) mention

that the imputation of categorical variables with more than 20 categories is often problematic. Therefore, if the range of the count variable lies between 0 and 20, using function mice.impute.polr to create the imputations might be a suitable option. In our simulation, we wanted to test whether this was also an appropriate strategy for zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) data.

Thirdly, Kleinke and Reinecke (2013) have already demonstrated that model-based MI, which creates multiple imputations by a correctly specified parametric model (here based on a ZIP or ZINB imputation model) produced unbiased statistical inferences when the data were in fact zero-inflated Poisson or zero-inflated negative binomial, respectively. However, we did not compare ZIP and ZINB imputation against predictive mean matching or ordered categorical regression-based multiple imputation.

In the present study, we explore, a) whether pmm can be safely used for the imputation of ZIP and ZINB data, b) the appropriateness of ordered categorical imputation in these scenarios, and c) a comparison of the results of these strategies against results based on correctly specified ZIP and ZINB imputation models with fitting distributional assumptions. To these ends, we re-analyzed the data from Kleinke and Reinecke (2013).

2. Method

Details about the simulation set-up are given in Kleinke and Reinecke (2013). Simulated data sets contained four variables. y was the dependent incomplete zero-inflated count variable which, depending on the respective condition, was either ZIP or ZINB. Data sets also included three continuous predictors: x_1, x_2, and z_1, with x_1 and x_2 being the covariates in the count part of the model, and z_1 the covariate in the zero-inflation part. Population parameters were set to $\beta_0 = 1$, $\beta_1 = .3$, $\beta_2 = .3$, $\gamma_0 = 0$, $\gamma_1 = 2$, with β being the parameters in the count part and γ the parameter in the zero part of the model. In the ZINB conditions, the dispersion parameter was set to 1. For both the ZIP and ZINB conditions, 1000 data sets with sample sizes N = 500; 1000; 10000 were simulated, respectively. MAR missingness in y was introduced as outlined in Kleinke and Reinecke (2013). Missing data were imputed using functions mice.impute.pmm (predictive mean matching) and mice.impute.polr (ordered categorical regression). Repeated data analysis and pooling of results was performed as described in Kleinke and Reinecke (2013).

3. Results

3.1. Predictive mean matching

The predictive mean matching results are shown in Tables 1 and 2. The tables display a) the average estimate of the respective parameter across the 1000 replications , b) their standard deviation, c) bias, which is defined as the average absolute difference between the simulated true parameter and its estimate across the 1000 replications, and d) 95% confidence interval coverage rates, i.e., the percentage of intervals that include the true parameter. Obviously, the average estimate should be close to the respective true parameter. Furthermore, the standard deviation of the estimates across the replications should be small (which in combination with an accurate point estimate reflects a consistently good estimation across the replications). Finally, coverage rates should be close to 95%. Schafer and Graham (2002) deem values below 90% as serious undercoverage.

Table 1. Performance of predictive mean matching MI when data are ZIP.

N	Parameter	EST	SD	Bias	CR
500	β_0	0.944	0.062	0.056	87.4
500	β_1	0.240	0.052	0.060	79.2
500	β_2	0.249	0.051	0.051	86.7
500	γ_0	−0.183	0.172	0.183	86.8
500	γ_1	1.719	0.235	0.281	79.1
1000	β_0	0.947	0.042	0.053	80.9
1000	β_1	0.243	0.037	0.057	67.2
1000	β_2	0.252	0.036	0.048	77.8
1000	γ_0	−0.185	0.121	0.185	74.6
1000	γ_1	1.707	0.165	0.293	63.3
10000	β_0	0.946	0.014	0.054	2.8
10000	β_1	0.244	0.011	0.056	0.7
10000	β_2	0.250	0.011	0.050	2.3
10000	γ_0	−0.191	0.038	0.191	0.3
10000	γ_1	1.699	0.052	0.302	0.2

N.B. N is the sample size. EST is the average parameter estimate across the replications, SD is the corresponding standard deviation. Bias is the difference between EST and the true population parameter, and CR is the percentage of 95% confidence intervals that include the true parameter.

When we first look at the results of the ZIP conditions, we see that point estimates $\hat{Q}$ are very similar, regardless of sample size. However, both parameters in the zero model and in the count model are biased to some extent. Furthermore, it should be noted that all coverage rates are below the acceptable 90% threshold and decrease dramatically with increasing sample size. This is because standard error estimates depend on sample size. An increasing sample size leads to a smaller standard error. Confidence intervals therefore become narrower, and even smaller biases can result in "significant" undercoverage.

Let us turn to the ZINB conditions. Here, biases were especially noticeable in the zero part of the model. Though parameters of the count part of the model were also underestimated (especially β_1), corresponding coverage rates were acceptably large, when $N = 500$ or $N = 1000$. However, in the large sample size condition, coverage fell below 90% for all parameters and, in the worst case, dropped to 3% for parameter γ_1.

Table 2. Performance of predictive mean matching MI when data are ZINB.

N	Parameter	EST	SD	Bias	CR
500	β_0	0.938	0.121	0.062	92.6
500	β_1	0.265	0.096	0.035	93.8

N	Parameter	EST	SD	Bias	CR
500	β_2	0.271	0.099	0.029	93.9
500	γ_0	−0.195	0.300	0.195	96.0
500	γ_1	1.729	0.326	0.271	84.9
1000	β_0	0.939	0.085	0.061	91.5
1000	β_1	0.266	0.068	0.034	92.2
1000	β_2	0.270	0.067	0.030	95.2
1000	γ_0	−0.194	0.211	0.194	92.0
1000	γ_1	1.709	0.220	0.291	77.7
10000	β_0	0.939	0.026	0.061	42.1
10000	β_1	0.269	0.022	0.031	69.8
10000	β_2	0.271	0.021	0.029	75.8
10000	γ_0	−0.196	0.065	0.196	20.6
10000	γ_1	1.694	0.066	0.306	3.0

N.B. N is the sample size. EST is the average parameter estimate across the replications, SD is the corresponding standard deviation. Bias is the difference between EST and the true population parameter, and CR is the percentage of 95% confidence intervals that include the true parameter.

3.2. Polytomous regression

The results of polytomous regression are shown in Tables 3 and 4. Coefficients in the count part of the model were only slightly biased. The coefficient in the zero-inflation part was more heavily biased. Coverage rates of most model parameters were usually acceptable, unless sample size was very large.

In this scenario, we conclude that where data were either zero-inflated Poisson or zero-inflated negative binomial and thus severely skewed, predictive mean matching produced biased statistical inferences. Polytomous regression in comparison yielded better results especially regarding the count part of the model. When we compare results to the ones reported in Kleinke and Reinecke (2013), we see that using an appropriate parametric imputation model with fitting distributional assumptions is clearly the better choice.

Table 3. Performance of polytomous regression MI when data are ZIP.

N	Parameter	EST	SD	Bias	CR
500	β_0	0.997	0.055	0.003	95.0
500	β_1	0.284	0.046	0.016	94.6
500	β_2	0.276	0.046	0.024	93.3
500	γ_0	-0.024	0.162	0.024	94.4
500	γ_1	1.943	0.252	0.057	93.9
1000	β_0	0.997	0.037	0.003	95.0

N	Parameter	EST	SD	Bias	CR
1000	β_1	0.284	0.034	0.016	92.5
1000	β_2	0.281	0.032	0.019	92.6
1000	γ_0	-0.016	0.115	0.016	94.4
1000	γ_1	1.932	0.175	0.068	91.5
10000	β_0	0.996	0.013	0.004	92.4
10000	β_1	0.283	0.010	0.017	63.3
10000	β_2	0.284	0.011	0.016	70.0
10000	γ_0	-0.004	0.036	0.004	94.4
10000	γ_1	1.921	0.053	0.079	70.7

N.B. N is the sample size. EST is the average parameter estimate across the replications, SD is the corresponding standard deviation. Bias is the difference between EST and the true population parameter, and CR is the percentage of 95% confidence intervals that include the true parameter.

Table 4. Performance of polytomous regression MI when data are ZINB.

N	Parameter	EST	SD	Bias	CR
500	β_0	1.038	0.116	-0.038	90.5
500	β_1	0.299	0.093	0.001	94.7
500	β_2	0.276	0.092	0.024	93.8
500	γ_0	0.079	0.261	-0.079	90.5
500	γ_1	1.781	0.289	0.219	85.8
1000	β_0	1.030	0.083	-0.030	91.6
1000	β_1	0.296	0.066	0.004	94.2
1000	β_2	0.281	0.064	0.019	93.6
1000	γ_0	0.080	0.184	-0.080	91.3
1000	γ_1	1.773	0.201	0.227	82.4
10000	β_0	1.022	0.026	-0.022	85.8
10000	β_1	0.292	0.021	0.008	92.3
10000	β_2	0.291	0.021	0.009	93.1
10000	γ_0	0.077	0.056	-0.077	71.6
10000	γ_1	1.782	0.062	0.218	16.4

N.B. N is the sample size. EST is the average parameter estimate across the replications, SD is the corresponding standard deviation. Bias is the difference between EST and the true population parameter, and CR is the percentage of 95% confidence intervals that include the true parameter.

4. Conclusions

Kleinke and Reinecke (2013) used an imputation model with fitting distributional assumptions (i.e. ZIP imputation for ZIP data and ZINB imputation for ZINB data) and obtained unbiased statistical inferences. Here, we used the same data and imputed them using proxy methods like predictive mean matching and (ordered) polytomous regression. These strategies, however, yielded biased statistical inferences in at least some scenarios.

The results of this simulation once again stress the need to find an appropriate imputation model that has a sufficiently good fit for the data at hand. If the imputation model is overly implausible or mis-specified, results will be biased.

It should be noted that this study was based on simulated data, which were ZIP and ZINB distributed. A ZIP or ZINB model will never have a perfect fit to 'real' empirical data. Applied researchers need to bear in mind that the quality of imputations will depend on how well one's empirical data can be modeled by mathematically convenient models (like in this case a ZIP or ZINB model). Usually, the worse the model fit between empirical data and the assumed data generating process is, the less "plausible" model-based imputations will be and the more bias is to be expected. Future research could extend the idea of predictive mean matching to various kinds of count models. Instead of using a normal heteroscedastic linear regression model to match the donor to the done (like the standard pmm function in mice), a count data regression model could be adopted. This could enhance the matching process and preserve many properties of the empirical data at hand. This might be a useful addition to the methodological toolbox for situations where none of the standard count models has a perfect fit for the empirical data at hand.

References

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions – a systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, *42*(4), 371–404.

Kleinke, K., & Reinecke, J. (2013). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, *67*(3), 311–336.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. W. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, *27* (1), 85–95.

Royston, P. (2009). Multiple Imputation of Missing Values: Further Update of Ice, with an Emphasis on Categorical Variables. *Stata Journal*, *9* (3), 466–77.

Schafer, J. L. (1999). NORM users' guide (Version 2). University Park: The Methodology Center, Penn State. https://scholarsphere.psu.edu/collections/v41687m23q

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall / CRC.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, *42*(1), 105–138.

Differences in longitudinal trajectories between groups - The Multi-Group Latent Growth Components approach

Benedikt Langenberg[1], Axel Mayer[1]

[1]*RWTH Aachen University, Germany*

Abstract

Purpose: In this article, we propose a multi-group approach for analyzing complex nonlinear longitudinal trajectories. *Method*: The approach is based on the latent growth components approach (LGCA) that offers a flexible framework for defining growth components and extends the same for the use with multiple groups. The approach benefits from known advantages of the LGCA and adds more capabilities from the multi-group framework, that is, (1) it can flexibly include complex nonlinear growth components, (2) incorporate a measurement model for the latent state variables and latent covariates, (3) it can model differences in growth components based on categorical covariates, and (4) treat covariates and group weights as fixed or stochastic. *Results and conclusions*: We demonstrate the approach using data from the Health and Retirement Study that includes individuals diagnosed with cancer. We analyze trajectories in depressive symptoms before and after the cancer diagnosis with respect to a subset of categorical covariates (i.e., groups). We further present the open-source R package semnova that implements the proposed approach and makes it conveniently accessible for applied researchers.

Keywords: Latent growth models; Longitudinal research; Average effects; Multi-group analysis; Latent growth components approach.

Funding: This work was supported by the German Research Foundation under Grant MA 7702/1-1. The data that is used in this article was funded by the American National Institute on Aging under Grant NIA U01AG009740.

Supplementary material: The software code that supports the findings of this study is openly available in github at https://github.com/langenberg/LangenbergMayer2020_EAM2020.

E-mail: benedikt.langenberg@uni-bielefeld.de

1. Introduction

There is strong demand in research for the analysis of complex trajectories of change over time. For instance, the study of change in patients' well-being measured multiple times before and after a cancer diagnosis is of great interest in medical research. Several have been proposed to address this challenge including *latent growth curve models* (McArdle, 1988; McArdle & Epstein, 1987; Meredith, 1993) and *latent change score models* (McArdle, 2009; McArdle & Hamagami, 2001; Raykov, 1999; Steyer et al., 1997). Latent growth curve models oftentimes aim at modeling polynomial trajectories of change, and latent change score models focus on modeling the change between two neighboring measurement occasions. Researchers, however, may have very particular hypotheses about the shape of change in patients' well-being. The *latent growth components model* (LGCA, Mayer et al., 2012) is a generalization of the aforementioned models and satisfies this need offering the researcher a convenient way to model complex trajectories of change.

All of the aforementioned models have in common that they were originally formulated as single-group models. That is, groups and categorical covariates are included in the model as dummy-coded predictors (e.g., Mayer et al., 2013). In this article, we extended the LGCA for use with multiple groups to model effects of categorical covariates (*multi-group latent growth components approach*, MG-LGCA). For this purpose, we built on causality theory to estimate the average of the effects of categorical covariates which is conceptually similar to the *EffectLiteR approach* (Mayer et al., 2016). We demonstrated the MG-LGCA by means of data from the Health and Retirement Study containing physical and emotional depressive symptoms of patients before and after a cancer diagnosis as well as multiple categorical and continuous predictors. We employed the same model and the same data that was used by Mayer et al. (2013) but used a multi-group model instead of dummy-coded categorical covariates. We conclude this article by briefly discussing findings from the analysis.

2. Method

2.1. Motivating Example

For this article, we used the same data from the Health and Retirement Study (N = 2,798, University of Michigan, 2020) that was used by Mayer et al. (2013). The data contained several items measuring depressive symptoms that were divided into two parcels of physical symptoms (included items: *Felt depressed, Effort*, *Sleep*, *Not get going*) and emotional symptoms (included items: *Happy, Lonely, Enjoy life, Sad*). For each of the variables, two measurements before the cancer diagnosis, two measurements after the diagnosis as well as one measurement in the year of the diagnosis were selected (more occasions are available, see, e.g., Infurna et al., 2013). The data consequently included five measurement occasions that were two years apart from each other. For illustration, the patients' gender (G, female vs. male) and marital status (M, married or partnered vs. not married or partnered) were used as categorical predictors.

2.2. The Single-Group Latent Growth Components Model

We use the multi-state model from Mayer et al. (2013) to build the MG-LGCA. The LGCA enables the researcher to define custom contrasts (i.e., growth components) by specifying a contrast matrix $\mathbf{C}$ that transforms the latent state variables $\boldsymbol{\eta}$ into $\boldsymbol{\pi}$:

$$\pi = \mathbf{C}\eta$$

This transformation cannot directly be implemented in SEM because the relevant part of the structural model takes the form $\boldsymbol{\eta} = \mathbf{B}^*\boldsymbol{\pi}$. To obtain the $\mathbf{B}^*$ matrix, the $\mathbf{C}$ matrix must be inverted:

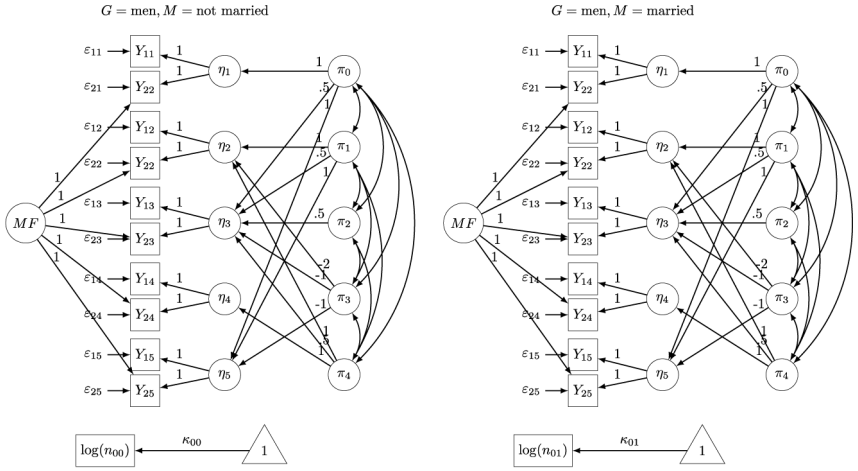
$$\eta = \mathrm{C}^{-1}\pi = \mathrm{B}^*\pi.$$

The $\mathbf{B}^*$ matrix is then incorporated into the matrix of structural coefficients of the SEM containing regressions between the latent variables (see Mayer et al., 2012, for details). Mayer et al. (2013) used this approach to formulate five growth components of change in patients' well-being before and after a cancer diagnosis. $\boldsymbol{\eta}$ corresponded to the latent state variables at the five measurement occasions measured by two items each (i.e., the two parcels). $\boldsymbol{\pi}$, on the other hand, corresponded to the growth components. The five components of interest were: (1) initial level π_0 at the first measurement occasion; (2) linear change component π_1; (3) reaction component π_2 defined as the difference between the average of the two measurement occasions before the diagnosis and the measurement in the year of the diagnosis; (4) adaptation component π_3 defined as the difference between the two measurement occasions before the diagnosis and the two measurement occasions after the diagnosis; (5) post-diagnosis level π_4 defined as depressive symptoms at the first occasion after the diagnosis. The contrast matrix $\mathbf{C}$ and the corresponding inverse $\mathbf{B}^*$ are given by:

$$\mathbf{C} = \begin{array}{c} \\ \text{growth components} \end{array} \begin{array}{c} \text{latent state variables} \\ \begin{array}{ccccc} \eta_1 & \eta_2 & \eta_3 & \eta_3 & \eta_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & -1 & 0 & 1 & 2 \\ -1 & -1 & 2 & 0 & 0 \\ -1 & -1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{array}{l} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{array},$$

$$\mathbf{B}^* = \mathbf{C}^{-1} = \begin{array}{c} \\ \text{latent state variables} \end{array} \begin{array}{c} \text{growth components} \\ \begin{array}{ccccc} \pi_0 & \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 & 1 \\ 0.5 & 0.5 & 0.5 & -1 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & 0 \end{pmatrix} \end{array} \begin{array}{l} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{array}.$$

Mayer et al. (2013) used a τ-equivalent measurement model for the two parcels at each measurement occasion fixing the intercepts of the parcels to zero and the loadings to one. An additional method factor was used to account for method effects of the parcels. For identification purposes, intercepts and residual (co-)variances of the $\boldsymbol{\eta}$ variables were fixed to zero while means and covariances of $\boldsymbol{\pi}$ were freely estimated. The model fit reported by Mayer et al. (2013) was $\chi^2(28)=87.979$, $p < .001$, *CFI* = 0.989, *RMSEA* = 0.027, *SRMR* = 0.021.



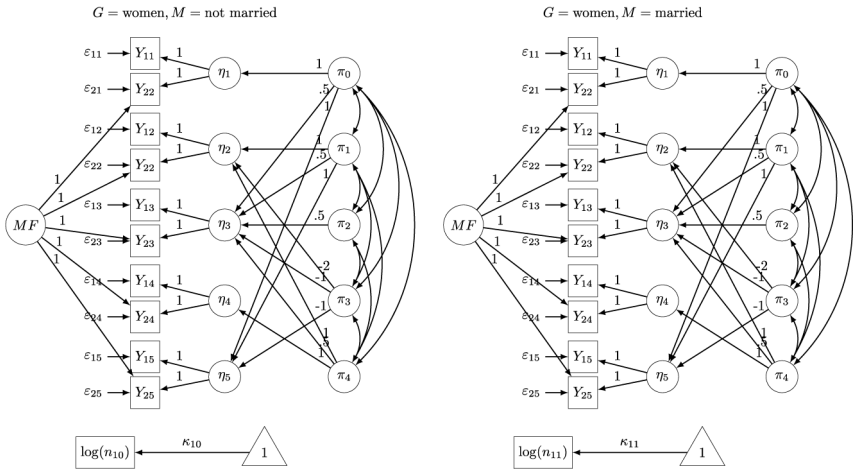


Figure 1. Complete model with latent state variables (η_i), growth components (π_j) and method factor (MF) separated by the four permutations of the categorical predictors (G, M). For the sake of readability, the covariances between the growth components π_j and the method factor have been omitted.

2.3. The Multi-Group Latent Growth Components Approach

In the MG-LGCM, separate models for every possible permutation of the categorical covariates were simultaneously estimated. The measurement model was invariant across the groups as well as the structural coefficients regressing the dependent variables $\boldsymbol{\eta}$ onto the growth components $\boldsymbol{\pi}$ and the means of the method factor. In this example, only an intercept was estimated for every growth component $\boldsymbol{\pi}$:

$$E(\pi_j \mid G = \text{g}, M = \text{m}) = \alpha_{j\text{gm}}.$$

$\alpha_{j\text{gm}}$ represented the mean of π_j in the corresponding group. Based on the parameter estimates of the separate models, a regression was formulated for each of the $\boldsymbol{\pi}$ variables onto the categorical covariates:

$$E(\pi_j \mid G = \text{g}, M = \text{m}) = \beta_{j00} + \beta_{j1\_} \cdot I_{G = \text{women}} + \beta_{j\text{-}1} \cdot I_{M = \text{married}} + \beta_{j11} \cdot I_{G = \text{women}} \cdot I_{M = \text{married}}.$$

$I_{G = \text{women}}$ was an indicator variable that equaled one of gender equals women. $I_{M = \text{married}}$ was an indicator variable that equaled one of marital status equals married. This regression was very similar to a single-group regression with dummy–coded categorical covariates and coefficients to be interpreted in the same way. β_{j00} was the mean of π_j in the group of men that were not married which served as a reference group. $\beta_{j1\_}$ represented the increase in π_j for unmarried women compared to unmarried men. $\beta_{j\text{-}1}$ was the increase in π_j for married men compared to unmarried men. β_{j11} represented the interaction of gender and marital status. We further defined the average effects of the categorical covariates using gender as an example by:

$$\text{AVE}_{G;\pi_j} = E[E(\pi_j \mid G = \text{women}, M) - E(\pi_j \mid G = \text{men}, M)]$$

The average effects of the categorical covariates were defined as the average of the group-specific effects weighted by the group probabilities. For the average effect of gender, these group probabilities corresponded to the unconditional distribution (i.e., unconditional probability) of the categorical covariate marital status (M). The group probabilities were also estimated from the

data and treated as stochastic which is very similar to the EffectLiteR approach (Mayer et al., 2016; Mayer & Thoemmes, 2019).

3. Results

With the regression for each of the growth components , it is now possible to calculate the (conditional) expectations and average (conditional) effects of interest. For this demonstration, we focused on the average effects of the categorical covariates and the unconditional expectation of the π_j variables. We estimated the model using the SEM R software package lavaan (Rosseel, 2012) with full maximum likelihood to account for missing values and a robust estimator. Table 1 contains the estimated regression coefficients $\boldsymbol{\beta}$, the average effects of gender and marital status and the unconditional expectation for each of the $\boldsymbol{\pi}$ variables of the MG-LGCA. The model fit is $\chi^2(115) = 232.277$, $p < .001$, $CFI = 0.981$, $RMSEA = 0.042$, $SRMR = 0.036$. Although, the χ^2-statistic is significant, the other fit indices are fairly good. Compared to the single-group multi-state model, the fit is slightly worse, but can still be considered comparable. Figure 2 shows the model implied means of the five measurement occasions for each combination of the categorical covariates. From Table 1 and Figure 2, it can be seen that married man have the lowest baseline of depressive symptoms $\pi_0\left(\hat{\beta}_{000} + \hat{\beta}_{0\_1}\right)$. There is, furthermore, a significant average effect of gender $\left(\widehat{\text{AVE}}_{G;\pi_0}\right)$ and marital status $\left(\widehat{\text{AVE}}_{M;\pi_0}\right)$ indicating that women show a higher baseline as well as unmarried participants. Married participants have on average a steeper linear trend $\pi_1\left(\widehat{\text{AVE}}_{M;\pi_1}\right)$ compared to unmarried participants. For the reaction growth component π_2, there is a significant average effect for marital status $\left(\widehat{\text{AVE}}_{M;\pi_2}\right)$. Married participants show a greater reaction component. The interaction between gender and marital status $\left(\hat{\beta}_{211}\right)$ is also significant for the reaction component. For the adaptation component π_3, there is again a significant average effect for marital status $\left(\widehat{\text{AVE}}_{M;\pi_3}\right)$. Married participants have higher depressive symptoms after the diagnosis while symptoms are the highest for married men $\left(\hat{\beta}_{300} + \hat{\beta}_{3\_1}\right)$. The post diagnosis growth component π_4 is the smallest for married men $\left(\hat{\beta}_{400} + \hat{\beta}_{4\_1}\right)$ which is, however, no surprise as this group has the lowest overall depressive symptoms. The post diagnosis level is on average lower for married participants $\left(\widehat{\text{AVE}}_{M;\pi_4}\right)$ and lower for men $\left(\widehat{\text{AVE}}_{G;\pi_4}\right)$.

Table 1. Regressions with predictors for the multi-group latent growth components model.

	$\boldsymbol{\pi_0}$		$\boldsymbol{\pi_1}$		$\boldsymbol{\pi_2}$	
	Estimate	SE	Estimate	SE	Estimate	SE
$\hat{E}(\pi_j)$	2.25*	0.07	1.50*	0.21	1.07*	0.13
$\widehat{\text{AVE}}_{G;\pi_j}$	0.30*	0.10	−0.33	0.32	0.16	0.21
$\widehat{\text{AVE}}_{M;\pi_j}$	−1.11*	0.12	1.07*	0.39	0.69*	0.23
$\hat{\beta}_{j00}$ (intercept)	2.92*	0.17	0.77	0.59	0.71*	0.33

	π_0		π_1		π_2	
	Estimate	SE	Estimate	SE	Estimate	SE
$\hat{\beta}_{j1\_}$ (women)	0.16	0.21	0.12	0.69	−0.47	0.39
$\hat{\beta}_{j\_1}$ (married)	−1.20	0.18	1.38*	0.62	0.26	0.35
$\hat{\beta}_{j11}$ (women, married)	0.20	0.23	−0.67	0.77	0.93*	0.46

	π_3		π_4	
	Estimate	SE	Estimate	SE
$\hat{E}(\pi_j)$	0.92*	0.12	2.69*	0.08
$\widehat{\text{AVE}}_{G;\pi_j}$	−0.23	0.20	0.27*	0.11
$\widehat{\text{AVE}}_{M;\pi_j}$	0.74*	0.24	−0.70*	0.13
$\hat{\beta}_{j00}$ (intercept)	0.40	0.35	3.02*	0.19
$\hat{\beta}_{j1\_}$ (women)	0.15	0.41	0.36	0.23
$\hat{\beta}_{j\_1}$ (married)	1.00*	0.37	−0.64*	0.20
$\hat{\beta}_{j11}$ (women, married)	−0.56	0.46	−0.14	0.26

Note: *$p < .05$.

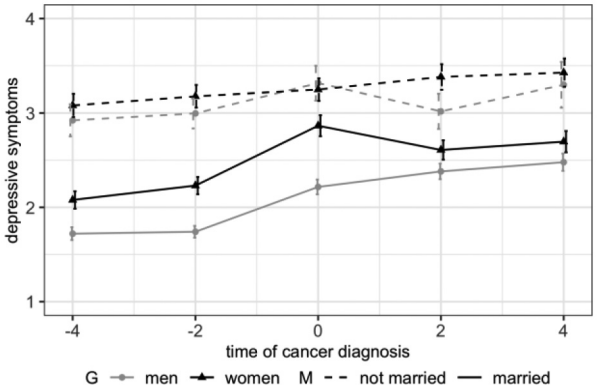


Figure 2. Model implied the mean of depressive symptoms for the categorical covariates. Time zero corresponds to the year of the diagnosis. Negative values correspond to years before the diagnosis and positive values after the diagnosis, respectively. Error bars indicate standard errors.

4. Conclusions

In this paper, we presented the multi-group extension to the latent growth components model. Using data from the Health and Retirement Study, we showed how to specify the MG-LGCA model for five growth components and two categorical covariates each with two levels.

The open-source R package semnova (https://github.com/langenberg/semnova) implements the LGCA and makes the analysis of complex custom growth components conveniently accessible to applied researchers. The semnova package was originally developed for latent repeated measures' analysis of variance which is closely related to latent growth curve modeling and the latent growth components approach, and also builds upon the LGCA for estimating the model. The package currently supports only the single-group LGCA, but the multi-group extension will soon be available.

References

Infurna, F. J., Gerstorf, D., & Ram, N. (2013). The nature and correlates of change in depressive symptoms with cancer diagnosis: Reaction and adaptation. Psy- chology and Aging, 28(2), 386–401. https://doi.org/10.1037/a0029775

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR ap- proach for analyzing average and conditional effects. Multivariate Behavioral Research, 51(2-3), 374–391. https://doi.org/10.1080/00273171.2016.1151334

Mayer, A., Geiser, C., Infurna, F. J., & Fiege, C. (2013). Modelling and predicting com- plex patterns of change using growth component models: An application to depression trajectories in cancer patients. European Journal of Developmental Psychology, 10(1), 40–59. https://doi.org/10.1080/17405629.2012.732721

Mayer, A., Steyer, R., & Mueller, H. (2012). A general approach to defining latent growth components. Structural Equation Modeling, 19(4), 513–533. https:// doi.org/10.1080/10705511 .2012.713242

Mayer, A., & Thoemmes, F. (2019). Analysis of variance models with stochastic group weights. Multivariate Behavioral Research, 54(4), 542–554. https://doi.org/10. 1080/00273171.2018.1548960

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longi- tudinal data. Annual Review of Psychology, 60(1), 577–605. https://doi.org/10. 1146/annurev. psych.60.110707.163612

McArdle, J. J., & Epstein, D. B. (1987). Latent growth curves within developmental structural equation models. Child Development, 58(1), 110–133.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. https://doi.org/10. 1037/10409-005

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525–543. https://doi.org/10.1007/BF02294825

Raykov, T. (1999). Are simple change scores obsolete? an approach to studying cor- relates and predictors of change. Applied Psychological Measurement, 23(2), 120–126. https://doi.org/10.1177/01466219922031248

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1–20. https://doi.org/10.18637/jss.v048.i02

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. Methods of Psychological Research, 2(1), 21– 33.

University of Michigan. (2020). Health and retirement study. https://hrs.isr.umich.edu

The effects of scaling, manifest residual variances, and sample size on the χ2-test statistic of the metric invariance model

Eric Klopp[1], Stefan Klößner[2]

[1]*Department of Education, Saarland University, Germany,*

[2]*Faculty of Educational and Social Sciences, University of Vechta, Germany*

Abstract

In measurement invariance (MI) analysis, the metric invariance model is central. One prominent scaling method restricts the loading of a reference indicator (RI) to unity in all groups to identify the metric MI model. It has recently been argued that this scaling method is problematic when the RI's loading in the population is not invariant, claiming that such a scenario contributes to the low power of the test statistic to detect metric MI violations. However, there are two further scaling methods, and it is obvious to ask whether the same concerns apply to these methods. We demonstrate that the scaling method used to identify the metric MI model generally does not affect the resulting test statistic by using Monte Carlo simulations. Moreover, we demonstrate that the magnitude of manifest residual variances, i.e., measurement error, affects the test statistic heavily; an overlooked effect in the literature. When MI is violated, the test statistic becomes smaller with increasing measurement error, decreasing the power to detect MI violation. Additionally, we show that the test statistic depends on the sample size when MI is violated, which counteracts the effect of increasing measurement error.

Keywords: Measurement invariance, metric, scaling method, sample size measurement error.

E-mail: e.klopp@mx.uni-saarland.de

1. Introduction

The metric measurement invariance (MI) model is the first step of testing MI in confirmatory factor analysis (CFA) models. In this paper, we assume a multiple group setting with one group being assigned the role of reference group. In the metric MI model, the loadings are set equally across two or more groups. If the metric MI model holds, i.e., if its test statistic is not statistically significant, then the well-known further steps in the measurement invariance testing procedure are undertaken (cf., Brown, 2015). However, as the CFA model entails latent variables, there is t need for a scaling method to identify the metric MI model. A widely used method is the Fixed Marker (FM) scaling method (cf., Kline, 2016, pp. 199-200), in which the loading of a reference indicator (RI) is set to 1 in each group. Every single indicator may serve as an RI so that there are as many as possible RIs as there are indicators. Usually, the same indicator is used as an RI in different groups.

In MI analysis, the use of RIs has been discussed critically. Brown (2015, p. 271) mentions that non-invariance of the RI may not be detected when testing the metric MI model because it is fixed to 1 in every group. As a consequence thereof, there is the implicit assumption that the RI is invariant across groups. Kline (2016, p. 405) argues that the restriction of RIs to 1 over the groups is tantamount to an assumption of invariance and that the RI is excluded from the metric MI test. However, this author also states that the RI choice should not affect the model fit most of the time. Finally, Cheung and Rensvold (1999, p. 8) claim that identification constraints imply an invariance assumption and present an example in which either using an invariant or non-invariant RI gives different results for the test statistic of a metric MI model.

Raykov, Marcoulides, Harrison, and Zhang (2020) raised concerns about the dependability of the FM scaling method in the case of non-invariant RI in a metric MI analysis. They presented an example of a one-factor model with 12 manifest indicators in two groups. Their population model contained one indicator that was non-invariant across the groups. In a Monte Carlo simulation using the non-invariant indicator as RI and a sample size of 400 per group, they demonstrated that the metric model's test statistic was non-significant on average. Thus, the model would have failed, on average, to detect the non-invariance. In other words, the metric MI model would, on average, be erroneously accepted. However, Raykov et al. (2020) did not consider choosing another RI.

Johnson, Meade, and DuVernet (2009, p. 654) found in a Monte Carlo simulation for the metric MI model that the metric MI model test was not affected by the invariance or non-invariance of the RI. The authors state that the differences between groups on the RIs were transferred to other indicators via the constraints on the RI to be equal to 1 in both groups, which in turn affected the estimated factor loadings of all other indicators by setting a scale. Furthermore, the authors summarize that the metric MI model tests were generally accurate in detecting non-invariance when the RI differed across groups regardless of the invariance or non-invariance of the other indicators.

Besides the FM scaling method, there are two more scaling methods (cf. Kline, 2016, pp. 199–200). In the Reference Group (RG) method, the estimated latent variance is fixed to 1 in the reference group and freely estimated in the other groups while having the estimated loadings equated over the groups (cf. Kline, 2016, p. 393; Lee, Little, & Preacher, 2011, p. 60). There is also the effects coding method (EC, Little, Slegers, & Card, 2006) that works with constraints on the estimated loadings. These constraints are such that the average of the estimated factor loadings equals 1. In the metric MI model, this restriction applies in a first group, and the estimated loadings are set equally across the other groups. Obviously, these scaling methods come without any assumptions about the invariance of an RI.

Based on the above, we aim to scrutinize two research questions in this paper:

1) Does any special choice of the RI or the general choice of the scaling method affect the metric MI model's test statistic?
2) What is the reason for the failure of the metric MI model's test statistic in Raykov et al. (2020) in detecting the model's non-invariance?

To answer the first question, we used the Raykov et al. (2020) example with Monte Carlo simulations to show that the same test statistic emerged regardless of the choice of any particular RI or the choice of any scaling method. In answer to the second question, we explored, again by Monte Carlo simulation, how sample size and the size of the manifest residual variances, i.e., measurement errors, affected the metric MI model's test statistic.

2. CFA model and simulation settings

In this paper, we assumed the usual multiple-group CFA model with G specifying the group

$$\boldsymbol{y}_G = \boldsymbol{\alpha}_G + \boldsymbol{\Lambda}_G \boldsymbol{\eta}_G + \boldsymbol{\delta}_G. \tag{1}$$

With the usual assumptions (e.g., Bollen, 1989) and the usual conformable matrices and vectors, the model-implied covariance matrix of the model was

$$\boldsymbol{\Sigma}_G = \boldsymbol{\Lambda}_G \boldsymbol{\Phi}_G \boldsymbol{\Lambda}'_G + \boldsymbol{\Theta}_G \tag{2}$$

and the mean structure was

$$E(\boldsymbol{y}_G) = \boldsymbol{\alpha}_G + \boldsymbol{\Lambda}_G \boldsymbol{\tau}_G. \tag{3}$$

To answer the research questions, we used the example provided by Raykov et al. (2020) that consists of a one-factor CFA model with 12 indicators and one latent variable, η_G, in two groups, $G \in \{1, 2\}$. The population model was

$$\boldsymbol{\Lambda}_1 = (1,1.25,1.25,1.25,1.5,1.5,1.5,1.5,1.75,1.75,1.75,2)^T \tag{4}$$

$$\boldsymbol{\Lambda}_2 = (1,1.25,1.25,1.25,1.5,1.5,1.5,1.5,1.75,1.75,1.75,2.5)^T \tag{5}$$

$$\boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2 \tag{6}$$

$$\theta_{\mathrm{i,i,G}} = 2, 1 \leq i \leq 12, \tag{7}$$

$$\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = (1,1.25,1.25,1.25,1.5,1.5,1.5,1.5,1.75,1.75,1.75,2)^T \tag{8}$$

$$\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = 0 \tag{9}$$

Obviously, the last indicator $\lambda_{12,1,G}$ is non-invariant between the two groups and violates the metric MI condition. To estimate this model, we used the FM, RG, and EC scaling methods. For the FM method, we wrote FM1 when the first indicator was the RI, FM2 when the second indicator was the RI, etc. so that there were 14 ways in total to scale the model. We named the model to which a scaling method had been applied, the scaled model. The scaled model, including the mean structure, had 130 degrees of freedom. From estimating a scaled model, a test statistic T is obtained which, under the assumption of the null hypothesis, follows a χ^2–distribution. For a nominal significance level of .05, the critical value of the test statistic was T_{crit} =157.610. We

used the population model from equations (5) to (10) and the maximum likelihood (ML) estimator for all simulations. The number of replications was 10,000. We used Mplus (Muthén & Muthén, 1998-2019) in version 8.3.

3. Effects of scaling

As mentioned above, to answer the first question, we used the Raykov et al. (2020) example. As reported above, the authors found that when using FM12, the test statistic was, on average, below the critical value of the test statistic. As we aimed to explore the effects of using different scaling methods on the metric MI model's test statistic, we expanded this example and used all 14 possible scaling methods. The result of the Monte Carlo simulation was that the resulting average test statistic was 150.394 with a standard deviation of SD=18.205, for all 14 scaling methods. A closer look reveals that the particular scaling method did not affect the test statistic for a specific random draw either. The log-likelihood for the first seed's draw was -18137.550, regardless of the scaling method. The same held for the log-likelihoods of the second and third seed's draw. These were -18028.464 and -18034.866, respectively. The answer to the first research question is that the choice of any special RI or the general choice of the scaling method does not affect the metric MI model's test statistic results.

From a theoretical perspective, this result emerges because the estimates obtained under the different scaling methods are equivalent. As shown in Klopp and Klößner (2021; cf., Klößner & Klopp, 2019), it is possible to convert the estimated parameter values obtained under a specific scaling method to the estimated parameter values obtained under any other scaling method without re-estimating the model. In other words, the parameters estimated using a specific scaling method that minimize the ML discrepancy function can be converted into the parameters that also minimize the ML discrepancy function under any other scaling method. Therefore, the choice of the scaling method does not affect the test statistic.

To sum up, our findings corroborate the results of Johnson et al. (2009) in that the choice of a particular RI does not affect the model's test statistic. Our findings also extend this result in that all scaling methods for the metric MI model are equivalent. However, Johnson et al. (2009) also mentioned that the test statistic could detect non-invariance in the model over their simulations. Therefore, the question still remains as to which specific features of the model in the Raykov et al. (2020) example yielded the failure of the metric MI model to detect the non-invariance on average. This led us to the second research question which we address in the following section.

4. Effects of sample size and measurement error

To answer the second research question, we drew on two issues mentioned in the literature that have received little attention in MI research. The first issue was mentioned by Heene, Hilbert, Draxler, Ziegler, and Bühner (2011). They demonstrated that the measurement error contributed to the size of the test statistic and that a large measurement error could mask the misfit of mis-specified models. According to Browne, MacCallum, Kim, Andersen, and Glaser (2002, p. 403), the test statistic itself has the little-known property that it is more sensitive to misfit when measurement errors are small than when they are large. These authors demonstrated that the measurement error and the test statistic are oppositely related, i.e., when the measurement error decreases, the test statistic increases. Therefore, one of the factors contributing to the failure to detect the deviation from metric MI in the Raykov et al. (2020) example may have been overly large measurement errors.

Meade and Bauer (2007) summarized research demonstrating that the precision of estimated factor loadings increases with increasing sample size. Increasing the precision of the estimated loadings, in turn, also leads to a reduction of noise which is caused by measurement error and yields a decrease of sample fluctuations.

Taken together, we can hypothesize that the relatively large measurement error and the small sample size in each group caused the failure to detect the non-invariance in the example. To test this hypothesis, we conducted a Monte Carlo experiment in which we varied the size of the measurement error and the sample size. Firstly, we varied the size of the measurement error. The original value was $\theta_{i,i,G} = 2$ for all the indicators. We added the three conditions $\theta_{i,i,G} = 0.5$, $\theta_{i,i,G} = 1$, and $\theta_{i,i,G} = 4$. Secondly, we varied the sample size. We used sample sizes of $N = 200$, $N = 400$, $N = 600$, $N = 800$, and $N = 1000$ in each group. Therefore, the model had the same number of degrees of freedom and the same critical test statistic value as the original example. The other settings of the Monte Carlo simulation are as mentioned above. The results are shown in Figure 1. The dashed line indicates the critical value of the test statistic. For the following interpretation of the results, we draw on the mean of the test statistic over the replications, i.e., the results indicate how the test statistic behaves on average in detecting the violation of the metric MI condition in the model.

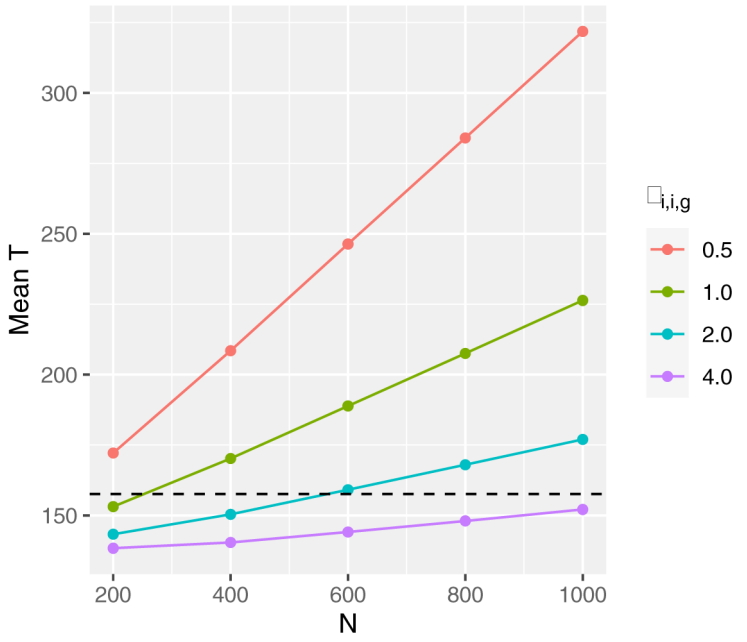


Figure 1. Mean of the test statistic T plotted against sample size for various values of the measurement error. The dashed line represents the critical value of the test statistic.

The results show that for $\theta_{i,i,G} = 2$ (light blue line), sample sizes lower than 600 do not provide enough power to detect the misspecification. The model has only enough power to detect the misspecification for a sample size well above 600. In the condition with $\theta_{i,i,G} = 4$, i.e., double the size of the measurement error, the results (violet line) show that the metric MI model did not have enough power to detect the misspecification for no sample size. In the condition with $\theta_{i,i,G} = 1$ (green line), i.e., half of the size of the measurement error as in the original examples, the model would only fail to detect non-invariance for a sample size of $N = 200$. Finally, in the condition with $\theta_{i,i,G} = 0.5$ (red line), the model shows on average enough power to detect the

metric MI violation for all sample sizes. In summary, the results show what follows from the theoretical reasoning: on average, the power to detect the misspecification increases with increasing sample size and increases with decreasing size of the measurement error. Finally, the answer to the second research question is that in the specific combination of sample size and size of the measurement error, the model will fail in simulations to detect the misspecification most of the time.

5. Conclusions

This study shows that choosing a particular RI or a scaling method is irrelevant for the metric model's test statistic. Thus, the study overcomes the concerns found in the literature about non-invariant RIs and extends the literature showing that the choice of the scaling method is arbitrary. Although the equivalence of the RI choice using the FM scaling methods in particular or the equivalence of the scaling methods in general is a well-known fact in single-group settings, this is not the case in multiple-group settings. We also hint at a neglected issue in the research on MI, i.e., the role of the measurement error. Our results demonstrate the importance of measurement error in the analysis of metric MI. Overly large measurement error may mask the erroneous assumption of metric MI by a non-significant test of the metric MI model.

The study also has its limitations. Firstly, we only demonstrated the equivalence of the scaling methods using a Monte Carlo simulation, providing a conceptual explanation. However, a sound mathematical proof would furnish a final word on the equivalence of scaling methods for the metric MI model. This proof would also be general and not depend on a specific example. Secondly, the exploration of the effects of measurement error on the power to detect non-invariance draws only on the relatively simple example of Raykov et al. (2020). Although the considerations about the role of the measurement error draw on the mathematical derivation by Browne et al. (2002), simulations with more complex models are necessary to obtain a more nuanced view on the effects of measurement errors on the analysis of metric MI. Thirdly, we only considered the ML estimator. However, the same reasoning concerning the relation between the test statistic and sample size, or measurement error, respectively, applies to other estimators that also yield a χ^2-distributed test statistic (cf., Browne, 2002). A fourth and obvious limitation is that the results concerning the equality of the scaling method hold for metric MI models, but not for partial metric MI models. In partial metric MI models, not all loadings are equal over the groups so that constraint interaction may occur (Klößner & Klopp, 2019), i.e., the test statistic may depend on the scaling method and/or the choice of the RI. Finally, we only drew on multiple-group contexts. However, the results can be generalized to metric invariance settings in longitudinal models.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2. Ed.). New York: The Guilford Press.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, 7(4), 403–421.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25(1), 1–27.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, 16, 319–336.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 642–657.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4. Ed.). New York: The Guilford Press.

Klößner, S. & Klopp, E. (2019). Explaining constraint interaction: How to interpret estimated model parameters under alternative scaling methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 143–155.

Klopp, E. & Klößner, S. (2021). The Impact of Scaling Methods on the Properties and Interpretation of Parameter Estimates in Structural Equation Models with Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 28, 182–206.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 59–72.

Meade, A., & Bauer, D. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 611–635.

Muthén, B., & Muthén, L. (1998-2019). *Mplus user's guide. 8th edition*. Los Angeles: Muthén & Muthén.

Raykov, T., Marcoulides, G. A., Harrison, M., & Zhang, M. (2020). On the dependability of a popular procedure for studying measurement invariance: A cause for concern? *Structural Equation Modeling: A Multidisciplinary Journal*, 27(4), 649–656.

Steiger, J. H. (2002). When constraints interact: a caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods*, 7(2), 210–227.

Personality study in otariids (*Otariidae*): the case of the fur seals (*Arctocepalus pusillus*) in Faunia

Ignacio Miguel Pardillo[1], Ángela Loeches[1], Ricardo Olmos[2], Ana María Fidalgo[1]

[1]*Biological and Health Psychology, Autonomous University of Madrid, Madrid, Spain*,

[2]*Methodological and Social Psychology, Autonomous University of Madrid, Madrid, Spain*

Abstract

Animal personality studies have proven to be necessary to control the quality of life of captive-bred individuals. Nevertheless, research in this field has been subject to certain problems. Firstly, the application of Exploratory Factor Analysis (EFA) for the extraction of personality factors is usually accompanied by incorrect practices for the extraction of latent variables, such as Principal Component Analysis, Kaiser's rule and Varimax rotation. Secondly, the small sample size of the work on animal personality does not allow us to meet the necessary assumptions. The purpose of this project is to test a new non-parametric data analysis strategy that is not affected by the small sample size typical of ethology studies, specifically the Hierarchical Cluster Analysis, and we compare its results with those offered by EFA. In addition, we describe the personality of six fur seals (*Arctocepalus pusillus*) living in Faunia, using an animal behavior coding method. The results show us that statistical techniques converged in the clustering of most behaviors; however, the factorial solution was not stable (the data matrix was defined as non-positive). Based on the results, we conclude the existence of three personality traits in the sample studied: extraversion, self-confidence and apathy.

Keywords: Animal personality; coding animal behavior method; Hierarchical Cluster Analysis; fur seals.

E-mail: ignacio.pardillo@estudiante.uam.es

1. Introduction

Personality, defined as a set of individual differences in behavior consistent across time and different contexts (Gosling, 2001), has always been considered a human attribute; however, the use of this term in other species is not an anthropomorphic exercise that should be neglected as it is also applicable to behavioral differences between individual animals. Its consideration has proven to be an essential tool for the management of captive-bred populations (Watters & Powell, 2012). On the one hand, individual welfare is a consequence of the interaction between the personalities of the subjects and the environment and, on the other hand, we can posit behavioral profiles that fit more closely with some of the main roles that animals play in zoos, including breeding and exhibition or performance in educational programs. For example, based on the pentafactorial model (Goldberg, 1981), we can consider variables such as Sociability and Agreeableness to select the individuals that are more likely to reproduce successfully with a new mate, or the Extraversion dimension to facilitate the effectiveness of animals' close interactions with humans. In short, the study of personality in animals becomes essential both for the full adjustment of individuals to the different tasks carried out in zoos, and for the evaluation of their welfare.

1.1. How to measure personality in animals

Numerous studies on personality in animals, including the famous meta-analysis carried out by researchers Gosling & John (1999), point out how its structure becomes similar to that proposed in the pentafactorial model (Goldberg, 1981). The most stable conclusion indicates that the traits best observed across the different groups of species analyzed are Extraversion, Neuroticism and Agreeableness, followed by Openness to experience. In addition, factors such as Dominance become highly relevant when describing groups of non-human animals. Despite the importance of these works, it should be noted that the simple use of traits extracted from human models cannot fully describe a particular animal species. For this reason, the design of species-specific personality factors is necessary (Gosling & John, 1999). In order to measure animal behavioral tendencies, the literature describes two main approaches (Watters & Powell, 2012): rating animal behavior and coding animal behavior. The first approach (rating method) originates from personality research in humans and is based on the opinion offered by experts who are familiar with the sample to be studied: they are given the possibility to attribute values to the behavioral traits of the individual, using questionnaires previously designed by the research team based on personality tests already developed (top-down method) or the behavioral repertoire of the target species (bottom-up method). The second approach (coding method) involves direct observation by a set of observers familiar with the sample, either in a natural or experimental environment, as well as the recording of various measures of species-specific behaviors (collected in an ethogram), including frequency, latency, and duration. While the "trend assessment" method is much faster and collects a larger number of experiences, the effect of the context where the experts familiar with the sample usually have contact with an animal makes behavioral coding a more objective technique (Vazire et al., 2007). Regardless of the method used, the ethologist obtains a series of item scores or real behavioral measures, to which different grouping techniques are applied. Usually, the researcher uses an Exploratory Factor Analysis (EFA) and applies the problematic "Little Jiffy" pack, which includes Principal Component Analysis, eigenvalues greater than one (Kaiser's rule) and Varimax rotation (Ferrando & Anguiano-Carrasco, 2010). In addition, studies on animal personality often apply this kind of parametric techniques to observations collected from a small sample, which may lead to non-compliance

with some necessary assumptions. Considering the above, it is necessary to search for new tools to extract variables that do not impose such restrictive assumptions. This paper proposes the use of Hierarchical Cluster Analysis, a multivariate technique that allows us to group variables without requiring any type of distribution.

1.2. Fur seals' personality

Fur seals (*Arctocepalus pusillus*) belong to the superfamily *Pinnipeda* and the family *Otariidae*. This carnivorous species lives most of its life in the oceans; however, they require land for the birth and development of their pups during the first stages of life, in an ecosystem characterized by the scarcity of safe breeding sites. Consequently, they develop polygamous and gregarious systems. The study of personality in fur seals involves an additional problem: as a research topic it is still unaddressed. Fortunately, we can rely on previous work carried out on other species that are phylogenetically very close, such as Californian sea lions (*Zalophus californianus*), which are also otariids. For example, Amber de Vere (2017), through the coding method, contemplated Extraversion and Routine activity factors in Californian sea lions. In addition, Ciardelli et al. (2017) found up to three personality dimensions in this species by applying the rating method: Extraversion/Impulsiveness, which includes adjectives such as "playfulness", "creative", "curious", "demandingness" and "aggressiveness" (similar to the Extraversion factor extracted by de Vere and the Extraversion and Openness dimensions in humans); Dominance/Confidence, which includes "security" and "fear" (similar to the Routine Activity factor extracted by de Vere); and Reactivity/Independence, which contains the terms "cooperative", "people-friendly" and "people-aggressive".

1.3. Purpose of the paper

The aim of this work is, on the one hand, to describe and study the personality of fur seals (*Arctocepalus pusillus*) in Faunia, applying a coding method; and on the other hand, to use Hierarchical Cluster Analysis as a new data analysis tool, which is less sensitive to non-compliance with some assumptions, and to compare results with those offered by other commonly used techniques, such as EFA. It should be added that the present work also aims to promote research on personality in fur seals, which is still a novel aspect. It could be used as a pilot approach for future studies on personality traits in this species, establishing an empirical basis on which to develop new research with other samples and populations.

2. Method

2.1. Sample

The sample consisted of five fur seals located in Faunia: one male and four females between 14 and 23 years of age. None of these individuals had serious diseases at the beginning of the study, except for two females with certain vision problems.

2.2. Data collection

Two observers carried out an *ad libitum* sampling of the animals' behaviors over a period of two months, in order to draw up a specific ethogram for the sample study (Table 1).

Table 1. Main categories of the ethogram

Behavior	Description
Aquatic locomotion	Immersion in the water.
Grooming	Combing or scratching with their body parts or the substrate.
Floating	Remaining motionless on the surface of the water without showing interest.
Land locomotion	Walking and crawling.
Resting	Resting and sleeping on land.
Nose contact	Intentional friction between the vibrissae of individuals.
Rubbing	Spontaneous contact between animal body parts.
Parallel swimming	Two animals swimming close together.
Aggression	Biting, hitting and growling directed at trainers or animals.
Observer interaction	Observation or exhibition of behaviors directed at people.
Playing	Interaction with environment or individuals with exaggerated movements.
Training	Exhibiting behaviors requested by trainers.
Open mouth	Animal opens its mouth more than 30º without closing it immediately.
Yawn	Animal opens its mouth 90º inhaling deeply.
Nodding	Repeated movement of the head in an up and down direction.
Swinging	Wiggling of the trunk and head to the right and left for 5s or more.
Scanning	Active observation of the surroundings.
Out of sight	The animal is outside the observer's visual range.
Other	

The remaining observations were made over a period of three months in three time slots: from 10:00h to 11:30h, before the zoo activities; from 15:00h to 16:30h, before the afternoon exhibitions and interactions; and from 17:00h to 18:30h, coinciding with the closing of the zoo. We made different observations with and without spectators and trainers, as well as before and after exhibitions or intimate interactions. We carried out between one and two individual continuous focal samplings each day, lasting between thirty and sixty minutes. The total observation time for each subject was over 150 minutes. During the first seven days, both investigators calculated the inter-rate reliability, and we obtained a Kappa coefficient greater than 0.70 for all scores, i.e., a substantial agreement (Landis and Koch, 1977).

2.3. Data Analysis

A Hierarchical Cluster Analysis was conducted using IMB SPSS 25.0 (IBM Corp., 2017). The measure used to obtain the distance matrix was the squared Euclidean distance and the clustering method used was between-group average linkage. Prior standardization of the variables was necessary, using a scale between zero and one. An Exploratory Factor Analysis was also applied, following the guidelines of animal personality studies: Principal Component Analysis, Kaiser's Rule and Oblimin rotation.

3. Results

3.1. Hierarchical Cluster Analysis

The clustering of the different behavioral categories required a total of seventeen steps. The dendrogram (Figure 1) suggests three clusters. The third group includes three categories that have been joined together in fifteen steps and, therefore, represents a more heterogeneous cluster than the others.

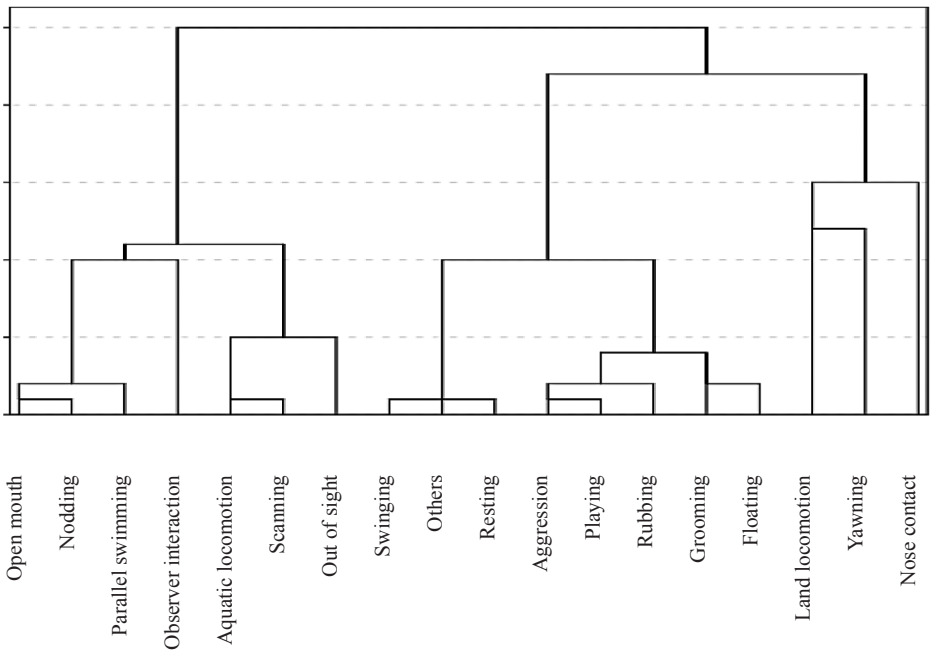


Figure 1. Dendrogram of the behaviors observed in the sample.

3.2. Exploratory Factor Analysis.

We applied the Shapiro-Wilk, KMO and Barlett Test for sphericity to check the adequacy of the data. The first test showed that the frequencies of the behavioral categories Ground locomotion, Play, Mouth opening, Head nodding and Others did not present a normal distribution. Barlett Test for sphericity defined the correlation matrix as non-positive, indicating that the use of EFA was not appropriate. Based on these results, we conclude that it is difficult not to draw biased conclusions from these analyses. However, the results obtained are described below, as they may be useful for a better understanding of the clusters described above. We conclude the presence of four factors (Table 3).

Table 3. Total variance explained.

Factor	Extraction Sums of Squared Loadings		
	Total	**% of variance**	**Cumulative %**
1	8.954	49.747	49.747
2	4.381	24.343	74.091
3	2.426	13.478	87.569
4	2.237	12.430	100.000

We looked at those weights greater than 0.4 (in absolute value) present in the pattern matrix (Table 4). Factors three and four showed certain similarities with cluster one described previously, as did factors one and two with cluster two.

Table 4. Pattern matrix.

Behaviors	Factor			
	1	2	3	4
Resting	0.865			
Land locomotion	0.602		0.612	0.525
Observer interaction	−0.939			
Swinging	0.936			
Others	1.012			
Floating	0.411	0.773		
Grooming		0.857		
Aggression		0.973		
Nose contact		−0.948		
Playing		0.995		
Rubbing		0.834	−0.415	
Aquatic locomotion			0.794	
Scanning			0.756	
Out of sight			0.946	
Synchronized swimming	−0.538		0.438	−0.518
Open mouth				−0.724
Yawn				1
Nodding				−0.752

4. Conclusions

Based on this information, we conclude the existence of three personality traits in the sample of fur seals (*Arctocepalus pusillus*) studied. 1) Extraversion: this cluster encompasses behavioral categories involving high motor activity, sociability and contact with other animals and humans, which concurs with the definitions of Extraversion made by Goldberg (1981), Ciardelli et al. (2017) and de Vere (2017). 2) Self-confidence: this cluster collects behavioral categories that imply calmness, security and dominance, which is coincident with the Dominance/Confidence dimension found by Ciardelli et al. (2017). 3) Apathy: this third group presents greater heterogeneity than the others; however, the behaviors collected imply disinterest in the environment and absence of anxiety in the short term. Fur seals that score high on the Extraversion trait (first cluster) are suitable for performing exhibitions or participating in close interactions with humans (accompanied by high apathy), as well as those that score high on

Extraversion and Self-confidence correlate positively with learning ability and enthusiasm for acquiring new skills (Ciardelli et al., 2017). It is important not to force a specific individual to perform for each activity carried out by the zoo. These conclusions enable us to check the tendency of the individual to stay in one pole or another of the traits found when they experience a change in their routines or participate in new activities.

Additionally, the method employed, based on the coding method and the use of Hierarchical Cluster Analysis, is applicable in future research on personality and welfare in captive-bred marine mammals. It is a correlational methodology and not based on questionnaires, which offers greater objectivity and avoids the appearance of response biases typical of studies that apply a rating method (Vazire et al., 2007). The results offered by the Exploratory Factor Analysis are not stable, because the matrix was defined as non-positive. However, we can conclude a certain convergence between this statistical method and the Hierarchical Cluster Analysis, which grants validity to the results extracted from this method which had not previously been used in this kind of studies. It is worth mentioning the convenience of testing other methods of data analysis that can work with the small sample size typical of ethology studies, since non-parametric tests lack high statistical power. On the other hand, we are testing other behavior measures such as the duration of some states. We also aim to provide convergence validity to the measures by applying a rating method. Finally, the stability of behavioral trends in the animals studied still needs to be assessed, and to this end, we aim to apply Growth Curve Modeling that assesses how much of the variability of the measures related to behavioral categories is explained by personality traits or by relevant variables, such as the time of day or the opening of the zoo to the public.

In conclusion, this work suggests the existence of three personality traits in the Faunia fur seal group, which enables us to select the individuals for the routines of the zoo and to verify the incidence of these on their welfare. In addition, we have tested the effectiveness of the method of observation and coding of animal behavior, as well as the Hierarchical Cluster Analysis as an alternative to Exploratory Factor Analysis.

References

Ferrando, P. J. & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología. *Papeles del psicólogo, 31*(1), 18–33.

Goldberg, L. R. (1981). Language and individual differences. The search for universal in personality lexicón. *Review of personality and Social Psychology*, *2*, 141–165.

Gosling, S. (2001). From mice to men: what can we learn about personality from animal research? *Psychology Bulletin, 127*, 45–86.

Gosling, S. D. & John, O. P. (1999). Personality dimensions in nonhuman animals: A cross-species review. *Current Directions in Psychological Science, 8*(3), 69–75. http://doi.org/10.1111/1467-8721.00017.

IBM Corp. Released (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp. Recuperado de https://www.ibm.com/analytics/es/es/technology/spss/.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174. http://doi.org/10.2307 / 2529310.

Vazire, S., Gosling, S. D., Dickey, A. S. & Schapiro, S. J. (2007). Measuring personality in nonhuman animals. In Robins, R. W., Fraley, R. C. y Krueger, R. F., *Handbook of research methods in personality psychology* (pp. 190–206). The Guilford Press.

Vere, A. J. d. (2017). *Do Pinnipeds Have Personality? Coding Harbor Seal (Phoca vitulina) and California Sea Lion (Zalophus californianus) Behavior Across Contexts* (Master's Theses). University of Southern Mississippi, Mississippi.

Watters, J. V. & Powell, D. M. (2012). Measuring Animal Personality for Use in Population Management in Zoos: Suggested Methods and Rationale. *Zoo Biology, 31*(1), 1–12. http://doi.org/10.1002/zoo.20379.

Improving meta-analytical estimation using *p*-uniform and Fisher's statistic

Juan I. Durán[1], Manuel Suero[2], Juan Botella[2]

[1]*Department of Psychology and Health, Universidad a Distancia de Madrid, Spain,*

[2]*Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Spain*

Abstract

The *p*-uniform method for meta-analytical estimation exploits the fact that when testing a true hypothesis, the distribution of *p*-values is uniform. It can be applied to the subset of statistically significant studies. It consists of finding the value of the parametric effect size for which the distribution of *p*-values is uniform, by fitting a statistic whose distribution is known under such assumption. The method looks for the value for which the statistic is equal to the expected value of such a distribution. The *puniform* package offers several options, including Fisher's statistic, but the implementation of that statistic in *puniform* is not optimal; we propose two changes. First, the degrees of freedom of the corresponding distribution should be $(2k-1)$ instead of $2k$. Second, when the effect size index is Cohen's *d* the *t* distribution should be used to compute the corresponding *p*-values, instead of the normal approximation. The simulation reported here provides evidence supporting this implementation, as it gives less biased estimates that are not accompanied by losses in efficacy (*RMSE*), or in the coverage of the confidence intervals. The proposed implementation is recommended to estimate the standardized mean difference when using the *p*-uniform method under a fixed effect-model.

Keywords: Fisher's method, Meta-analysis, p-uniform, Publication bias.

Funding: Spanish Ministry of Science and Innovation (project reference: PSI2017-82490-P).

E-mail: ignacio.pardillo@estudiante.uam.es

1. Introduction

The meta-analytical method known as *p*-uniform was devised to circumvent the problem of publication bias (PB). It assumes that although there is a significant level of PB, at least the subset of studies with statistically significant results is reasonably intact, or the losses are random (the probability of publication is independent of the *p*-values). Then, if other sources of noise apart from the PB (e.g., the QRP; see Carter, Schönbrodt, Gervais, & Hilgard, 2019) are not present, the *p*-uniform method can provide good pooled estimates of the effect size.

The core idea underlying the *p*-uniform method takes advantage of the well-known fact that if a true null hypothesis is tested then the *p*-values follow a uniform distribution (see Ulrich & Miller, 2018). The *p*-uniform estimation procedure consists of finding the parametric value for which the empirical distribution of *p*-values when testing that value best approximates a uniform distribution. Originally van Assen, van Aert, and Wicherts (2015) developed *p*-uniform for a scenario that is better described with a single parametric value (fixed-effect model; Borenstein, Hedges, Higgins, & Rothstein, 2010). Later, they provided new steps designed for scenarios that are better described with a distribution of parametric values (random-effects model; van Aert, 2020a). Here we only cover the fixed effect model.

One of the tools developed to estimate the parametric effect size under a fixed-effect model is the *R* package *puniform* (van Aert, 2020b), which includes several options for evaluating the fit. These authors recommended one based on the Irwin-Hall distribution (van Aert, Wicherts, & van Assen, 2016) but they also included others, such as Fisher's statistic, whose performance turned out to be somewhat worse. We believe that their implementation of Fisher's statistic was not optimal and that by improving their implementation of that method we can provide very good estimates, even better than those provided by *puniform*, both with the Irwin-Hall distribution and with their implementation of Fisher's method. Let's suppose a random sample of *k p*-values obtained in independent tests of the same hypothesis. If the hypothesis is true, the *p*-values follow a U(0; 1) distribution. Fisher's statistic is,

$$S_F = -2\sum_{i=1}^{k} Log\left(p_i\right) \tag{1}$$

If the *p*-values are uniformly distributed, then S_F follows a chi-square distribution with 2*k* degrees of freedom (Fisher, 1931). Van Aert (2020b) implemented Fisher's method in the *puniform* package in that way. We believe that for estimation processes it is more accurate to assume (2*k*-1) degrees of freedom (df) for the chi-squared distribution. The reason is that the procedure used for the estimation imposes a linear restriction by setting the fitting statistic as being equal to its expected value. Then, the correct distribution is the same, but a degree of freedom is lost for each restriction (Cochran, 1952; Fisher, 1931), even if these restrictions are non-homogeneous (Satterwhite, 1942). Therefore, in this case, the reference distribution must be χ^2 with (2*k*-1) degrees of freedom. The *p*-values follow a uniform distribution if they are obtained with the true parametric value. But this estimation process involves an unknown parameter, which is the one that we are trying to estimate. Then, the estimate should look for the value of the parameter in which Fisher's statistic is equal to the expected value of the χ^2 distribution with (2*k*-1) df.

The implementation of Fisher's procedure in *puniform* has a related problem when applied to the standardized mean difference (δ). The *p*-values are obtained through approximation to the normal distribution, instead of the *t* distribution. The difference is small when the sample sizes are large, but often in psychology the sample sizes are rather small. The *p*-uniform method for

estimating the effect size consists of obtaining the estimated value of the parameter $\hat{\delta}$, for which the *p*-values yield a value of Fisher's statistic equal to the expected value of its distribution. For any value of $\hat{\delta}$, the *p*-value of a study corresponding to the right tail is calculated in *puniform* through (Φ is the cumulative standard normal function)[1],

$$p = 1 - \Phi\left(\left(d - \hat{\delta}\right) \Big/ \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2} + \frac{d^2}{2 \cdot \left(n_1 + n_2\right)}}\right) \tag{2}$$

Then the *p*-values must be transformed, since when selecting the studies according to their statistical significance, the range of possible *p*-values is no longer between 0 and 1. The *p*-values under H_0 cannot be greater than 0.05. When estimating the true effect size, the *p*-values have a greater range, but are still incomplete. Therefore, the *p*-values must be scaled to the interval [0; 1] before applying the fitting procedures. The procedure employed by van Assen, van Aert and Wicherts (2015) essentially consisted of dividing the *p*-value by the maximum possible *p*-value, i.e., the right-hand area of the distribution at the point limiting the rejection area. When the hypothesis tested is true, the scaling value is .05, but when the parameter is different (in the expected direction), the scaling value is larger than .05 (it equals the power of the test for the parametric value assessed).

In short, in the *puniform* implementation of Fisher's procedure the *k p*-values are obtained with (2), and the Fisher's statistic is calculated following (1). Numerical methods are used to obtain the value of $\hat{\delta}$ with which Fisher's statistic equals the assumed expected value of its distribution (2*k*). We believe that for the two reasons outlined above, when the *p*-uniform procedure has been evaluated by taking Fisher's statistic as a baseline, the estimates of the effect size are systematically biased, especially when δ and *k* are small (van Assen et al, 2015).

Van Aert, et al. (2016) concluded by recommending the method based on the Irwin-Hall distribution, since it provided better estimates compared to their implementation of Fisher's statistic and other methods[2]. Again assuming a uniform distribution for the *p* values, the Irwin-Hall distribution describes the probability distribution of the sum of the *k p*-values, $S_{IH} = \Sigma p_i$. The exact probabilities of the S_{IH} values can be calculated but are computationally demanding. However, when $k \geq 10$ the S_{IH} distribution approaches the normal distribution, with the expected value $k/2$ and variance $k/12$ (Johnson, Kotz, & Balakrishnan, 1994; Kocak, 2017). As meta-analyses in psychology are rarely performed with $k < 10$ studies, this method is often used through approximation. Estimation through Fisher's procedure has seldom been assessed, as in simulation studies the choice for *p*-uniform is usually the Irwin-Hall distribution (e.g., Carter et al, 2019; McShane, Böckenholt, & Hansen, 2016).

Our present goal is to propose a corrected implementation to estimate the effect size, δ, through use of the *p*-uniform method and Fisher's statistic. This proposal consists of: (a) a better choice of the degrees of freedom; and (b) the use of an expression equivalent to (2) to assess the fit, in which the exact Student´s *t* distribution is employed instead of the normal approximation

1 The performance of this implementation is additionally reduced because the denominator of (2) uses the value of the sample statistic, *d*. We believe that it should be replaced by the value itself that is evaluated as a parametric value, $\hat{\delta}$, since the variance of *d* depends on the parameter.

2 This recommendation is also based on the fact that it is more robust to heterogeneity and the presence of outliers.

(see below). To support this proposal, we carried out a simulation study assessing three estimation procedures: our implementation of Fisher's method and two *puniform* options, Fisher and Irwin-Hall, as references for comparison. We chose Fisher's method as implemented in *puniform* for comparison to show that the corrected procedure is in fact an improvement. We also chose the *puniform*'s Irwin-Hall method for comparison, to check whether the corrected procedure also outperforms the procedure recommended by its developers. From here we will refer to the three methods i.e., Fisher corrected, Fisher, and Irwin-Hall, as S_{Fc}, S_F and S_{IH}, respectively.

We only cover the scenario in which the meta-analyst decides to select the significant studies. Typically, the meta-analyst makes this decision on a well-founded suspicion that there is a non-negligible significant *PB* compared to non-significant studies. Therefore, the size of the set of studies, *k*, is the number of significant studies, which are the only ones that enter the estimation process.

2. Method

We simulated the results of primary studies by randomly generating two independent samples of values and calculating their means and variances. Each replication had a set of *k* primary studies with significant results, i.e., the sample of independent estimates entering meta-analysis.

We set the sample size ($n_1 = n_2$) of the main body of results at 15, so that the total size of the study ($N = n_1 + n_2$) was 30. However, the supplementary material[3] shows the results for two alternative total sizes, 60 and 120. With those larger sample sizes, the differences between the three methods were smaller than with $N = 30$. The simulation for the condition reported here consisted of generating 15 values from the N(0; 1) distribution and 15 from the N(δ; 1) distribution. The δ values defined 5 conditions: from 0.25 to 1.25 in steps of 0.25. The sizes of the meta-analyses were set at $k = 10$ to 50 in steps of 10. Therefore, the conditions of the simulation consisted of all combinations of 3-factor levels. We generated results of 5000 replications (meta-analyses) for each condition. With the means and variances of each primary study we calculated the test statistic, *T*:

$$T = \left(\bar{X}_1 - \bar{X}_2\right) \Big/ \left(S_{pooled} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \tag{3}$$

In order to reproduce a scenario in which the meta-analyst only retains the significant studies, the primary studies generated that were not statistically significant in a one-sided, right test of the null hypothesis (H_0) were eliminated, with a significance level of $\alpha = .05$. We defined a test as statistically significant if $T > {}_{.95}t_{(N-2)}$, where ${}_{.95}t_{(N-2)}$ was the 95th percentile of the Student's *t* distribution with (*N*-2) df. If the result of a primary study was not statistically significant a completely new one was generated. The process continued until the number *k* of significant studies of the condition was reached.

We applied the three chosen procedures to each replication to estimate δ from the simulated data. They can be categorized within the *p*-uniform method, but with different fitting criteria. Two of them were implemented by the authors of the *p*-uniform method and the third one is the one proposed here. For the first two we used the *puniform* function of the *puniform* package (van Aert, 2020b) from *R* (R core team, 2019). The inputs were the sample sizes, n_1 and n_2, and

3 Supplementary material: https://osf.io/nwprh/

the value of the test statistic for two independent means, already calculated with (3). We ran the estimations according to the criteria S_{IH} (based on the Irwin-Hall distribution) and S_F (based on Fisher's method). The output from *puniform* for each of these criteria included the point estimate (δ), the estimated standard error, and the confidence interval (95%).

The criterion for fitting proposed here involved two differences with *puniform*: (a) the reference value for the estimate was the expected value of the chi-square distribution with (2*k*−1) df, instead of 2*k*; (b) the *p*-values were obtained from the $t_{N\text{-}2}$ distribution instead of the normal approximation. The distribution of the test's statistic was a non-central Student's *t* with the non-centrality parameter $\delta \cdot \text{sqrt}[n_1 \cdot n_2/(n_1 + n_2)]$. When the null hypothesis was true, $\delta = 0$ and the distribution was the central Student's *t*. The *p*-values were then re-scaled in order to obtain uniform distribution in the range [0:1]. This was done here following van Assen, van Aert and Wicherts (2015), but with the *t* distribution (see the appendix in the supplementary material).

In order to estimate the effect size, δ, we replaced it with its estimate ($\hat{\delta}$) in the non-centrality parameter (see the appendix in the supplementary material):

$$\frac{1-F\left[T_i;(N-2),\hat{\delta}\cdot\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}\right]}{1-F\left[{}_{1-\alpha}t_{N-2};(N-2),\hat{\delta}\cdot\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}\right]} \tag{4}$$

Where T_i is the test statistic for two independent means, equation (3), and ${}_{1-\alpha}t_{N-2}$ is the $(1-\alpha)\cdot 100^{\text{th}}$ percentile of the non-central Student's *t* distribution with $(N-2)$ df.

More specifically, we looked for the parametric value, $\hat{\delta}$, for which the re-scaled *p*-values best fitted the uniform distribution. Thus, Fisher's statistic was calculated with the *p*-values obtained with (4) and set to being equal to the expected value of the reference distribution (2*k*−1). The estimated value, $\hat{\delta}$, was the one that gave a value of the Fisher's statistic being equal to its expected value, (2*k*-1), through the R CRAN function *uniroot*. We also obtained the limits of the 95% confidence interval by performing a similar search through the R CRAN function *uniroot*, but matching Fisher's statistic to the values of the χ^2 distribution with (2*k*-1) *df* corresponding to the 2.5th and 97.5th percentiles (see the supplementary material for the *R* syntax).

Once the 5,000 estimates of $\hat{\delta}$ were obtained according to each fitting criteria we calculated the following for the three criteria and for each condition (Burton, Altman, Royston, & Holder, 2006): (a) the mean of the estimates, $\bar{\hat{\delta}} = \Sigma\hat{\delta}_i / 5000$; (b) the estimated bias as the difference between the average of the estimates and the parameter, $(\bar{\hat{\delta}} - \delta)$; (c) the *RMSE*, as the square root of the empirical variance of the estimates, $\sqrt{\sum\left(\hat{\delta}_i - \bar{\hat{\delta}}\right)^2 \Big/ 5000}$; and (d) the coverage of the confidence intervals as the proportion of replications that included the true parametric value, $(LL \leq \delta \leq UL)$.

3. Results

Some of the main results for the $N = 30$ conditions appear in figure 1, left (the rest of the figures and all the data points are in the supplementary material). The bias was much smaller with the proposed implementation of Fisher's procedure than with the two procedures implemented in

puniform. On the one hand, it is better than the one provided by *puniform* with its implementation of Fisher's criterion, which we expected if the *t* distribution was employed and when our choice of the degrees of freedom was correct. On the other hand, it is better than the one provided by *puniform* with the Irwin-Hall criterion, which is the one recommended by its developers (van Aert, Wicherts, & van Assen, 2016). Although the bias was largely reduced in all conditions, the difference with the other methods was greater when the number of studies was small (10 or 20). The absolute bias was as large with large δ values (1.0 and 1.25) as it was with small δ values, but the relative bias was in fact much smaller in those conditions. The results in the conditions with larger sample sizes (60 and 120) reflect the same trends, although the difference between the bias among the three conditions was smaller in almost all conditions (see the supplementary material).

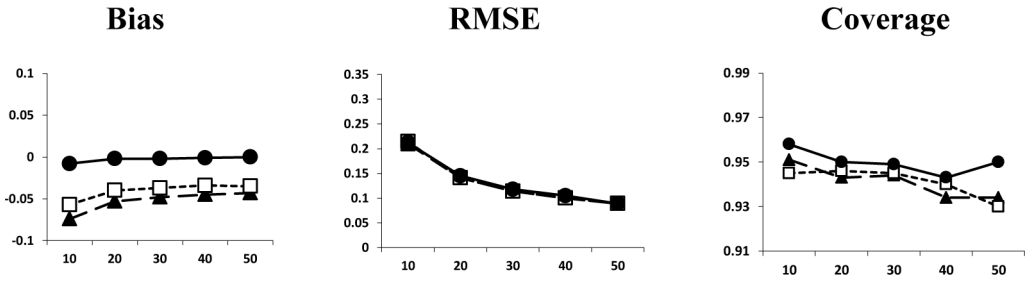


Figure 1. Values of bias, RMSE and coverage of the confidence intervals, for each number of studies (k, on the X axis), with δ = 0.75 and N = 30 (triangles, F; squares, IH; circles, F corrected).

The efficacy, measured through the *RMSE* (figure 1, central), was nearly the same with all methods in almost all conditions (including those with $N = 60$ and 120). The only difference appeared with $\delta = 0.25$ and $k = 10$, where S_{Fc} outperformed the other two methods in the three N conditions. The coverage was close to the nominal value (.95) both with the proposed method and with *puniform* through the two alternative fitting criteria. None of the three was uniformly better. Our method had worse coverage with $\delta = 0.25$ and $k = 20$ or 30, as well as with $\delta = 0.50$ and $k = 10$ or 20. However, with high values of δ (1.0 and 1.25) the coverage deviated very little from 0.95 in the whole range studied here. With $N = 60$ or 120, our procedure was the only one with coverages close to .95 in the whole range of values of δ and k studied here.

4. Conclusions

The procedure proposed here provides estimates generally better than *puniform*, especially in terms of bias. It is based on Fisher's statistic but with the df corrected ($2k$-1) and the *p*-values obtained from the exact *t* distribution. The smaller bias is not accompanied by declines in other features. Specifically, (a) the efficacy was essentially equivalent, and (b) the coverage was better in most conditions. Our corrected Fisher's method should become the recommended one for fixed-effect models. Of course, as the number of studies increases, the difference between using $2k$ or ($2k$-1) for the estimation becomes smaller.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effects and random-effects models for meta-analysis. *Research Synthesis Methods*, 1, 97–111.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25, 4279–4292.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of mathematical statistics*, 315-345.

Fisher, R. A. (1931). *Statistical Methods for Research Work*. Edinburg and London: Oliver and Boyd.

Johnson, N., Kotz, S, & Balakrishnan, N. (1994). *Continuous Univariate Distributions (Second Edition, Volume 2)*. New York: Wiley.

Kocak, M. (2017). Meta-analysis of univariate P-values. *Communications in Statistics - Simulation and Computation, 46*:2, 1257–1265.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11, 730–749.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Satterthwaite, F. E. (1942). Linear restrictions on chi square. *Annals of Mathematical Statistics*, 13 (3), 326–331.

Ulrich, R., & Miller, J. (2018). Some properties of p-curves, with an application to gradual publication bias. *Psychological Methods*, 23(3), 546.

Van Aert, R. C. M. (2020a, August 11). *Correcting for publication bias in a meta-analysis with the p-uniform* method*. Retrieved from osf.io/ebq6m.

van Aert, R. C. M. (2020b). *Puniform: Meta-Analysis methods correcting for publication bias*. R package version 0.2.2. https://CRAN.R-project.org/package=puniform

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: reservations and recommendations for applying p-uniform and p-curve. *Perspectives in Psychological Science*, 11, 713–729.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20, 293–309.

“In medio virtus”: Searching for the factor structure between fit and parsimony

Pablo Nájera[1], Francisco J. Abad[1], Miguel A. Sorrel[1]

[1]*Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Spain*

Abstract

Factor analysis is arguably the most common procedure for scale validation. Confirmatory factor analysis (CFA) was for a long time the preferred technique; recently, exploratory structural equation modeling (ESEM) has become more popular because of its flexibility and better model fit. However, these features can also be a drawback since practitioners might be tempted to retain well-fitting models with little theoretical interpretation. Moreover, the greater complexity of ESEM might lead to unstable results with low sample sizes. Several alternatives have recently been developed with the aim of offering a middle-ground solution between fit and parsimony. However, a comparison of these alternatives is yet to be made in order to provide practical guidelines. A simulation study was conducted to compare the performance of CFA, CFA with modifications, ESEM, ESEM-based CFA (CFA_E), and Bayesian SEM (BSEM) in terms of parameter estimation accuracy. CFA and BSEM could not properly recover the internal structure under the presence of cross-loadings, overestimating factor correlations. On the contrary, ESEM showed an underestimation tendency for factor correlations. The middle-ground solutions, especially CFA_E, managed to reduce the bias of parameter estimates. Practical guidelines are discussed.

Keywords: Factor analysis; validation; CFA; ESEM; BSEM.

Funding: This study has been supported by the Ministerio de Economía y Competitividad and European Social Fund (PSI2017-85022-P), the UAM-IIC Chair «Psychometric Models and Applications», the Asociación Española de Metodología de las Ciencias del Comportamiento (AEMCCO), and the European Association of Methodology (EAM).

E-mail: pablo.najera@uam.es

1. Introduction

Factor analysis is arguably the most common procedure for scale validation in psychological and educational research. The development of confirmatory factor analysis (CFA) has become the most recommended technique for scale validation due to its parsimony and alignment with the hypothesis testing approach. However, in the last two decades, several researchers have pointed out some important limitations derived from the overly restrictive structure of CFA models. To name a few, CFA models have been found to provide biased parameter estimates under slightly mis-specified models, with a particularly concerning tendency to greatly overestimate factor correlations (Asparouhov & Muthén, 2009; Marsh et al., 2014). As a result, interest in exploratory factor analysis has recently increased, especially after the development of exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009). Among other advantages, ESEM shows greater flexibility than CFA by allowing all factor loadings to be unrestricted. This usually results in considerably better model fit. The popularity of ESEM can be reflected by the vast number of scales and questionnaires that have been re-analyzed with this technique, sometimes leading to substantially different interpretations from those previously found with CFA (e.g., Garrido et al., 2020). However, all these desirable features of ESEM come with a risk. Namely, applied researchers might be tempted to retain these well-fitting models, even though they lack theoretical interpretability. This practice is susceptible of capitalization on chance; that is, capturing the idiosyncrasies of a particular sample, thus hindering the generalization of the conclusions to other contexts (MacCallum et al., 1992). Moreover, the greater complexity of ESEM might lead to unstable results with low sample sizes, in addition to blending the definition of the factors (Marsh et al., 2020).

CFA and ESEM can be conceptualized as the two poles of a continuum. In this vein, the more precise terminology of *restricted* and *unrestricted* factor analysis (see Ferrando, 2021) better reflects this idea. Thus, it can be argued that the strengths derived from the restrictiveness of CFA (e.g., parsimony, theoretical interpretability) are the main shortcomings of ESEM, while the advantages derived from the flexibility of ESEM (e.g., better model fit, less biased parameter estimates) coincide with the drawbacks of CFA. Given the circumstances, it might seem worthy to explore the middle-ground terrain between the two poles of the continuum, with the aim of finding a balance between fit and parsimony that can mitigate the limitations of both options while enhancing their advantages. Of course, the opposite could also occur (i.e., combining more pitfalls than benefits from CFA and ESEM). With the former outcome in mind, several factor analytic techniques have recently been developed, such as Bayesian structural equation modeling (BSEM; Muthén & Asparouhov, 2012), regularized structural equation modeling (RegSEM; Jacobucci et al., 2016), and the objectively refined target matrix procedure (RETAM; Lorenzo-Seva & Ferrando, 2020). Thus, these methods joined the pool of already existing middle-ground techniques, which included the use of modification indices for sequential model modification in CFA, as well as the target rotation procedures for exploratory models (see Browne, 2001).

The growing increase of available options for factor analysis is undoubtedly valuable, as it provides more specialized tools for scale validation studies. However, the usefulness of such developments is certainly subject to a certain degree of mastery of the different factor analytic techniques. A larger number of techniques requires a greater level of study and mastery, and practitioners might find themselves lost in all the available options. Thus, it is necessary to clarify this potential paradox of choice by systematically comparing the performance of several factor analytic techniques under a comprehensive and unified set of conditions. The Filling this

gap, which is precisely the purpose of the present study will enable applied guidelines to be specified for scale validation studies.

2. Method

A Monte Carlo simulation study was conducted to compare the performance of five factor analytic techniques: CFA, CFA with sequential model modification using modification indices (CFA_{MI}), ESEM, BSEM, and a CFA based on the statistically significant loadings from the ESEM (CFA_{E}). A *p*-value of 0.001 was used as a cutoff point for modification indices in CFA_{MI}, while an informative prior of $N(0, 0.01)$ was used for non-target loadings in BSEM. All techniques were implemented with the *MplusAutomation* package (Hallquist & Wiley, 2018) of R software (R Core Team, 2021) and self-developed functions.

Six independent variables were systematically manipulated, including sample size ($N = 300$, 1000), number of factors ($K = 3, 5$), number of items per factor ($JK = 4, 8$), magnitude of main loadings ($ML = 0.5, 0.7$), magnitude of cross-loadings ($CL = 0.15, 0.30$), and number of cross-loadings per factor ($CLK = 1, 2$). Factor correlations were fixed at 0.5, which is a moderate correlation often found in applied studies (e.g., Wiesner & Schanding, 2013).

Standardized continuous variables were generated following the common factor model and the procedure proposed by Cudeck and Browne (1992), which enables a known degree of model misfit to be introduced at the population level. Namely, a population root mean squared error of approximation (RMSEA) of 0.05 was used in the present study. One hundred datasets were generated per condition.

The performance of the factor analytic techniques was evaluated in terms of parameter estimation accuracy, which was calculated as the bias and root-mean-squared-error (RMSE) of main loadings, cross-loadings, zero-loadings (i.e., the factor loadings that equal 0 in the population), and factor correlations.

3. Results

Overall, all techniques achieved a more accurate parameter estimation with greater main loadings ($ML = 0.7$), larger sample sizes ($N = 1000$), and a higher number of items per factor ($JK = 8$). There were no relevant interactions between the independent variables and the factor analytic techniques. Consequently, Figure 1 summarizes the overall parameter estimation accuracy of the major loadings, cross-loadings, zero-loadings, and factor correlations. The techniques have been ordered along the confirmatory-exploratory continuum, where clear tendencies are seen for parameter estimation along such continuum.

First, all techniques properly recovered the main loadings, with a very similar overall estimation accuracy ($.056 \leq RMSE \leq .067$). CFA and BSEM performed similarly across conditions. By definition, CFA provided the worst cross-loading estimates and the most accurate zero-loading estimates. Similarly, BSEM tended to underestimate cross-loadings ($Bias = -.094$ and $RMSE = .099$). Both techniques obtained inflated factor correlations ($Bias \geq .078$ and $RMSE \geq .100$). On the contrary, ESEM obtained accurate cross-loading estimates ($Bias = .010$ and $RMSE = .067$) and a slightly poorer recovery of zero-loadings ($Bias = .016$ and $RMSE = .064$) compared to the more confirmatory techniques. Moreover, it showed a noticeable factor correlation underestimation tendency ($Bias = -.117$), resulting in the least accurate factor correlation estimation ($RMSE = .137$).

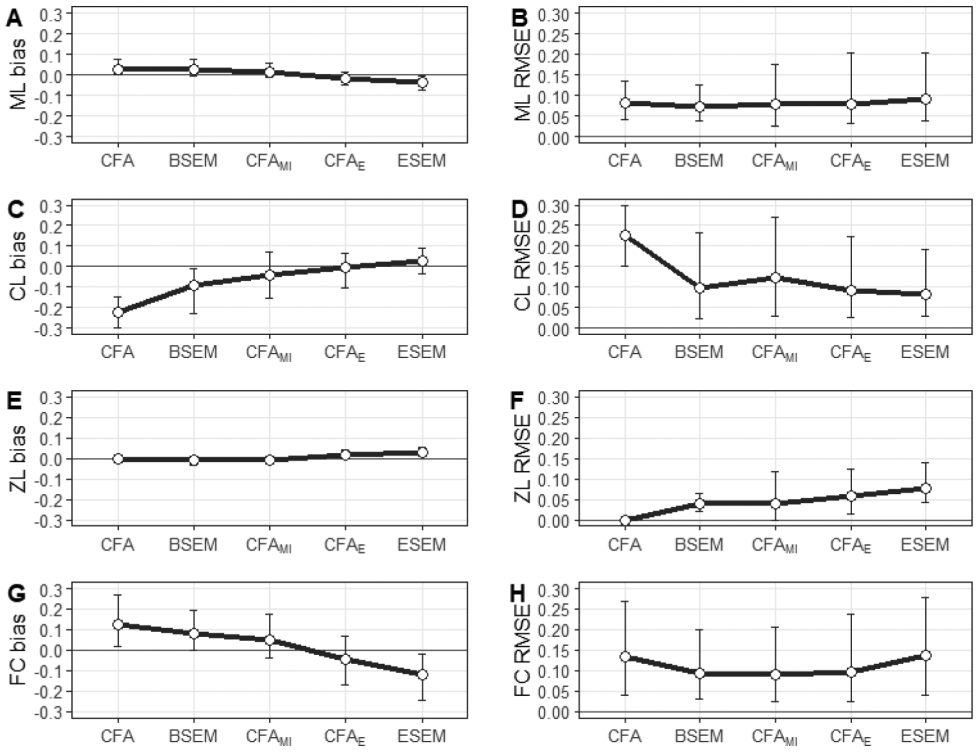


Figure 1. Parameter estimation accuracy. Panel A: major-loadings bias; Panel B: major-loadings RMSE; Panel C: cross-loadings bias; Panel D: cross-loadings RMSE; Panel E: zero-loadings bias; Panel F: zero-loadings RMSE; Panel G: factor correlations bias; Panel H: factor correlations RMSE.

Regarding the middle-ground solutions, CFA_{MI} managed to mitigate the biased estimations of CFA, with an overall better recovery of cross-loadings ($\Delta RMSE$ = –.122 and –.016, respectively) and factor correlations ($\Delta RMSE$ = –.044 and –.018, respectively). Closer to the exploratory side, CFA_E increased the overall accuracy of zero-loading estimates ($\Delta RMSE$ = –.021) compared to ESEM. Moreover, CFA_E substantially improved the recovery of factor correlations ($\Delta RMSE$ = –.041). Overall, CFA_E showed the least biased estimations of all the techniques.

4. Conclusions

Given the lack of studies dedicated to systematically comparing the performance of the most popular and novel factor analysis techniques under a unified set of conditions, the present investigation had the main purpose of shedding some light on this topic with the ultimate goal of providing applied guidelines for scale validation studies.

By means of a simulation study, it was shown that, in line with previous literature (Asparouhov & Muthén, 2009; Marsh et al., 2014), the overly restrictive specifications of the traditional CFA make it unable to recover the internal structure of slightly mis-specified models (i.e., with few cross-loadings). The poor performance of BSEM was less expected. It provided similar estimates to those obtained by CFA, which indicates that the informative priors used for non-target loadings, with a standard deviation of 0.01, were too restrictive. In this vein, Muthén and Asparouhov (2012) pointed out the importance of conducting sensitivity analyses to evaluate the impact of the prior choice, although this might entail a practical burden in applied research (MacCallum et al., 2012).

On the other pole of the continuum, ESEM has been considered as the golden standard for scale validation studies in the last decade. The present investigation does not support this statement, since ESEM failed to properly recover the internal structure of the generating models under the explored conditions, which included a moderate factor correlation of 0.5. In this scenario, it provided inflated cross-loadings and zero-loadings in addition to underestimated factor correlations. In applied research, lower factor correlations might give the impression of clearly differentiated constructs, which is often found as a desirable outcome (e.g., Marsh et al., 2020). This, together with the satisfactory model fit usually obtained by ESEM, might make practitioners retain these models without further considerations. Under these settings, ESEM might lead to imprecise conclusions. Unfortunately, due to their complexity, psychological constructs tend to be interrelated, and moderate factor correlations are consistently found in many fields (e.g., Wiesner & Schanding, 2013). Under these circumstances, the present study does not support the notion that ESEM is the most appropriate factor analytic technique.

The parsimony of CFA_E managed to provide less biased and more accurate estimates than ESEM. Accordingly, its use is recommended whenever the constructs under measurement are expected to be moderately correlated, especially in situations where a relatively novel knowledge domain is being covered or the constructs under measurement are not clearly defined, and thus the use of procedures that require *a priori* specification of the internal structure (i.e., CFA, CFA_{MI}, BSEM) becomes a challenge. In these settings, the CFA_E, which does not rely on any pre-specified model, might be a sound option to explore the internal structure of the scale.

The findings and implications of the present research should be interpreted within the limits of the present simulation study and the conditions explored. Moreover, this investigation should be further extended to other factor analytic techniques that could not be incorporated in the simulation study due to their high computation time, as was the case for ESEM with target rotation, the RETAM procedure, and RegSEM. Finally, it is very important to note the importance of evaluating the theoretical interpretability of all factor analysis models. Just as a meaningless parameter should not be introduced in a CFA although it shows a large modification index, an uninterpretable cross-loading should not be blindly accepted in an ESEM. As MacCallum et al. (1992) noticed, purely relying on data-driven procedures to assess the internal structure of a scale without considering the theoretical interpretability of such structures has a high probability of leading to capitalization on chance and non-generalizable models with little practical utility.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 1–14. https://doi.org/10.1080/10705511.2020.1764360

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, *57*(3), 357–369. https://doi.org/10.1007/BF02295424

Ferrando, P. J. (2021). Seven decades of factor analysis: From Yela to the present day. *Psicothema*, *33*(3), 378–385. https://doi.org/10.7334/psicothema2021.24

Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H. F., Legaz, E., Ferrís, G., & Rojo-Moreno, L. (2020). Is small still beautiful for the Strengths and Difficulties Questionnaire? Novel findings using exploratory structural equation modeling. *Assessment*, *27*(6), 1349–1367. https://doi.org/10.1177/1073191118780461

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Lorenzo-Seva, U., & Ferrando, P. J. (2020). Unrestricted factor analysis of multidimensional test items based on an objectively refined target matrix. *Behavior Research Methods*, *52*, 116–130. https://doi.org/10.3758/s13428-019-01209-1

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, *17*(3), 340–345. https://doi.org/10.1037/a0027131

MacCallum, R. C., Roznowski, M., & Necowitz, B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504.

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and Set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, *55*(1), 102–119. https://doi.org/10.1080/00273171.2019.1602503

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

R Core Team (2021). *R (Version 4.0) [Computer Software]*. Vienna, Austria: R Foundation for Statistical Computing.

Wiesner, M., & Schanding, G. T. (2013). Exploratory structural equation modeling, bifactor models, and standard confirmatory factor analysis models: Application to the BASC-2 Behavioral and Emotional Screening System Teacher Form. *Journal of School Psychology*, *51*, 751–763. https://doi.org/10.1016/j.jsp.2013.09.001

Bayesian versus frequentist approaches in multilevel single-case designs: on power and type I error rate

Cristina Rodríguez-Prada[1], Ricardo Olmos[1], José Ángel Martínez-Huertas[1]

[1]*Department of Social Psychology and Methodology, School of Psychology, Autonomous University of Madrid (Spain)*

Abstract

Single-case designs aim to analyze the effect of interventions on one or a limited number of subjects by collecting measurements over time. While traditionally the possibility of analyzing this data with inferential tools was not foreseen, multilevel models have been gaining relevance in the last decade. In these models, repeated measures are nested in a higher grouping level, namely the subject or individual. This work compares frequentist and Bayesian statistical perspectives on estimation methods and their effects on power and type I error rate, and the proportion of times that the correct model is selected. A Monte Carlo simulation single-case AB study was carried out. A wide range of conditions was manipulated: the number of subjects, the number of repeated measures, the intervention effect size and the population model that generated the data. Results showed that the frequentist methods performed well for simpler models, while the Bayesian framework was the most consistent and recommended when few subjects were available.

Keywords: Multilevel designs; multilevel modeling; statistical applications; simulation study; Bayesian statistics; frequentist approach (maximum 6).

Funding: This study has been supported by Ayudas al Fomento de la Investigación en Másteres Oficiales-UAM 2020-2021 (grant awarded to Cristina Rodríguez-Prada).

E-mail: cristina.rodriguezp02@estudiante.uam.es

1. Introduction

Traditionally, quantitative data analysis in single-case designs has been scarce. One of the associated problems is the small number of subjects in these studies. Another problem is that classical models ignore the statistical dependence between subjects' repeated measures (e.g., ANOVAs or t-tests are discouraged in this type of design). However, in recent years, there has been a proliferation of studies highlighting that multilevel regression models are an interesting statistical approach to analyze these studies (Baek et al., 2013; Baek et al., 2020; Moeyaert et al., 2017).

Multilevel regression models, also known as hierarchical models or mixed models, are well suited to single-case designs because they can model different sources of variability at different levels. In these studies, it is common to model the effect of the intervention or change in the dependent variable between the baseline and the treatment phase (i.e., within-subject variability or change is modeled), but it is also possible to model differences between subjects, either differences at baseline (i.e., intercept variance) or individual differences in the treatment effect (i.e., slope variance). Between-subject variability arises from what are known in these models as level 2 units, represented by the subjects. Within-subject variability, on the other hand, arises in the level 1 units, which would be the subjects' repeated measurements. In multilevel models, level 1 units are nested within level 2 units. In single-case designs, repeated measurements are nested within each subject.

A virtue of these models lies in the flexibility for modeling the aforementioned variance components (i.e., within-subject variance, or intercept or slope variances). Modeling and estimating all these sources of variability enriches clinical practice: for example, by exploring why some clients change earlier, more or are better than others.

However, estimating random effects is challenging because of their complexity, especially when the sample size is small. While maximum likelihood methods are the best known and used by conventional software, they require strong assumptions that are severely compromised in small samples. Studies indicate that covariance parameter estimates can be biased in single-case designs (Moeyaert et al., 2017; Baek et al, 2020). Much of the recent literature advocates the use of Bayesian estimation methods, where the incorporation of uncertainty and prior knowledge through *a priori* distributions can compensate for the limited information available in the data (i.e., few level 2 units).

Part of the current debate focuses on the selection of suitable priors. While the preferred choice has been non-informative prior distributions for fixed effects and level 1 residual variance (e.g., normal distributions with mean 0 and wide variances, or the gamma-inverse distribution for the residual variance), there is less consensus for random parameters. At present, there seems to be more support for the use of *a priori* weakly informative distributions, but this is not a clear-cut issue.

This study aims to compare the frequentist and Bayesian frameworks (with different families of prior distributions) in terms of power, type I error rate and the proportion of times that both frameworks make a correct model selection through the indicators AIC, BIC -for the frequentist methods- and WAIC -for the Bayesian methods-. These objectives allow us to answer: (1) which estimation method performs better in terms of a good balance between type I error rate and statistical power; and (2) which framework is the most suitable to properly select the model that generates the data, a decisive issue prior to interpreting its parameters.

2. Method

The present simulation study worked exclusively with type AB single-case designs, where A refers to measurements taken at the baseline (where the subject has not yet received any treatment) and B is the intervention phase (where measurements of the subject are collected during the intervention). Statistical power, type I error rate and the proportion of times a fit index selects a model correctly were used as dependent variables of the study.

The simulated data were generated following the structure of four multilevel models which differed in the number of random effects. The simplest of the simulated models was called "minimal" and lacked random effects (nor intercept variance nor slope variance). A model called 'partial-intercept' included a random effect for the intercept (between-subject differences in baseline). The 'partial-slope' model included the variance parameter for slopes. The fourth and last model, called 'maximal', contained both random effects: variances for the intercept and for the slopes. The covariance between slopes and intercepts was set to 0 in the simulations.

The most complex model used in the simulations was:

$$y_{ij} = \gamma_{00} + \gamma_{10} D_{ij} + u_{0j} + u_{1j} D_{ij} + e_{ij}$$

where y_{ij} represents the dependent variable measured at moment i for subject j, D_{ij} is the treatment (0 for baseline condition and 1 for treatment condition), γ_{00} is the general intercept, γ_{10} is the fixed effect of the intervention, u_{0j} is the random effect for subject j in the baseline condition, u_{1j} is the random effect for subject j for the slope treatment effect, and e_{ij} is the level-1 residual at the i observation for subject j.

500 replications were performed for each of the 144 simulation conditions in the R programming environment with the RStudio interface (RStudio Team, 2019).

The *MASS* (Venables and Ripley, 2002), *lme4* (Bates et al., 2015) and *nlme* (Pinheiro et al., 2021) packages were used to estimate the frequentist models, selecting REML as the estimation method. The *brms* library (Bürkner, 2017) was used for Bayesian estimation.

2.1. Simulation conditions

In the simulation, some parameters were set to fixed values: γ_{00} was set to 5 throughout the simulation, $e_{ij} \sim N(0,1)$ and σ_e^2 were set to 1. The manipulated conditions can be found in Table 1, together with the studied prior distributions. While the prior distributions for fixed parameters γ_{00} and γ_{10} were set throughout the simulation to non-informative Normal(0,10^6), and σ_e^2 to non-informative Inverse-gamma(0.001,0.001), we studied several prior distributions for σ_{U0}^2 and for σ_{U1}^2.

Table 1. Simulation conditions of the study

Variables/parameters	Manipulated values
N. subjects	3, 5, 7
N. repeated measurements	10, 20, 30, 40
ES (γ_{10})	0 (null), 1.15 (medium), 2.70 (large)
Population model	Maximal, Partial (random intercepts), Partial (random slopes), Maximal

Variables/parameters	**Manipulated values**			
	Scenario 1	***Scenario 2***	***Scenario 3***	**'Prior' type**
γ_{00}, γ_{10}	Normal ~(0, 10^6) (Moeyaert et al., 2017)			Non-informative
e_{ij}	Inverse-gamma~(0.001, 0.001)			Non-informative
u_{0j}, u_{1j}	Half-Cauchy~(0,10)	Half-Cauchy~(0,20)	*Half-Cauchy~(0,50)*	Weakly-informative
	Half-normal~(0,10)	Half-normal~(0,20)	*Half-normal~(0,50)*	Weakly-informative
	Uniform(0, 100)			

2.2. Data analysis

Concerning power and type I error rate, the p-value (<0.05) in the frequentist perspective and the 95% quantiles of the credible intervals in the Bayesian perspective (if they did not enclose the value 0, a statistically significant effect was considered to be present) were used to assess the effect of the intervention.

In terms of selecting the correct model according to the BIC, AIC and WAIC indices, each model generated was always analyzed with the four analysis models and each of the three indicators was calculated. The lowest value was the one that indicated the chosen model.

3. Results

3.1. Power, type I error rate and model selection in Bayesian methods

The differences are very small among the seven Bayesian methods used with respect to power, Type I error rate and correct model selection. All selected priors show very similar behavior (Table 2).

The power is not excessively high on average (0.61). The Type I error rate is somewhat conservative (0.033) and the proportion of times the correct model was selected with the WAIC index is 84%. The small differences point to the better performance of the half-Cauchy(0, 10) prior which was chosen for further analysis following a parsimony criterion.

Table 2. Descriptive statistics (means and standard deviations) of power, type I error rate and correct model selection for Bayesian methods.

	Power	**Type I error rate**	**Hit rate (model selection)**
	Mean (SD)	Mean (SD)	Mean (SD)
Uniform ~ (0, 100)	0.605 (0.49)	0.032 (0.18)	0.841 (0.37)
Half-Cauchy ~ (0, 10)	0.624 (0.48)	0.033 (0.18)	0.841 (0.37)
Half-Cauchy ~ (0, 20)	0.611 (0.49)	0.033 (0.18)	0.840 (0.37)
Half-Cauchy ~ (0, 50)	0.607 (0.49)	0.033 (0.18)	0.841 (0.36)
Half-normal ~ (0, 10)	0.618 (0.49)	0.033 (0.18)	0.842 (0.36)
Half-normal ~ (0, 20)	0.609 (0.49)	0.033 (0.18)	0.841 (0.37)
Half-normal ~ (0, 50)	0.607 (0.49)	0.033 (0.18)	0.842 (0.36)

3.2. Model selection by relative fit indices

There are no major differences between AIC, BIC and WAIC with respect to the proportion of correct model selection ($F(1.709, 121968.106)=510.138$, $p < 0.001$; $\eta^2_{partial} = 0.007$).

An interaction effect between fit indices and the simulated model (i.e., minimal, partial-intercept, partial-slope and maximal) nuances this conclusion. The selection of the correct model is highly dependent on the simulated population model ($F(5.126,121968.106) = 4379.944$, $p < 0.001$; $\eta^2_{partial} = 0.155$. When it is the 'maximal' one, WAIC detects it correctly; when the simulated model is less complex, the WAIC index has to choose them to a lesser extent than when the correct model is the maximal one. Frequentist indices choose simpler models. AIC is mostly stable (Figure 1).

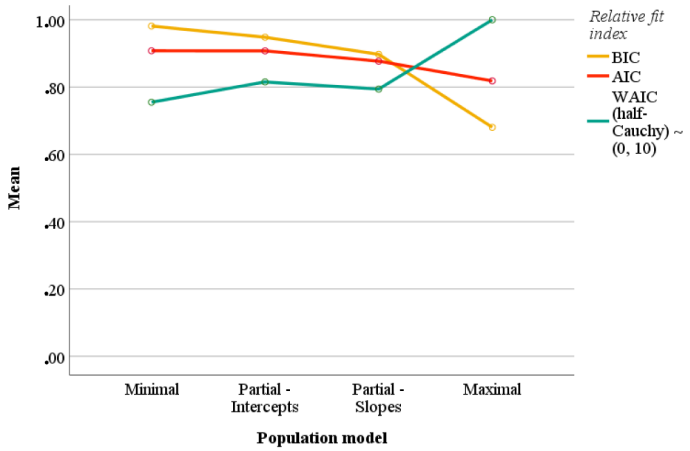


Figure 1. Proportion of hits for each relative fit index according to the population model

The higher the number of subjects and the higher the number of repeated measurements, the higher the proportion of correct model selection, with 5 being the minimum value at which all indexes reach values above 0.8. WAIC does not depend so much on these conditions, while frequentists report larger fluctuations.

A detailed descriptive analysis of the errors (i.e., when the relative fit index did not select the correct model) showed that the three indices taken into account had different error patterns. While BIC tends to over-penalize complex models and tends to select the simpler models (the 'minimal' and 'partial intercepts models), WAIC tends to select the 'maximal' model almost half of the time. AIC spreads its errors more evenly, although it tends to select the random intercepts model more often (Figure 2).

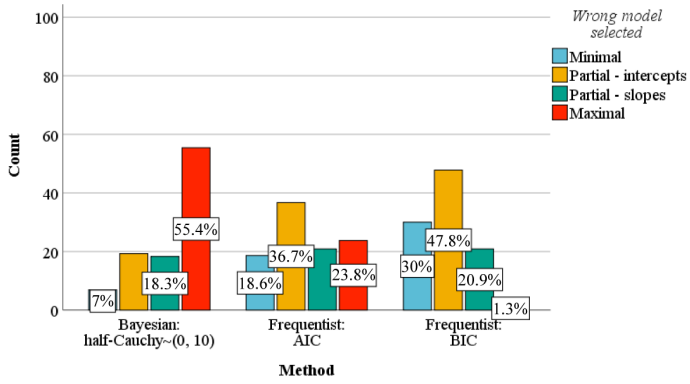


Figure 2. Percentage of times of wrong model selection for each relative fit index

4. Conclusions

The Bayesian methods showed no statistically significant differences between non-informative or weakly informative priors in terms of performance on power, Type I error rate and model selection. The choice between them was not overly determinative. At the descriptive level, the half-Cauchy and half-normal priors performed best, which is in line with the recommendations of Moeyaert et al. (2017) and Baek et al. (2020).

No major differences were found between the three relative fit indices regarding model selection. This would correspond to the equivalence between the Bayesian-frequentist methods established by Moeyaert et al. (2017), although it is true that Bayesian tends to favor complex models and frequentist favors simpler models. Taking this into account, we can say that in more complex scenarios, Bayesian methods can be a good alternative due to the difficulties that frequentist methods present for the estimation of random effects in complex models, which are precisely the elements that characterize them.

Although the performances were better for the simpler models used in this study, it is questionable what practical impact they have. They may be models that fit well but are scarcely found or do not adequately represent reality (such as the 'minimal' model). Simulating relatively simple models may partially explain the better performance of frequentist methods compared to Bayesian methods. The use of more complex models and the comparison between methods could reveal an advantage of Bayesian methods over frequentist ones. Sensitivity tests would need to be done on the selection agreement of the same model with various relative fit indices. Therefore, the recommendation for the applied researcher would be to choose one estimation method or another depending on the complexity of the assumed model.

References

Baek, E., Beretvas, S. N., Noortgate, W. V. den, & Ferron, J. M. (2020). Brief Research Report: Bayesian Versus REML Estimations With Noninformative Priors in Multilevel Single-Case Data. *The Journal of Experimental Education*, *88*(4), 698–710. https://doi.org/10.1080/00220973.2018.1527280

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods*, *45*(1), 65–74. https://doi.org/10.3758/s13428-012-0231-z

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods*, *22*(4), 760–778. https://doi.org/10.1037/met0000136

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R-Core Team. (2021). *nlme: Linear and Nonlinear Mixed Effects Models* (3.1–152) [Computer software]. https://CRAN.R-project.org/package=nlme

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4.ª ed.). Springer-Verlag. https://doi.org/10.1007/978-0-387-21706-2

Two-tier Exploratory Modeling

Marcos Jiménez-Henríquez[1], Francisco J. Abad[1], Eduardo Garcia-Garzon[2], Luis Eduardo Garrido[3]

[1]*Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, Spain,*

[2]*Universidad Camilo José Cela, Spain,*

[3]*Pontificia Universidad Católica Madre y Maestra, Dominican Republic*

Abstract

The two-tier model is a generalization of the bi-factor model that aims to estimate a primary layer of general factors and a secondary layer of group factors. However, unlike in the bi-factor case, an exploratory method to estimate the two-tier model remains to be proposed. We developed such a method (GSLiD) by rotating the loading matrix with a target rotation criterion that facilitates the differential identifiability of the general and group factors, with the target being updated upon each rotation until an optimal one is obtained. Furthermore, a Newton-based algorithm was used to quickly estimate these large factor patterns. Results from a Monte Carlo simulation suggest our method outperforms the Schmid-Leiman alternative and is robust to challenging conditions involving cross-loadings and pure items. Thereby, we supply an R package to make this class of models readily available for substantive research. Finally, we used GSLiD to assess the two-tier structure of a reduced version of the Personality Inventory for DSM-5 Short Form.

Keywords: Bi-factor; exploratory factor analysis; two-tier; target rotation.

Funding: This research was supported by Grant PSI2017-85022-P (Ministerio de Ciencia, Innovación y Universidades, Spain) and the UAM IIC Chair Psychometric Models and Applications.

E-mail: marcosjnezhquez@gmail.com

1. Introduction

Bi-factor modeling is an increasingly popular strategy to conceptualize psychological constructs (Reise, 2012). For instance, it has been argued to prompt the understanding of complex phenomena like intelligence (Beaujean, 2015), personality (Abad et al., 2018) and psychopathology (Bornovalova et al., 2020). Its distinctive feature is that it addresses within-item multidimensionality by allowing the indicators to load simultaneously on one general factor (e.g., emotional stability) and narrower group factors (e.g., anxiety and depression). Currently, the exploratory estimation of these structures is an active research area with proposals involving the use of analytic rotation criteria (Jennrich & Bentler, 2011) and target matrices on the factor loadings (Abad et al., 2017; Garcia-Garzon et al., 2019; Lorenzo-Seva & Ferrando, 2019; Waller, 2018).

A limitation of the bi-factor model is that it enables a single general factor. As a consequence, applied researchers analyzing large factor structures may find themselves constrained to fit an independent bi-factor model to each domain of the data. We thus propose a generalization of the bi-factor model, termed the two-tier model, in which multiple general factors and group factors are jointly considered (Figure 1). The two-tier model can accommodate several bi-factor structures within a unique model, making it well suited to uncover complex relations among broad traits that otherwise would remain hidden in large factor structures. A key feature of the two-tier model that we propose is that it is fully exploratory, allowing all items to cross-load in both layers of general and group factors. Furthermore, factors within the same layer are allowed to freely correlate with each other.

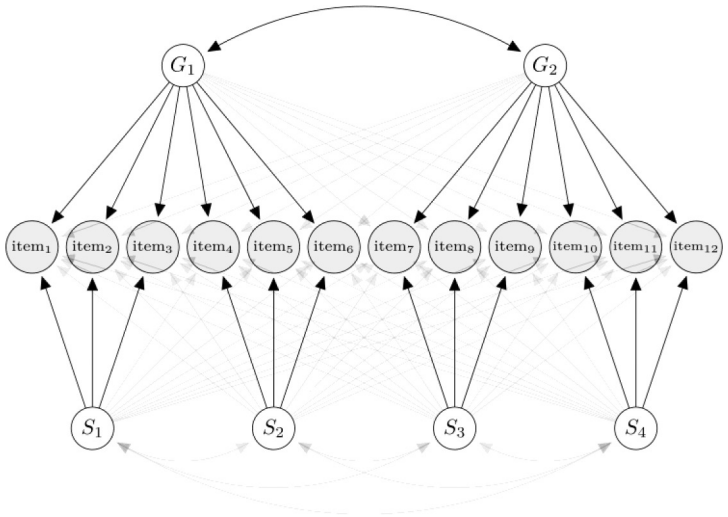


Figure 1. Illustration of the two-tier exploratory model with two general dimensions and four group factors for twelve indicators. Dark arrows correspond to expected path loadings while light arrows indicate possible cross-loadings and correlations. The correlations between the factors of both tiers are also estimated in this exploratory approach but are not displayed for greater clarity.

1.1. The Schmid-Leiman transformation

The Schmid-Leiman transformation (SL) gives a straightforward approximation to a two-tier configuration in an exploratory manner. It is based on the following hierarchical representation of the empirical correlation matrix **R**,

$$\mathbf{R} = \Lambda_1 \Phi_1 \Lambda_1^{\top} + \Psi_1, \tag{1}$$

$$\Phi_1 = \Lambda_2 \Phi_2 \Lambda_2^{\top} + \Psi_2, \tag{2}$$

where $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ denote a loading matrix, a correlation matrix among factors, and a diagonal matrix of uniquenesses, respectively.

Replacing (2) in (1) and expanding, we obtain $\mathbf{R} = \Lambda_1 \Lambda_2 \Phi_2 \Lambda_2^{\top} \Lambda_1^{\top} + \Lambda_1 \Psi_2 \Lambda_1^{\top} + \Psi_1$,

which can be arranged as $\mathbf{R} = (\Lambda_1 \Lambda_2 \Phi_2^{1/2} \vdots \Lambda_1 \Psi_2^{1/2})(\Lambda_1 \Lambda_2 \Phi_2^{1/2} \vdots \Lambda_1 \Psi_2^{1/2})^{\top} + \Psi_1$,

where $(\mathbf{X} \vdots \mathbf{Y})$ denotes the column-wise concatenation of matrices $\mathbf{X}$ and $\mathbf{Y}$ with same row dimension. Finally, we can obtain a two-tier exploratory model configuration with correlated general factors by setting

$$\Lambda_{\text{SL}} = (\Lambda_1 \Lambda_2 \vdots \Lambda_1 \Psi_2^{1/2}), \tag{5}$$

$$\Phi_{\text{SL}} = \begin{pmatrix} \Phi_2 & 0 \\ 0 & \mathbf{I} \end{pmatrix}. \tag{6}$$

Unfortunately, SL results in a rank-deficient solution by imposing linear dependencies on the factor loading matrix (Mansolf & Reise, 2016; Waller, 2018). In the bi-factor case, this transformation is useful to identify independent cluster structures (McDonald, 2000) but is unable to recover realistic bi-factor structures, which are likely to include cross-loadings among the group factors and pure item loadings on the general factors (Abad et al., 2017; Reise et al., 2011). As we expect the same situation to occur in the two-tier structure, we suggest a novel method that aims to estimate two-tier exploratory models for the first time and deals with cross-loadings and pure items. The description of this algorithm, which we have termed the *Generalized Schmid-Leiman Iterative Difference-based Target Rotation* (GSLiD), is given in the next section.

1.2. The Generalized Schmid-Leiman Iterative Difference-based Target Rotation

We propose an iterative target rotation procedure that automatically refines the target matrix for the loadings while taking into account the presence of two layers of general and group factors. This iterative procedure can be regarded as a generalization of the SLiD algorithm developed by Garcia-Garzon et al. (2019), which has been applied with success in exploratory bi-factor modeling (Garcia-Garzon et al., 2021).

Let $\mathbf{A}$ be a $p \times q$ matrix of unrotated factor loadings with p manifest variables and q common factors. The rotation problem is conceptualized as the estimation of a transformation matrix $\mathbf{X}$ so that the rotated factor solution $\mathbf{\Lambda} = \mathbf{A}\mathbf{X}^{-T}$ minimizes some complexity functions to provide a more interpretable loading matrix pattern. When $\mathbf{X}$ is constrained to the oblique manifold of rotation matrices, $\mathcal{OB} = \{\mathbf{X} \in \mathbb{R}^{q \times q} : \text{ddiag}(\mathbf{\Phi} = \mathbf{X}^T\mathbf{X}) = \mathbf{I}\}$, where ddiag($\mathbf{X}$) returns a diagonal matrix with the diagonal elements of $\mathbf{X}$ the off-diagonal elements of $\mathbf{\Phi}$ correspond to the correlations between the factors.

Until recently, all complexity functions only concerned the rotated loading matrix $\mathbf{\Lambda}$**.** However, Zhang et al. (2019) proposed a new complexity function based on partially specified targets for both factor loadings and factor correlations (e.g., the extended target criterion). This criterion was successfully applied to identify multitrait-multimethod structures where the

correlations among trait factors and method factors are freely estimated, but the correlations between them are penalized the more they deviate from zero. The rotation problem can be defined as finding the solution to

$$\underset{\mathbf{x}\in\mathcal{OB}}{argmin} \quad \frac{1}{2}\| \mathbf{W}_{\Lambda} \odot (\boldsymbol{\Lambda} - \mathbf{T}_{\Lambda}) \|^2 + \frac{\mathrm{w}}{4}\| \mathbf{W}_{\boldsymbol{\Phi}} \odot (\boldsymbol{\Phi} - \mathbf{T}_{\boldsymbol{\Phi}}) \|^2, \tag{7}$$

where $\odot$ is the Hadamard product. $\mathbf{W}_{\Lambda}$ and $\mathbf{T}_{\Lambda}$ denote the weight and target matrices for the loading matrix, respectively, while $\mathbf{W}_{\Phi}$ and $\mathbf{T}_{\Phi}$ bear the analog interpretation for the factor correlations. $\mathbf{T}_{\Phi}$ must be symmetric, and $\mathbf{W}_{\Phi}$ must be off-diagonal symmetric with non-negative elements. The scalar *w* ponders the relative contribution of the second term of (7) to solve the minimization problem.

To efficiently rotate a factor solution with an arbitrary number of correlated general factors, we propose setting an initial partially specified target on the factor loadings based on SL, as described above. The target matrix is then updated upon each rotation until it matches the target created in a previous iteration. This update is performed separately for the general and group factors to distinguish between broad and narrower traits and consists of calculating, for each factor, the mean of the one-lagged differences between the sorted squared normalized loadings. These values are then used as cut-offs to create the new target matrix. Such automatic determination of the target has been shown to improve on the demarcation of subjective cut-points in complex structures with many small cross-loadings (Garcia-Garzon et al., 2019). An illustration of this updating method can be found in Table 1 of Garcia-Garzon et al. (2019). Conversely, the targets for the factor correlations remain constant in the GSLiD algorithm. They must be provided by the researcher.

2. Method

To test the estimation accuracy of GSLiD, we ran an extensive simulation involving many variables of interest. The simulation can be considered an extension of the one found in Abad et al. (2017). In this case, two additional variables were considered: (1) the number of general factors and (2) the correlation between the general factors. Thus, nine variables were manipulated in a Monte Carlo simulation to accomplish a fully crossed design that amounts to 7776 conditions, each replicated 50 times. The variables and their levels were: (1) number of general factors (N.GF: 2, 3, 4, 5); (2) correlation between the general factors (COR.GF: 0, 0.5); (3) sample size (N: 500, 1000, 2000); (4) variables per group factor (VAR.GRF: 4, 5, 6); (5) number of group factors defining each general factor (NUM.GRF: 4, 5, 6); (6) cross-loadings between the group factors (CROSS.GRF: no, yes); (7) factor loadings on the group factors (LOAD.GRF: low, medium, high); (8) factor loadings on the general factors (LOAD.GF: low, medium, high); and (9) pure indicators of the general factor (PURE.GF: no, yes).

With this simulation, we tried to investigate the stability of the methods in the presence of two well-known disturbances of the simple structure, namely cross-loadings between the group factors and pure item loadings on the general factors. The combinations of these variables recreate the four types of structures investigated in Abad et al. (2017): (IC) Independent cluster structure: neither cross-loadings nor pure indicators are present; (ICB) Independent cluster basis: cross-loadings but not pure indicators are present; (ICP) Independent cluster pure: pure indicators but not cross-loadings are present; and (ICBP) Independent cluster pure basis: both cross-loadings and pure indicators are present.

The performance of the SL and GSLiD methods were compared in two outcomes: the average Tucker's factor congruence coefficients (ACC) between the simulated and estimated factor

loadings and the root mean square error between the true and estimated correlations among the general factors ($\hat{\boldsymbol{\Phi}}_{\mathrm{g}}\mathrm{RMSR}$).

$$\mathrm{ACC} = \frac{1}{\mathrm{q}}\sum_{j}\frac{\sum_{\mathrm{i}}\hat{\lambda}_{\mathrm{ij}}\lambda_{\mathrm{ij}}}{\sqrt{\sum_{\mathrm{i}}\widehat{\lambda_{\mathrm{ij}}^{2}}\sum_{\mathrm{i}}\lambda_{\mathrm{ij}}^{2}}}, \quad \hat{\boldsymbol{\Phi}}_{\mathrm{g}}\mathrm{RMSR} = \sqrt{\sum_{\mathrm{i}>\mathrm{j}}\frac{\left(\phi_{\mathrm{g}_{\mathrm{ij}}} - \hat{\phi}_{\mathrm{g}_{\mathrm{ij}}}\right)^{2}}{\mathrm{g}\left(\mathrm{g}-1\right)/2}}, \tag{9}$$

where **g** denotes the number of general factors.

For each condition, we generated 50 population structures from which a random sample was drawn from a multivariate normal distribution. ANOVAs estimating up to third-order interactions among all the variables, treated as factors, were carried out for each combination of outcome and method. The partial omega squared Ω^{2}_{prtl} was then used as an effect size measuring the importance of each coefficient.

3. Results

The results of the ANOVA on the ACC (Table 3) confirmed that GSLiD was substantially less sensitive than SL to most of the variables, with the latter being largely influenced by the presence of pure items and cross-loadings ($\Omega^{2}_{prtl}[PURE.GF] = .90$; $\Omega^{2}_{prtl}[CROSS.GRF] = .80$), which were also involved in several high two-way interactions. Whereas SL slightly overcame GSLiD in the independent cluster structure (IC: ACC [SL] = .975; ACC [GSLiD] = .965), it provided worse results in the remaining structures. Figure 2 displays the third-order interaction between pure items, cross-loadings, and the number of variables per group factor. GSLiD was stable under all conditions except ICBP structures with four indicators per group factor, while SL underperformed in the presence of pure items (ICP), especially when they occurred simultaneously with cross-loadings in the ICBP structures ($\Omega^{2}_{prtl}[CROSS.GRF \times PURE.GF] = .62$).

Concerning the recovery of the correlations between the general factors, all marginal $\hat{\boldsymbol{\Phi}}_{\mathrm{g}}\mathrm{RMSR}$s were approximately twice as small for GSLiD as for SL, improving the correlation estimates across all four structure types. In total, 23 out of 25 marginal RMSEs were smaller than .05 for GLSiD, while SL produced an average RMSE below this threshold only under the orthogonal general factor level ($\hat{\boldsymbol{\Phi}}_{\mathrm{g}}\mathrm{RMSR}[\mathrm{COR.GF} = 0] = .031$).

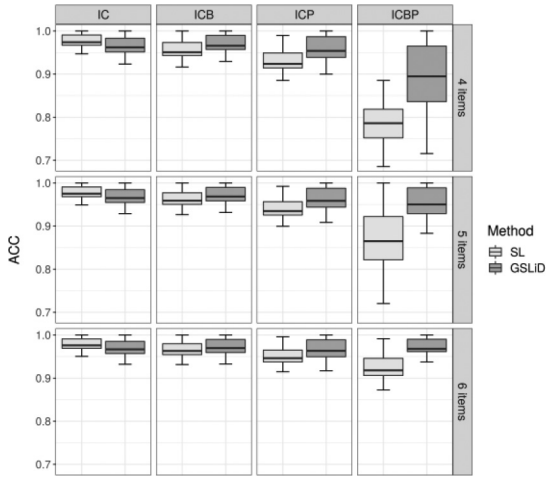


Figure 2. Interaction PURE.GF × CROSS.GRF × VAR.GRF on the ACC for GSLiD and SL. Each box displays the interquartile range, and the middle line denotes the mean value. The error bars delineate the range of typical values, defined as the median of ACC ± 1.5 times the interquartile range.

4. Conclusions

The two-tier model is an extension of the bi-factor model that estimates factor structures involving more than one general factor. Until now, researchers have been restricted to separately analyzing general dimensions in order to study such complex structures, ignoring the presence of cross-loadings and other potential patterns related to the structure of the general factors. Consequently, current methods may not resemble important aspects of an underlying two-tier structure. To overcome this situation, we developed an algorithm (GSLiD) to reliably estimate, for the first time, two-tier exploratory models. GSLiD is a flexible method that accommodates the presence of cross-loadings and factor correlations both among general and group factors and guarantees few convergence problems.

References

Abad, F. J., Garcia-Garzon, E., Garrido, L. E., & Barrada, J. R. (2017). Iteration of partially specified target matrices: Application to the bi-factor case. *Multivariate Behavioral Research*, *52*(4), 416–429. https://doi.org/10.1080/00273171.2017.1301244

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment*, *25*(8), 959–977. https://doi.org/10.1177/1073191116667547

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, *3*(4, 4), 121–136. https://doi.org/10.3390/jintelligence3040121

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18–27. https://doi.org/10.1016/j.biopsych.2020.01.013

Garcia-Garzon, E., Abad, F. J., & Garrido, L. E. (2019). Improving bi-factor exploratory modeling. *Methodology*, *15*(2), 45–55. https://doi.org/10.1027/1614-2241/a000163

Garcia-Garzon, E., Abad, F. J., & Garrido, L. E. (2021). On omega hierarchical estimation: A comparison of exploratory bi-factor analysis algorithms. *Multivariate Behavioral Research*, *56*(1), 101–119. https://doi.org/10.1080/00273171.2020.1736977

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*(4), 537–549. https://doi.org/10.1007/s11336-011-9218-4

Lorenzo-Seva, U., & Ferrando, P. J. (2019). A general approach for fitting pure exploratory bifactor models. Multivariate Behavioral Research, 54(1), 15–30. https://doi.org/10.1080/00273171.2018.1484339

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2(2), 57–64. https://doi.org/10.1027/1614-2241.2.2.57

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. Multivariate Behavioral Research, 51(5), 698–717. https://doi.org/10.1080/00273171.2016.1215898

McDonald, R. P. (2000). A basis for multidimensional item response theory. Applied Psychological Measurement, 24(2), 99–114. https://doi.org/10.1177/01466210022031552

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. Educational and Psychological Measurement, 71(4), 684–711. https://doi.org/10.1177/0013164410378690

Waller, N. G. (2018). Direct Schmid–Leiman transformations and rank-deficient loadings matrices. Psychometrika, 83(4), 858–870. https://doi.org/10.1007/s11336-017-9599-0

Zhang, G., Hattori, M., Trichtinger, L. A., & Wang, X. (2019). Target rotation with both factor loadings and factor correlations. Psychological Methods, 24(3), 390–402. https://doi.org/10.1037/met0000198

Predicting tweet emotionality via Latent Semantic Analysis

Diego Iglesias[1], Miguel Sorrel[1], Ricardo Olmos[1]

[1]*Departmento de Psicología Social y Metodologia, Universidad Autónoma de Madrid, España*

Abstract

In this paper we evaluate whether Latent Semantic Analysis (LSA) could be used to analyze tweets. First, we describe how a vector space model such as LSA simulates human semantics analyzing a large corpus of texts and creating a *K* dimensional semantic space in which words are represented as vectors. Then, we present an empirical study in which we studied whether the information contained in the latent semantic space is purely semantic and abstract. To this end, we tested whether neural network models receiving LSA information as input could predict the emotionality (i.e., emotional category and valence) of short written texts such as tweets. In the case of the emotional category, 38% of the tweets were correctly classified. For emotional valence, predictions improved up to a 71% hit rate. These results suggest that LSA can capture embodied features such as the emotionality of short written texts. Therefore, it may contain more than semantic and abstract information.

Keywords: Latent Semantic Analysis (LSA); sentiment analysis; tweets; symbols; embodiment.

Author Note

We would like to give special thanks to Universidad Autónoma de Madrid for granting the 'Ayuda al Fomento de la Investigación en Másteres Oficiales – UAM'. The first author was a recipient of this scholarship.

E-mail: diego.iglesiaso@estudiante.uam.es

1. Introduction

This paper aims to illustrate how a vector space model like Latent Semantic Analysis (LSA) simulates human semantics. First, we define the theoretical basis of the model. Then, we analyze a relevant issue for psychology such as the representational power of words.

LSA is a psychological model that represents human semantics analyzing the relationship between different linguistic units in an automatic way (Landauer & Dumais, 1997). The central idea of LSA is that the aggregate of all the contexts in which a word appears and does not appear provides information about how words are semantically related to each other (Deerwester et al., 1990).

In order to represent these semantic relationships, LSA starts by analyzing a large corpus of text which define what is known as the linguistic corpus. The function of the linguistic corpus is to emulate the knowledge that a real person may have. The first step of the analysis consists of segmenting the texts that form the linguistic corpus into units with meaning: words and documents (paragraphs). This identifies which words appear in which documents. The result of this process is stored in a document-term matrix in which the terms (words) are placed in the rows and the documents (contexts) in the columns. The cells of the matrix indicate the number of times (frequency) that a word appears in a document. This matrix only reflects superficial aspects of language such as co-occurrences between words and documents (first-order relationships). This does not represent the interest of LSA. For example, words like "car" and "automobile", which have the same meaning, do not seem to be semantically related (Table 1). After all, synonyms like "car" and "automobile" never co-occur, when one is used it replaces the other.

Table 1. Example of a document-term matrix.

	Doc 1	Doc 2	Doc 3	Doc N
Car	0	1	0	0
Automobile	1	0	0	1
Road	1	1	0	0
Fuel	0	1	0	1
Word M	1	1	0	1

Applying the singular value decomposition (SVD), a linear decomposition method of matrices into independent principal components, LSA tries to find the number of dimensions that represent the most relevant information in the original matrix. In this way, it generates an *K* dimensional latent semantic space capable of detecting higher order semantic relations (Landauer & Dumais, 1997), such as the indirect relation between "car" and "automobile" mediated by words like "road", "fuel", etc.

In the latent semantic space, words are represented as vectors with coordinates on the *K* dimensions retained (Figure 1). Sentences and documents are represented as vectors that reflect the result of the sum of the vectors of the words in it. It is important to note that these dimensions or components are independent and meaningless (Deerwester et al., 1990).

The number of dimensions that form a specific latent semantic space is an empirical question. Based on an inductive trial and error method, the number of dimensions retained are those with which the LSA obtains the best results in tasks in which it simulates human behavior (Quesada, 2007). In practice, the best results have been obtained with 300 dimensions, which seems to be

the number of dimensions that best represents human semantics (Landauer & Dumais, 1997; Landauer et al., 1998; Redher et al., 1998).

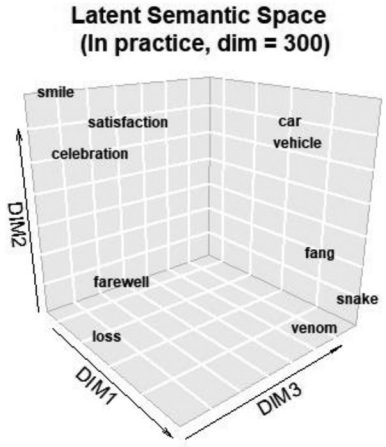


Figure 1. Graphical representation of the latent semantic space. Although a three-dimensional representation is shown here (i.e., K= 3), in practice it is typical to extract 300 dimensions.

The words represented in the latent semantic space are abstract symbols, i.e., amodal representations. That is, there is no natural relationship between the symbol "snake" and a real snake. Apparently, these symbols contain only semantic (i.e., abstract) information. However, Landauer (1999) argues that symbols are richer in information. After all, the way humans use words like "snake" or "venom" reflects much (if not all) of what they have learned from their real-world experience. Louwerse (2011, 2018), with the symbol interdependence hypothesis also defends the representational power of symbols. This hypothesis suggests that the corporeal properties of the real world are encoded in both modal representations (visual, auditory, emotional, etc.) and amodal representations (symbols).

From this hypothesis, Martínez-Huertas et al. (2021) studied whether neural network models receiving LSA information as input could detect the dimensional (valence and arousal) and discrete (happiness, sadness, fear, disgust, anger) emotional content of a sample of words. They found satisfactory results in both cases. These findings support the ideas of Landauer (1999) and Louwerse (2011, 2018) and highlight that the human's corporeal relationship with the external world is encoded in language. Therefore, extracting embodied features from symbols (in this case, emotionality) is possible using computational models such as LSA.

In congruence with Landauer (1999), Louwerse (2011, 2018) and Martínez-Huertas et al. (2021) in the present study we test the ability of LSA to evaluate the emotional content of short written texts. Based on the results obtained by LSA evaluating the emotional content of words (Martinez-Huertas et al., 2021), we expect LSA to be able to evaluate the emotionality of a sample of short written texts (tweets) as a human would do. Should this be found, this study would be providing evidence in agreement with the positions that emphasize the representative power of symbols.

2. Method

2.1. Tweet sampling and vectorization in latent semantic space

When writing a tweet, any Twitter user (Twitter, Inc, 2021) can use the hashtag (#) to indicate a certain topic. Therefore, we considered that the hashtag could define the emotional category and valence expressed by a tweet. These are the dependent variables or criterion of our study. For

example, tweets tagged with *#feliz* (*#happy* in Spanish) are texts which are expected to express happiness (VD_1) and have a positive valence (VD_2). We operationalized the criterion in this way to automate the data collection process as much as possible.

Based on these assumptions, using R software (R Core Team, 2021) we connected to the Twitter API and collected N = 3,457 tweets of which 1,031 were of positive valence ($n_{\#happy}$= 1031) and 2,426 were of negative valence ($n_{\#sad}$= 482, $n_{\#fear}$= 826, $n_{\#anger}$= 17, $n_{\#disgust}$= 1,101). Then, we used the software Gallito Studio (Jorge-Botana et al., 2013) to generate the 300 dimensions for the LSA semantic space and vectorize the tweets (i.e., project them in the latent semantic space). The semantic space was created from a journalistic corpus composed of 150,802 documents and 23,835 words. The data came from the Spanish newspapers of El País and El Mundo with articles written in 2019.

2.2 Neural Network

Once the tweets were vectorized in the semantic space, we used neural network models (Martínez-Huertas et al., 2021) trained to propagate the emotionality of the words using the latent semantic space information. These models, in the training phase, using human rated-words from the data sets of emoFinder (Fraga et al., 2018) were used to relate word vectors (LSA information) with their emotional features (degree of happiness, fear, disgust, etc. rated by humans). Based on these associations, in the testing phase, the models have to infer the emotionality of a set of no rated words.

In this study, the models receive the LSA coordinates of the tweet as input and gives the degree of happiness, sadness, fear, anger and disgust contained on the text on a scale of 0 to 1 as output. Then, we classified each tweet in an emotional category and valence based on the maximum score in the emotional output. For example, a tweet with the highest score in disgust was classified as disgusting (categorical approach) and negative (valence).

3. Results

3.1. Emotional category

In the case of the emotional category, 38% of the tweets were correctly classified according to their hashtag (i.e., criterion). These results are shown in Table 2. As can be noted from the table, the only emotional category with an acceptable hit rate is #happy, with 74% hits.

Table 2. Confusion matrix for the emotional category classification.

		Criterion					
		Disgust	**Anger**	**Happiness**	**Fear**	**Sadness**	**% Hits**
Prediction	**#Disgust**	173	1	40	90	55	16
	#Anger	163	6	38	117	54	35
	#Happiness	318	6	758	254	146	74
	#Fear	140	1	39	188	54	23
	#Sadness	307	3	156	177	173	36
	N	1101	17	1031	826	482	

Note. The main diagonal reflects the hits for each category. The sum of the diagonal divided by the total number of tweets reflects the global hit rate of 38%.

3.2. Emotional valence

For the emotional valence the predictions improved, with 71% of the tweets being correctly classified. As shown in Table 3, positive and negative valence have acceptable hit rates. Positive tweets are those in the emotional category "happy". Accordingly, the hit rate remained at 74% for this category. On the other hand, the hit rate of the negative ones increased to 70%.

Table 3. Confusion matrix for the emotional valence classification.

		Criterion		
		Negative	**Positive**	**% Hits**
Prediction	**Negative**	1702	273	70
	Positive	724	758	74
	N	2426	1031	

Note. The main diagonal reflects the hits for each valence. The sum of the diagonal divided by the total number of tweets reflects the global hit rate of 71%.

4. Conclusions

As can be seen in the results section, we found large differences between the hit rates on the categorical and valence approaches (38% vs 71%). In order to find out what could explain these differences we further explored the criterion variables in order to assess their quality. Specifically, we studied whether the hashtag could really represent the emotionality of a tweet. To address this, we searched the most frequent words used with each of the hashtags. The results for two of the hashtags (#disgust and #happy) are shown in Figure 2.

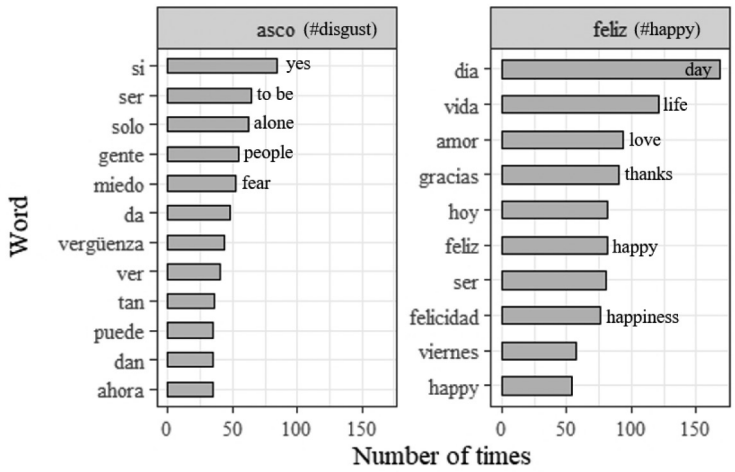


Figure 2. Most frequent words around hashtags #feliz (#happy) and #asco (#disgust).

Analyzing Figure 2, it seems that *#happy* really seems to express happiness. Words with positive connotations such as "love", "thanks", "happy" and "happiness" are the most frequent in tweets tagged with *#happy*. However, "yes", "to be", "alone" and "people", which are the most frequent words around the hashtag *#disgust*, do not seem to have a connotation of disgust. The rest of the negative hashtags worked in the same way. We also noticed an overlap between the

words used in the different negative hashtags. For example, "fear" is one of the most frequent words around *#disgust*, which is an added difficulty for the analysis we conducted in this study. When all the negative hashtags are taken together conforming the negative valence dimension the problems described above disappear. Words like "terror", "panic", "sad" and "alone", which have a negative connotation, are the most frequent words used in tweets with negative valence. This could explain the differences found between the results in the categorical and valence approaches.

It is also worth mentioning that the neural network models used in this study were trained to propagate the emotionality of words, not the emotionality of tweets. In spite of this, the neural network models based on the latent semantic space information achieved a considerably high hit rate (71%).

The limitations described above can be overcome in future research by means of training neural network models using tweets and improving the criterion validity (e.g., validating a sample of tweets by human judgement). Considering this, the results obtained in this study are promising and indicate in some way that words, apparently only abstract symbols, can capture much more information than is commonly believed.

References

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Jorge-Botana, G., Olmos, R., & Barroso, A. (2013). Gallito 2.0: A natural language processing tool to support research on discourse. En *Proceedings of the 23th Annual Meeting of the Society for Text and Discourse*, Valencia, España, 16–18 Julio.

Landauer, T. K. (1999). Latent semantic analysis (LSA), a disembodied learning machine, acquires human word meaning vicariously from language alone. *Behavioral and brain sciences*, *22*(4), 624–625.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573–589.

Martínez-Huertas, J. Á., Jorge-Botana, G., Luzón, J. M., & Olmos, R. (2021). Redundancy, isomorphism, and propagative mechanisms between emotional and amodal representations of words: A computational study. *Memory & Cognition*, *49*(2), 219–234.

Quesada, J. (2007). Creating your own LSA spaces. In Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). *Handbook of latent semantic analysis*, 71–85.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998) Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*(2–3), 337–354. https://doi.org/10.1080/01638539809545031

Twitter, Inc. (2021). *Twitter*. Microsoft Store. https://www.microsoft.com/es-es/p/twitter/9wzdncrfj140?activetab=pivot:overviewtab

Long-term longitudinal data collection and analysis in highly dynamic systems using mobile crowdsensing and mobile agents

Stefan Bosse[1]

[1]*Department of Mathematics and Computer Science, University of Bremen, Germany*

Abstract

Data collection from the real world is still a challenge. Classical surveys only deliver snapshots at specific times and a small sub-set of sensors that can be accessed. Simulation depends on real-world data, too. Surveys are typically participatory and rely on models and survey plans. Crowdsensing is typically opportunistic and self-organizing. Mobile crowdsensing enables continuous monitoring of sensors. To enable continuous longitudinal data sampling, an agent-based mobile crowdsensing approach is introduced that can be coupled with agent-based simulation and digital twin methods. Two use cases show the improvement of accuracy of system observables depending on individual micro-scale behavior or calibrated sensors. The first use case addresses dynamic time-series prediction and surrogate modelling, the second use case addresses the influence of variation in segregation simulation by digital twins. Both use cases show the still existing gap between real world and virtual simulation worlds caused by insufficient or distorted sampling and missing sensor calibration.

Keywords: Mobile crowdsensing; agent-based methods, machine learning, closed-loop simulation.

E-mail: sbosse@uni-bremen.de

1. Introduction

Accessing the state of the real world, e.g., state variables and aggregate observables of society, is still a challenge. Only a small sub-set of sensors can be accessed, e.g., using a planned survey to access people's opinions. But any survey is limited to a snapshot on a longitudinal time axis. Even if surveys are performed repeatedly there are gaps in the longitudinal data dimension. This work addresses new methodologies for longitudinal data collection and aggregation that can be used for:

1) Data analysis and data mining with a statistical background;
2) Data- and event-driven simulation;
3) Automated prediction and classification using machine learning;
4) Time-series analysis and prediction.

Machine learning can be considered as a kind of simulation of real data by inter- and extrapolation. All four domains depend on the strength and statistical quality on the vertical and horizontal (longitudinal time) scale, and incremental longitudinal data sampling is still a challenge. In social sciences, surveys play an important role to get a snapshot of the real world. Typical applications of classical longitudinal surveys are (Lynn, 2009):

- Surveys of businesses
- Surveys of school-leavers, graduates or trainees
- Household panel surveys
- Birth cohort studies
- Epidemiological studies
- Social networking
- Socio-technical systems

Surveys are typically participatory and rely on models and survey plans. Crowdsensing is typically opportunistic and self-organizing. Mobile crowdsensing enables continuous monitoring of sensors, as shown in Fig. 1.

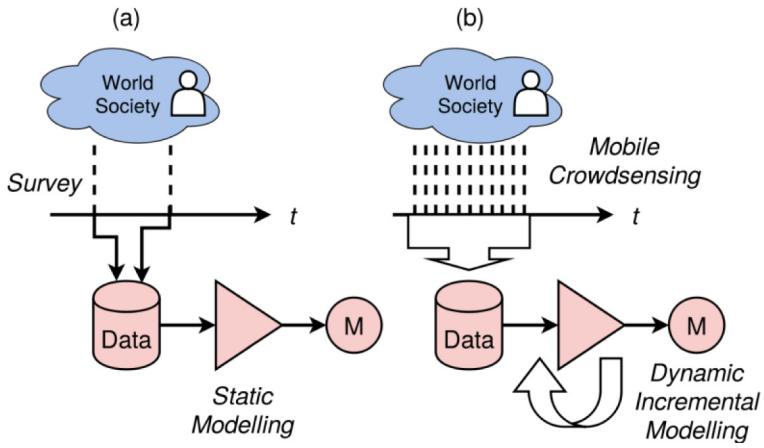


Figure 1. (a) Traditional survey-based data sampling and static modelling using participatory mechanisms (b) Continuous crowdsensing-based data-driven modelling using opportunistic mechanisms

The main issue with longitudinal sampling is the curse of dimensionality of longitudinal data (LD):

$$LD = P \times O \times L \times V \times t$$

with P: Persons/Entities, O: Occasions, L: Locations/Places, V: Sensor variables, t: time.

The sampling in time space (horizontal axis) can be performed periodically (polling), event-based, or randomized. Sampling in variable space (vertical axis) is affected by bias, fraud, distortion, noise, failure, missing data, and impurity. The errors in longitudinal data sampling are related to four major error classes contributing to coverage errors, sampling errors, non-response errors, and measurement errors, in general. The first three error contributions belong to errors that cannot be observed directly and require auxiliary variables, whilst the measurement error can be observed directly by statistical analysis. Mobile crowdsensing can help to reduce coverage, sampling, and non-response errors and to extend the data space with environmental/context sensor variables (auxiliary variables). Information (models, parameters, aggregate variables) can be derived from on-line or off-line data mining using statistical or approximating machine learning methods. Additionally, or exclusively, on-line or off-line simulation can be performed to derive aggregate variables (details in Bosse et al., 2019A).

2. Method

To overcome limitations introduced by classical surveys and longitudinal data sampling, a unified approach is used that is based on agents connecting the real-world environment with simulation in a bidirectional way. This concept is composed of different methodologies: Agent-based Modelling (ABM), Agent-based Simulation (AB S), Agent-based Computation (ABC, mobile software agents), Mobile agent-based Crowdsensing (MCWS), and Machine Learning used for Surrogate Modelling (SM). The mobile crowdsensing addressed in this work is characterized by (and shown in Fig. 2):

- Event-driven or request-reply-based sampling;
- Usage of mobile agents for sensor sampling (mobile devices);
- Performing micro surveys (dynamic/conditional scripts) via chat dialogs and environmental sensing (e.g., in a smart city context, see Alvear et al.,2018).

Two agent super classes are used: physical simulation agents (red) and computational software agents (blue, simulation and real-world) interacting with each other (Bosse et al., 2019A). The advantage of coupling physical agents in the simulation with computational agents that can perform real-world environmental and crowdsensing is the capability to synchronize the simulation world continuously and to create digital twin agents in the simulation with real-world data (e.g., by ad-hoc micro surveys). Another advantage of real-world coupled simulation is the synchronization of the simulation with real-world data and performing simulation to get future snapshots, i.e., implementing a future time machine. ABC crowdsensing can be used: 1. To update simulations in real-time, i.e., introducing variance by digital twins, 2. For fork simulation runs with time-compressing speed-up, and 3. Creating simulation snapshots for future world evolution, e.g., used for weather forecasting. Virtual sensors implemented by mobile or stationary agents are a central part of longitudinal data sampling and data reduction methodology (including calibration).

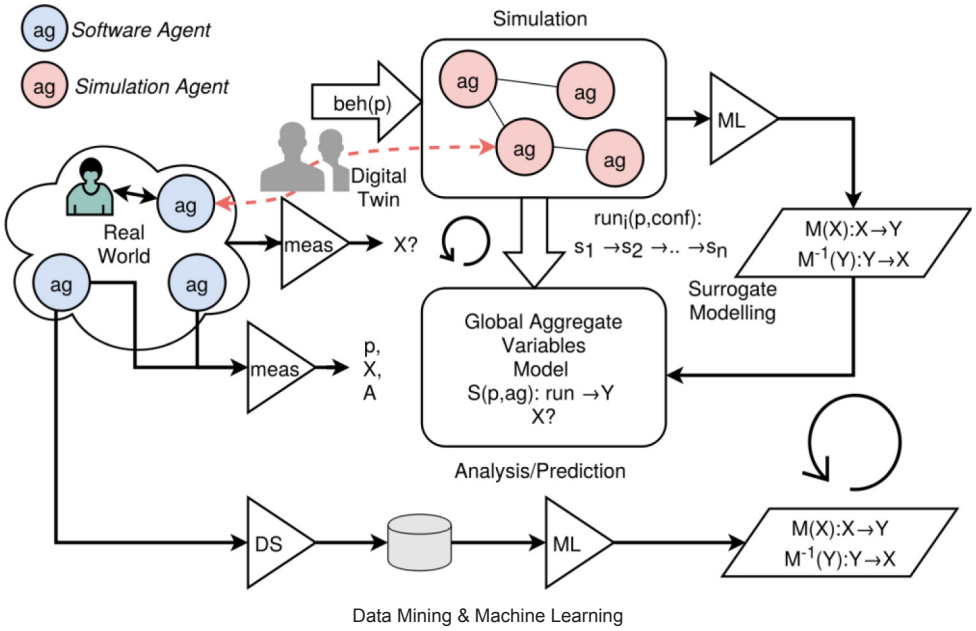


Figure 2. Unified Agent Methodology for longitudinal data mining, modelling of complex systems, and simulation

3. Results

3.1. Use case: pandemic simulation and time-series prediction

The goal of the first use case is a dynamic time-series prediction of infection cases in pandemic situations with the following methodologies (details in Bosse, 2021):

1) Data mining of already existing institutional longitudinal data and machine learning (time-series extrapolation),
2) Surrogate modelling of ABS using data from simulation and auxiliary data from mobile crowdsensing (crowd behavior, decision-making, and opinions) / Simulation seed with data from 1.

There is an institutional data mining and machine prediction using data from the Robert Koch Institute (notifications of weekly infection cases), and a time-series prediction by an LSTM-ANN. The real-world sensor data is biased and distorted/uncalibrated due to unknown test sampling distributions over time. It is not possible to derive a time-series prediction model that can predict the future development of the aggregate observable with meaningful accuracy. It is possible to predict future development using crowd-driven simulation and surrogate modelling with domain-partitioned parameterized statistical cellular automata simulation and time-series prediction by LSTM-ANN, as shown in Fig. 3.

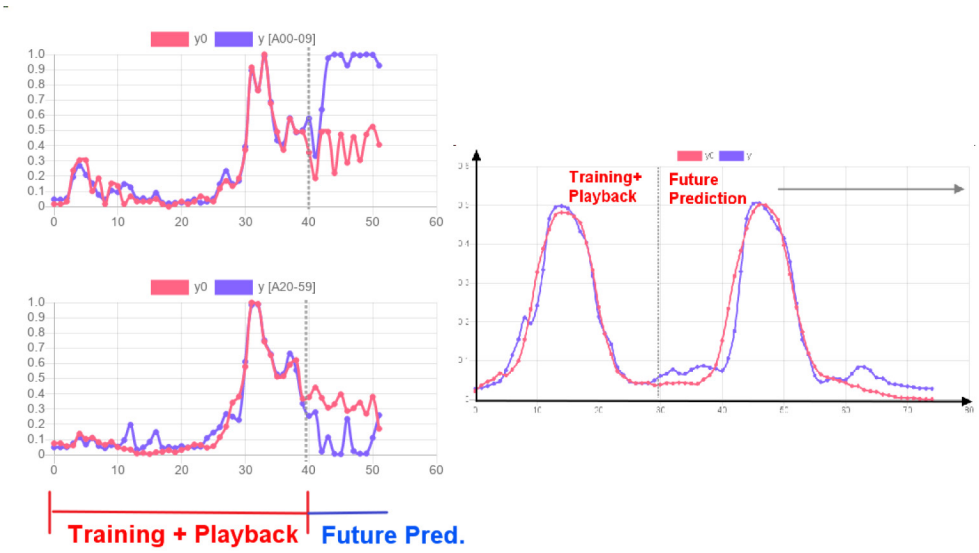


Figure 3. (Left) Time-series prediction of aggregate variable infection cases from real-world data (Right) Prediction using simulation data

3.2. Use case: simulation of segregation

The goal of the second use case is the study of segregation effects (cluster groups) with individual (variant) behavior based on mobility and social networking using the following methodologies (details in Bosse et al., 2019B):

1) Agent-based simulation with parameterized mobility and interaction models;
2) Agent-based crowdsensing performing micro surveys via mobile devices and chat dialogues finally creating digital twins introducing behavior model variance.

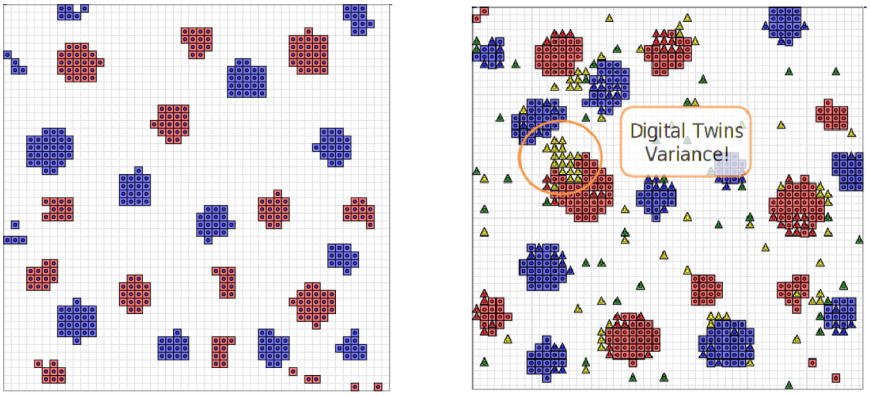


Figure 4. (Left) Results from closed segregation simulation (Right) Results from open simulation with updated real-world data (Bosse et al., 2019B)

Two simulation runs were compared, as shown in Fig 4. Firstly, a closed simulation was used with static agent behavior that poses a static segregation behavior model with typical group cluster formation of the same class. Secondly, an open simulation was used with in-line

crowdsensing and digital twins introducing behavior variance at micro level. The global outcome shows different cluster formations. Individualism had an significant impact on global aggregate variables.

4. Conclusions

Longitudinal data sampling and analysis is a challenge with respect to bias, distortion, sampling intervals, impure variables, missing calibration, dimensionality and data volume. Agent-based methods with a unified agent model features mobile crowdsensing sampling environmental and user data on a micro-scale level, tight coupling of simulation (ABS) with the real world (human-in-the-loop), incremental data collection by software agents that synchronizes simulation with the real world, while simulation snapshots and forking enables the prediction of future world evolution.

Agents can pose the following roles: physical agents in simulation, computational agents performing crowdsensing (physical sensors), and computational agents performing sensor aggregation, event detection, and data reduction (virtual sensors), finally providing continuous longitudinal data sampling.

References

Alvear, O., Calafate, C. T., Cano J.-C., & Manzoni, P. (2018). Crowdsensing in Smart Cities: Overview, Platforms, and Environment Sensing Issues, Sensors, vol. 18, no. 460.

Bosse S. (2021). Surrogate Predictive and Multi-domain Modelling of Complex Systems by fusion of Agent-based Simulation, Cellular Automata, and Machine Learning, p. 79–85, Proc. of the SIMUL 2021 Conference, The Thirteenth International Conference on Advances in System Simulation, IARIA, 3. - 7. October 2021, Barcelona, Spain, & Online.

Bosse, S., Engel, U. (2019A). Real-time Human-in-the-loop Simulation with Mobile Agents, Chat Bots, and Crowd Sensing for Smart Cities, Sensors (MDPI), doi: 10.3390/s19204356

Bosse, S., Engel, U. (2019B). Combining Crowd Sensing and Social Data Mining with Agent-based Simulation using Mobile Agents towards Augmented Virtuality, Proc. of the Social Simulation Conference, 24-27.9.2019, Mainz, Germany.

Lynn, P. (2009). Methods for Longitudinal Surveys, in Methodology of Longitudinal Surveys, p. 1–19.

On the nature of group factors in bifactor structures

Alicia Franco-Martínez[1,2], Daniel Ondé[2], Jesús M. Alvarado[2]

[1]*Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, Spain*

[2]*Departamento de Psicobiología y Metodología en Ciencias del Comportamiento, Universidad Complutense de Madrid, Spain*

Abstract

The bifactor model is commonly used to assess the degree to which a measure can be considered essentially unidimensional (in the presence of some degree of multidimensionality). In this work, we have changed the focus of attention, evaluating the degree to which group factors can be treated as subscales (in presence of a certain degree of common variance to all the items). This type of approach has not received much attention to date. We conducted a Monte Carlo simulation study in which we manipulated the general factor to present limited common variance (i.e., factor loadings of .10, .30 and .50). The results show an adequate parameter recovery of the group factors in most conditions, even when the factor loadings of the general factor are extremely low. Omega hierarchical is less affected by the general factor strength (or weakness) than parameter recovery and explained common variance. To consider that a group factor is an essentially unidimensional subscale must present factor loadings > .60 and, simultaneously, the factor loadings of the general factor must be around .30.

Keywords: Bifactor model; general factor; group factors; explained common variance, omega hierarchical.

E-mail: alicia.franco@estudiante.uam.es

1. Introduction

The bifactor model is an analytic strategy that has received much attention in recent years to evaluate the internal structure of data collected through tests. This strategy consists of dividing the responses of each item into two sources of variation: one that reflects common variance to all items and another that reflects variance that is specifically due to a cluster of items within a test (for example, a content domain). In bifactor models, the first source of variation is referred to as the general factor (F_{Gen}) and the second source of variation as group factors (F_{Gr}).

According to Reise et al. (2010), the need to decompose the observed variance of a test into two sources is because researchers generally write items to evaluate a single construct (general or unitary construct), while at the same time they need items with heterogeneous content to represent the construct. For this reason, most applications of the bifactor model have focused on evaluating the degree to which the different measures evaluated can be considered as essentially unidimensional (in presence of a certain degree of multidimensionality). To assess whether a measure can be considered as essentially unidimensional, various model-based indices have been proposed (for example, Reise, 2012; Rodriguez et al., 2016a, 2016b). The most popular criteria are currently the following: explained common variance (ECV) and the hierarchical omega coefficient (ω_H) of the general factor (> .60 and > .70, respectively).

When the general factor of a bifactor model shows an adequate level of factor determination (i.e., ECV > .60 and ω_H > .70), group factors can only explain a limited part of the total variance, even being a residual (meaningless, uninterpretable) part of the model. However, although there is an important tradition in Psychology related to unitary construct evaluation (i.e., essentially unidimensional F_{Gen}), this scenario is not the only possible one. In several studies, multidimensional constructs are still evaluated using correlated factor models in which the empirical correlation between factors is not very high. Other studies evaluate general factors as method factors (for example, acquiescence and faking behavior). Therefore, it is worth asking what happens to the group factors when the variance common to all the items is limited. This issue has not received much attention to date.

Our main goal is to evaluate the degree of determination of group factors in the presence of common variance running among all items in bifactor structures. In doing this, the specific goals are: (1) to study the relationship between model-based indices (ECV and ω_H) applied to the group factors and the parameter recovery, (2) to assess whether the goodness-of-fit indices are good predictors of parameter recovery, and (3) to assess the impact of different conditions: sample size, data distribution, number of items per group factor, and strength of the general factor.

2. Method

In this work, we conducted a simulation study based on the conditions proposed by Reise et al. (2013). The model structure from which we simulated our data was a bifactor with one general factor (F_{Gen}) and three group factors (one with strong loadings from .60 to .80, F_{Gr1}, one with moderate loadings from .45 to .55, F_{Gr2}, and one with weak loadings from .15 to .35, F_{Gr3}). We also varied the number of items of the F_{Gr3} from 4 to 7, while the other two group factors remained fixed to 6 for the F_{Gr1} and 4 items for the F_{Gr2}. This led us to specify four models regarding the PUC value (percentage of uncontaminated correlations): from .691 to .703 (see Reise et al., 2013). Our chief goal was to explore the consequences of working with a weak F_{Gen}, so we simulated conditions of factor loadings from .10 to .50. As opposed to Reise et al. (2013), where the data were continuous and normally distributed, here we generated discrete data (5-point response categories) with three distributions (symmetrical, moderately skewed and extremely skewed). In total, 4 conditions of PUC, 5 conditions of F_{Gen} factor loadings and 3 distributions:

60 simulated conditions (fixed effects), each replicated 1,000 times. We generated a random sample size between 100 and 1,000 observations for each replication.

Once generated the datasets, we applied a classic bifactor analysis with the DWLS (Diagonalized Weighted Least Square) estimation method, which is adequate for categorical data. The dependent variables extracted from each bifactor analysis were the following:

For parameter recovery we calculated the relative bias (RB): one for the loadings of the F_{GEN}, and three for the F_{Gr} expressed as:

$$RB = \frac{\lambda_{(e)} - \lambda_{(t)}}{\lambda_{(t)}} \times 100,$$

being $\lambda_{(e)}$, the estimated loading, and $\lambda_{(t)}$, the true simulated loading.

Regarding the model-based indices proposed in Rodriguez et al. (2016a, 2016b), we calculated the Explained Common Variance (ECV) for the F_{Gen},

$$ECV_{Gen} = \frac{\sum \lambda_{Gen}^2}{\sum \lambda_{Gen}^2 + \sum \lambda_{Gr1}^2 + \sum \lambda_{Gr2}^2 + \sum \lambda_{Gr3}^2}.$$

and for the three F_{Gr},

$$ECV_{Gr} = \frac{\sum \lambda_{Gen}^2}{\sum \lambda_{Gr}^2 + \sum \lambda_{Gr1}^2 + \sum \lambda_{Gr2}^2 + \sum \lambda_{Gr3}^2}.$$

And the Omega hierarchical, for the F_{GEN} (ω_H):

$$\omega_H = \frac{(\sum \lambda_{Gen})^2}{(\sum \lambda_{Gen})^2 + (\sum \lambda_{Gr1})^2 + (\sum \lambda_{Gr2})^2 + (\sum \lambda_{Gr3})^2 + \sum \psi},$$

being ψ the error variances, and for the three F_{Gr} (ω_{HS}):

$$\omega_{HS} = \frac{(\sum \lambda_{Gr})^2}{(\sum \lambda_{Gen})^2 + (\sum \lambda_{Gr1})^2 + (\sum \lambda_{Gr2})^2 + (\sum \lambda_{Gr3})^2 + \sum \psi}.$$

Lastly, to evaluate the fit of the models, we calculated the most frequently used goodness-of-fit indices: the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), and the comparative fit index (CFI), for which values below .08, below .06 and over .95, respectively, were indicators of adequate fit (Hu & Bentler, 1999).

3. Results

3.1. Parameter recovery and model-based indices

For the sake of simplicity, we will only represent the results for the conditions where F_{Gen} loadings are .10, .30 and .50, F_{Gr3} has 4 items and symmetrical data distribution (the remaining

conditions did not sensitively differ from these). *Figure 1* shows the average RB for each factor across the different scenarios.

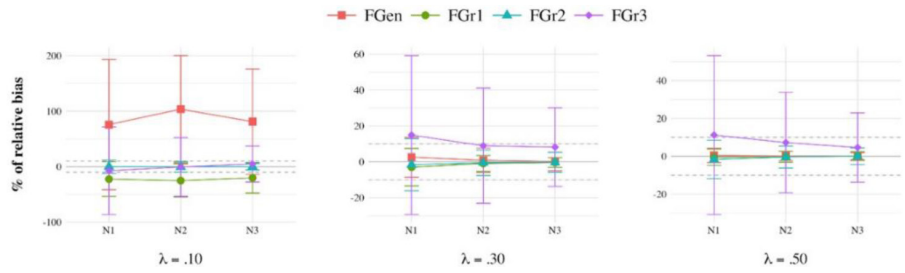


Figure 1. RB average (%) between simulated and estimated factor loadings in the general factor (FGen) and the three group factors (FGr1, FGr2, and FGr3), across simulated strength for the general factor (λ) and sample size (N1: 100-400, N2: 401-700, and N3: 701-1000).

As we expected, when the factor loadings of the general factor were .30 or .50, the parameter recovery was adequate for every factor, except for the F_{Gr3}, since it was the weakest one (4 items and loadings from .15 to .35). However, this pattern changed abruptly when F_{Gen} loadings were .10. Now, the F_{Gen} is extremely overestimated, reaching RBs' means of 100%, whereas the F_{Gr1} is relatively underestimated.

The ECVs, illustrated in Figure 2, are in line with the performance of RB. Additionally, we notice that when F_{Gen} loadings are .30 or .50, the ECV for the F_{Gr1} is systematically overestimated, while the ECV for the F_{Gr2} is systematically underestimated. The weakest factor (F_{Gr3}) shows ECV average values which are very similar to the simulated ECV values, reflecting poor explained common variance. Only under some conditions (i.e., factor loadings of .30 in the general factor and N > 400), does F_{Gr1} reach ECV values above the .60 recommended by Reise et al. (2013). The rest of group factors cannot reach this value in any condition.

The ω_H, showed in *Figure 3*, follows a similar pattern to the ECVs, but when the F_{Gen} is weak, ω_H is not as affected as ECVs or RBs. Only when factor loadings in F_{Gen} are .30 or .50 (in some solutions when factor loadings in F_{Gen} are .10), does F_{Gr1} reach ω_H values above the .70 recommended by Reise et al. (2013). The rest of group factors cannot reach this value in any condition, although F_{Gr2} comes close when F_{Gen} factor loadings are .10.

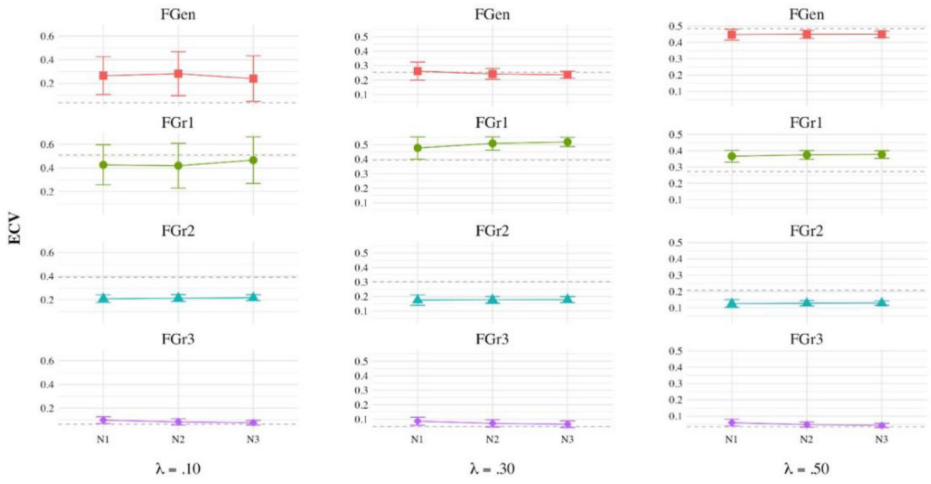


Figure 2. ECV average for the general factor (FGen) and the three group factors (FGr1, FGr2, and FGr3), across simulated strength for the general factor (λ) and sample size (N1: 100-400, N2: 401-700, and N3: 701-1000). The dotted grey lines represent the true simulated ECV values.

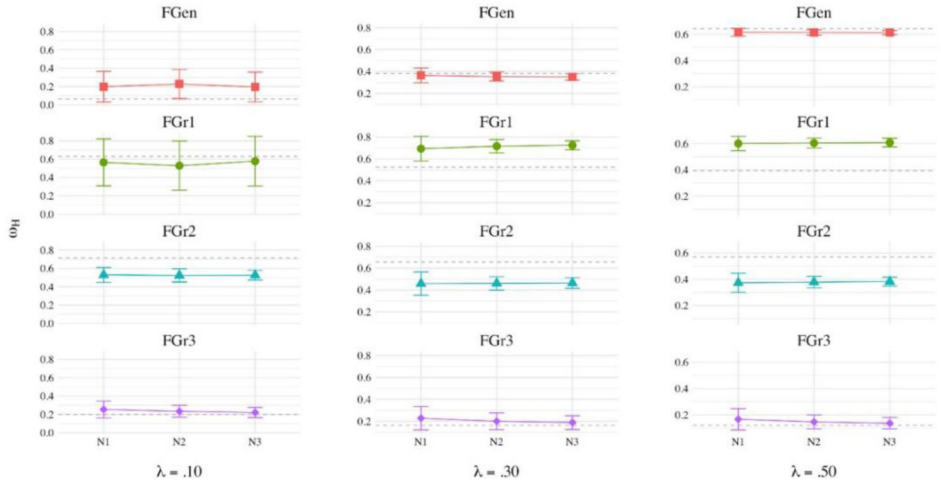


Figure 3. Values of ω_H for the general factor (FGen) and the three group factors (FGr1, FGr2, and FGr3), across simulated strength for the general factor (λ) and sample size (N1: 100-400, N2: 401-700, and N3: 701-1000). The dotted grey lines represent the true simulated ECV values.

3.2. Parameter recovery and goodness-of-fit

All goodness-of-fit indices suggested an excellent fit, regardless of simulated conditions. Even when loadings for the F_{Gen} were .10, and this factor was highly overestimated (as mentioned before), goodness-of-fit was preserved.

4. Conclusions

Our results show that only when the general factor have moderate-low average factor loadings (i.e., .30 - .50) can a group factor obtain adequate model-based indices (ECV > .60 and ω_H > .70). These values are considered good indicators of unidimensionality and reliability (Reise et al., 2013) so this type of group factor could be interpreted meaningfully and used as a subscale (i.e., using unit-weighted composite scores). However, the factor loadings of this type of group factor must be high (in our case, between .60 and .80 for F_{Grl}), and the factor loadings of the general factor should be neither too high nor too low.

With high factor loadings in the general factor, low common variance still needs to be explained by group factors. When the correct model includes a weak general factor, apparently the bifactor estimation method tries to capture a strong general factor. To gain this variance, the bifactor takes some from the strongest group factor, underestimating it. We have noticed that, while ECVs are also sensitive to this problem, ω_H is not: although the loadings for the general factor are being overestimated, the empirical ω_H is still a good estimation of the true . To this, it must be added that extremely low factor loadings are difficult to estimate, resulting in high standard errors on many occasions (see, for example, Ondé & Alvarado, 2020). In these situations, even the bifactor model can present serious estimation problems. Another discovery in this study was that none of the goodness-of-fit indices selected suggested a poor fit when these estimation problems arose. In the light of this performance, we encourage applied researchers not to trust the whole model selection decision over the goodness-of-fit indices: reporting an excellent fit may be masking other problems of recovery bias.

Finally, the parameter recovery was more stable as the sample size increased (we can see this in the error bars of the three figures shown in this work). However, skewing the data distribution

or increasing the number of items in the F_{Gr3} did not affect the results already presented. The number of items per factor could be a more influential variable if the factor loadings of F_{Gr3} were higher (like in F_{Gr2}, for example), improving the ECV and the ω_H average values. To verify this possibility, and to check and expand on the conclusions exposed, additional simulation studies are required.

References

Hu, L. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Ondé, D., & Alvarado, J. (2020). Reconsidering the Conditions for Conducting Confirmatory Factor Analysis. *The Spanish journal of psychology*, *23*, E55. https://doi.org/10.1017/SJP.2020.56

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research*, 47(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and psychological measurement*, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of personality assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137. https://doi.org/10.1037/met0000045

Analysis of the psychometric properties of a social self-efficacy scale in adolescents in Spain, from a gender perspective.

Vanesa Salado[1], Sara Luna[1], Mª Carmen Moreno[2] y Francisco Rivera[1]

[1]*Department of Experimental Psychology, University of Seville, Spain,*

[2]*Department of Developmental and Educational Psychology, University of Seville, Spain*

Abstract

Self-efficacy is the beliefs of an individual about their abilities to perform actions that allow them to achieve a desired performance that will regulate their behavior. The main objective of this study is to analyze the psychometric properties of an 8-item scale inspired by the one developed by Muris, P. (2001). The sample consisted of 5773 young people between the ages of 11 and 18. The responses were collected through a multistage sampling stratified by conglomerates. The psychometric properties of the scale related to the reliability evidence showed acceptable alpha coefficients for both the global sample and those segmented by sex. The analyses of the evidence of internal validity referring to the structure of the questionnaire showed the unidimensionality of the instrument as the best possible structure. Regarding the evidence of validity related to other variables, the perception of the friend support scale and the perception of the classmate support scale showed a moderate correlation with the measure of self-efficacy, being slightly higher in the group of boys than in the group of girls when the sample was segmented. Future research can segment the sample into age groups to ascertain how this scale behaves at an evolutionary level.

Keywords: Self-efficacy, Spanish adolescents, evidence of reliability, evidence of validity.

Funding: This study has been supported by the project entitled "Barometer of Childhood and Opinion of Adolescents" developed through collaboration between the Spanish Committee of UNICEF and the University of Seville. The methodological development of the study and the questionnaire itself were proposed by the same research team during 2017, thanks to competitive research funding by UNICEF (ref. 3122/0294) obtained by the research team. The continuity of the project was possible thanks to a contract 68/83 (ref. 3592/0991) with UNICEF for the systematization and implementation of the barometer nationally in 2019 (first edition) and again in 2020 (second edition), both at regional level (thanks to an ERDF project) and national level (another 68/83 contract with UNICEF), as well as its longitudinal monitoring granted

by the Ministry of Economy, Knowledge, Business and Universities (Ministry of Economy, Knowledge, Business and University) (ERDF Project, ref. US-1266024). In addition, this work had the support of the Ministry of Science, Innovation and Universities (Ministry of Science, Innovation and Universities) through a scholarship that Vanesa Salado Navarro received in the University Teacher Training Program (ref. FPU19 / 00023).

E-mail: saladonavarrovanesa@gmail.com

1. Introduction

Social self-efficacy is the individual's ability to develop healthy relationships with others (Schunk & Pajares, 2009). Social self-efficacy is a subdimension of general self-efficacy as proposed by Bandura et al. (1999) which, together with academic self-efficacy and self-regulatory efficacy form a measure to ascertain the beliefs that people have about their abilities to organize and carry out actions that affect their lives (Bandura, 1997, 2006).

Social self-efficacy is a predictor of pro-social behaviors (Wentzel, 2014) and better quality of relationships between young people (Raskauskas et al., 2015). In this sense, healthy social relationships promote greater coping with the daily situations of adolescents and, as a result, greater reinforcement of social self-efficacy (Benight & Bandura, 2004).

With regard to the gender perspective, it has been found that girls between the ages of 10 and 12 report greater social self-efficacy than boys (Coleman, 2003). However, in the ages between 13 and 18 years there are no gender differences in this variable (Bacchini & Magliulo, 2003).

From the psychometric point of view, the understate of social self-efficacy has scarcely been evaluated. Muris (2001) constructed a questionnaire with the three subdimensions proposed by Bandura (1999) and evaluated the evidence of reliability and validity of each of them. For the scale of social self-efficacy, Muris (2001) reported adequate internal consistency and the one-dimensionality of the instrument with the elimination of item 8. However, the studies by Zullig et al. (2011) did not eliminate item 8 to consider the one-dimensionality of the instrument. On the other hand, taking into account the relationship of this scale with other variables, as explained at the beginning of the introduction, is essential to ascertain its predictive power and, therefore, the evidence of external validity in relation to other variables.

Therefore, this study aims to analyze the psychometric properties of the 8-item social self-efficacy subscale inspired by the self-efficacy scale developed by Muris (2001) in Spanish adolescents from a gender perspective.

2. Method

2.1. Study Design and Participants

This research was conducted by the OPINA Barometer (Barómetro OPINA) project, carried out by a research team from the University of Seville (Spain) in collaboration with UNICEF. The transversal study evaluated adolescents' opinions and concerns, their knowledge about sociopolitical issues, their implication as citizens, as well as their wellbeing. Participants were selected using multistage random sampling stratified by conglomerates. For the present study, 5773 participants were selected who were between 11 to 18 years of age, of which 47.7% were boys and 52.3% were girls. Data was collected through an anonymous online questionnaire administered at school and with informed consent from the school staff, parents/legal guardians, and students. The questionnaire was approved by the University of Seville's Ethical Committee (Comité Ético de Experimentación de la Universidad de Sevilla) in accordance with the standards of the 1964 Helsinki Declaration and its subsequent modifications.

2.2. Instruments

The instrument used was the Opinion Barometer of Childhood and Adolescence (Barómetro de Opinión de la Infancia y la Adolescencia) (Moreno et al., 2017). In addition to gender, the following variables were selected from the questionnaire:

- Social self-efficacy was inspired by the subdimension of social self-efficacy developed by Muris (2001). This subscale had 8 items. Some of the items were: "How well can you express your opinions when other classmates disagree with you?", "How well can you become friends with other children?", "How well can you have a chat with an unfamiliar person?", " How well can you work in harmony with your classmates?" etc. The response options oscillated between 1 (not at all) and 5 (very well).
- Friend support was evaluated through the friend subscale of the Multidimensional Scale of Perceived Social Support (MSPSS; Zimet et al., 1988). This subscale consists of four items: "My friends really try to help me", "I can count on my friends when things go wrong", "I have friends with whom I can share my joys and sorrows", and "I can talk about my problems with my friends". Responses ranged on a 7-point Likert scale from 1 (very strongly disagree) to 7 (very strongly agree).
- Classmate support was evaluated using the subscale of classmates from the Perceived Support from Teachers and Classmates Scale (Torsheim et al., 2000). Classmate support was measured by three items: "The students in my class enjoy being together", "Most of the students in my class are kind and helpful", and "Other students accept me as I am". The responses were recorded on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

2.3. Data Analysis

Firstly, the Student t-test was employed to examine mean differences in the examined variable and their items across gender, with the estimation of effect size through Cohen's d - small effect for values around 0.30, moderate effect for values between 0.30 to 0.50, and strong effect when values were equal to or higher than 0.50 -. (Cohen,1988). On the other hand, internal consistency was estimated through Cronbach's alpha, accepting a value of at least 0.70 for group comparisons (Nunnally and Bernstein, 1994). In addition, the correlations of each of the items were established with the test for global and segmented samples.

A confirmatory factor analysis of the 8-item scale was performed to determine the validity evidence regarding the structure of the questionnaire. Chi-square ($\chi 2$) was used to evaluate the fit of the model in the global sample and segmented by gender as well as other fit indices: Comparative Fit Index (CFI), with values greater than 0.90 considered acceptable and greater than 0.95 were considered excellent; the root mean square error of approximation (RMSEA) and the residual standard mean square root (SRMR). For these indices, values close to or less than 0.08 and 0.05 were considered indicators of an acceptable model fit, respectively. In addition, a cut-off point of a factorial load (β) greater than 0.30 was taken into account to preserve an item (Vega et al., 2019). Finally, to verify the evidence of the external validity of the scale and its relationship with other variables, the Pearson correlation coefficient was used. The correlations were repeated in the gender-segmented sample and the differences between the correlations were interpreted using Fisher's Z and Cohen's Q analyses for effect size (Cohen, 1988).

IBM SPSS Statistics 26.0 was used with a 95% confidence level for all analyses except the confirmatory factor analysis, which was performed using the Jasp 0.14 statistical program.

3. Results

3.1. Mean comparison analysis

Table 1 results from the student's t test showed differences between boys and girls in social self-efficacy and in items 2, 3 and 5, with the scores obtained for boys being higher than for girls with negligible effect size.

Table 1. Comparison analysis of means by gender in the Social self-efficacy scale.

	Descriptive statistics				Significance tests and Effect size
	$\bar{x}$	*SD*	$\bar{x}$	*SD*	
	Boys		Girls		
Social self-efficacy	3.64	.80	3.59	.80	$t_{(5771)} = 2.31$, $p = .021$; $d = .06$
Item 1	3.51	1.25	3.45	1.28	$t_{(5771)} = 1.86$, $p = .061$; $d = .05$
Item 2	3.82	1.15	3.74	1.19	$t_{(5771)} = 2.68$, $p = .007$; $d = .07$
Item 3	3.38	1.24	3.27	1.30	$t_{(5771)} = 3.08$, $p = .002$; $d = .08$
Item 4	3.69	1.06	3.74	1.09	$t_{(5771)} = -1.48$, $p = .065$; $d = .05$
Item 5	3.51	1.23	3.41	1.28	$t_{(5771)} = 2.93$, $p = .003$; $d = .08$
Item 6	3.68	1.19	3.60	1.24	$t_{(5771)} = 2.59$, $p = .010$; $d = .06$
Item 7	3.85	1.07	3.83	1.10	$t_{(5771)} = .645$, $p = .519$; $d = .02$
Item 8	3.66	1.27	3.67	1.27	$t_{(5771)} = -.092$, $p = .927$; $d = .01$

Note: SD; Standard Deviation.

3.2. Study of score reliability

Cronbach's alpha for the social self-efficacy scale was 0.82: 0.83 for the sample of boys and 0.81 for the group of girls, meeting the criteria defined *a priori* of alpha greater than 0.7. The item-total correlation analyses of the scale showed high item-total correlations in the total sample except with the item 8 which were moderate. In addition, in no case, would the reliability of the instrument improve by suppressing some of the items that compose it.

3.3. Analysis of evidence of internal validity referred to the structure of the questionnaire and analysis of evidence of external validity referring to the relationship with other variables

Table 2 presents the absolute goodness-of-fit indices (Chi-square), as well as the approximate ones (CFI, RMSEA, and SRMR) for the social self-efficacy model in the global sample and by gender. All the indices showed excellent goodness of fit of the data to the model. In addition, the standardized coefficients were greater than 0.5, except for item 8 which had a load towards the factor of 0.29.

Table 2. Goodness-of-fit indices for the global and segmented samples by gender.

	Global	Boys	Girls
$\chi^{2\,a}$ / ^{b}df	43.24	16.32	29.03
NNFI [c]	.91	.93	.89
CFI [d]	.94	.95	.92
IFI [e]	.94	.95	.92
RMSA [f] (CI 95%)[g]	.08	.07	.09
SRMS [h]	.04	.03	.05

Note: [a] χ^2, chi squared; [b] *df*, degree of freedom; [c] NNFI, non-normed Fit Index; [d] CFI, comparative fit index; [e] IFI, incremental fit index; [f] RMSA, root mean squared error; [g] CI, confidence interval; [h] SRMR, standardized room mean squared residual.

Finally, the analyses referring to the relationship with other variables showed positive and moderate correlations in most of the items between the self-efficacy scale and the scales of perceived support of friends and classmates, in the global sample. When segmented by gender, the correlations decreased slightly but remained positive.

4. Conclusions

The main objective of this research was to examine the evidence of reliability and validity of the social self-efficacy scale, developed by Muris (2001 in Spanish adolescents.

First, the results obtained in the mean comparisons showed higher scores for boys in terms of social self-efficacy than for girls, unlike the study with adolescents by Zullig et al. (2011), which found a higher score for girls.

From the point of view of the evidence of reliability, the scale of social self-efficacy showed a satisfactory internal consistency both in the global sample and when segmented by gender, as reflected in the studies of Muris (2001) and Zullig et al. (2011) for this sub-dimension. In relation to the validity analyses related to the structure, all the items satisfactorily loaded a factor in the global and segmented sample with loads greater than .50, except in item 8 - "How well do you succeed in preventing quarrels with other children?" - with a load of .29. Although in the research of Muris (2001) this item was eliminated for not exceeding .37, we consider that greater exploration of the sample is necessary to make that decision, since all the adjustment indicators showed one adequate index to consider the one-dimensional common structure.

Finally, the relationships established between self-efficacy and other variables of adolescent development such as the perceived support of classmates and friends showed a positive association with these two study variables. Research confirms the importance of social self-efficacy in developing adequate supportive relationships and avoiding stressful situations (Caprara et al. 2010; Barchia & Bussey, 2010).

References

Bacchini, D., & Magliulo, F. (2003). Self-image and perceived self-efficacy during adolescence. *Journal of youth and adolescence, 32*(5), 337–349.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bandura, A., Pastorelli, C., Barbaranelli, C., & Caprara, G. V. (1999). Self-efficacy pathways to childhood depression. *Journal of Personality and Social Psychology, 76*, 258–269.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science,1,* 164–18.

Barchia, K., & Bussey, K. (2010). The psychological impact of peer victimization: Exploring social cognitive mediators of depression. Journal of Adolescence, 33, 615–623.

Benight, C. C., & Bandura, A. (2004). Social cognitive theory of posttraumatic recovery: The role of perceived self-efficacy. *Some Behaviour research and therapy*, *42*(10), 1129–1148.

Caprara, G. V., Gerbino, M., Paciello, M., Di Giunta, L., & Pastorelli, C. (2010). Counteracting depression and delinquency in late adolescence: The role of regulatory emotional and interpersonal self-efficacy beliefs. *European Psychologist*, *15*, 34–48.

Cohen, J. (1988). *Statistical power analysis Jbr the behavioral*. Sciences. Hillsdale (NJ): Lawrence Erlbaum Associates, 18–74.

Coleman, P. K. (2003). Perceptions of parent-child attachment, social self-efficacy, and peer relationships in middle childhood. *Infant and Child Development: An International Journal of Research and Practice, 12*(4), 351–368.

Moreno, C., Rivera, F., Ramos, P., Sánchez, I., Jiménez, A., García, I., Moreno-Maldonado, C., Paniagua, C., Villafuerte, A., Ciria, E., Abate, M., Morgan, A. (2017). *Barómetro de Opinión de la Infancia: Manual para su uso*. Madrid: UNICEF Comité Español. Retrieved from: https:// www. unicef. es/publicacion/ barometro de opinion de la infancia manual para su uso. Accessed 11 Dec 2018

Muris, P. (2001). A brief questionnaire for measuring self-efficacy in youths. *Journal of Psychopathology and behavioral Assessment*, *23(*3), 145–149.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory 3rd edn*. New York: McGraw-Hill.

Raskauskas, J., Rubiano, S., Offen, I., & Wayland, A. K. (2015). Do social self-efficacy and self-esteem moderate the relationship between peer victimization and academic performance? *Social Psychology of Education*, 18(2), 297–314.

Torsheim, T., Cavallo, F., Levin, K. A., Schnohr, C., Mazur, J., Niclasen, B., & Currie, C. (2016). Psychometric validation of the revised family affluence scale: A latent variable approach. *Child Indicators Research*, *9*(3), 771–784

Schunk, D., & Pajares, F. (2009). *Self-Efficacy theory*. In K. R. Wentzel & A. Wigfield (Eds.), Handbook of motivation at school (pp. 35–53). Routledge

Wentzel, K. R. (2014). *Prosocial behavior and peer relations in adolescence.* In L.M. Padilla-Walker & G. Carlo (Eds.), Prosocial development: A multidimensional approach (pp. 178–200). New York, NY: Oxford University Press

Zimet, G. D., Dahlem, N. W., & Farley, G. K. (1988). The multidimensional scale of perceived social support. *Journal of Personality Assessment*, *52*(1), 30–4

Zullig, K. J., Teoli, D. A., & Valois, R. F. (2011). Evaluating a brief measure of social self-efficacy among US adolescents. *Psychological reports, 109*(3), 907–920.

Validation of the Spanish version of the Goal Motives Questionnaire in athletes

Natalia Martínez-González[1], Francisco L. Atienza[2], Isabel Balaguer[1]

[1]*Faculty of Psychology, Department of Social Psychology, University of Valencia, Valencia, Spain,*

[2]*Faculty of Psychology, Department of Personality, Evaluation and Psychological Treatment, University of Valencia, Valencia, Spain*

Abstract

Having knowledge of the type of goal motives and their consequences becomes especially relevant in a sport context, where understanding all the variables involved in the process of pursuing goals could be fundamental to achieve success. In this study, a version of the Goal Motives Questionnaire was created within the framework of the Self-concordance model (SCM) to assess goal motives among Spanish athletes. Participants were 332 male and female athletes aged between 17 and 32 (M = 20.55; SD = 2.44). Confirmatory factor analyses (CFA) revealed that a model of two correlated factors provided the best fit to the data and Cronbach's alpha coefficients showed acceptable levels of internal reliability. Correlation with well- and ill- being indicators, as in previous literature, showed that autonomous goal motives significantly and positively correlated with subjective vitality and negatively with physical and emotional exhaustion, whereas controlled goal motives correlated significantly and negatively with subjective vitality and positively with physical and emotional exhaustion. The results of this study suggest that the Spanish version of this questionnaire is adequate and valid for use in a sport context.

Keywords: goal; motives; self-concordance; athlete.

Funding: This study has been partially supported by the Spanish Ministry of Education, Culture and Sport via a Grant for Training and Research awarded to Natalia Martínez-González (FPU/04671).

E-mail: natalia.martinez@uv.es

1. Introduction

According to the Self-Concordance Model (SCM) developed by Sheldon and Elliot (1999), there are different motives underlying the goals that people pursue, depending on whether these motives are more or less concordant with the person. When goals are aligned with the individual's values and interests, goal motives are called autonomous, whereas if goals are guided by internal or external pressures they are known as controlled. In sport, there are literature that has revealed that autonomous goal motives are related to goal attainment (Smith et al., 2007) and well-being indicators (Healy et al., 2014; Smith et al., 2011). Conversely, controlled goal motives are related to higher levels of ill-being (Gaudreau and Baaten, 2016) and unrelated to goal attainment (Healy et al., 2014).

Despite the fact that there is considerable previous research and this theoretical framework is well-established in sport, to date, the methodology has not been adapted to assess goal motives in Spanish athletes. Given that, in sport contexts, athletes are continually driven by goals, having an instrument adapted to Spanish athletes that provides insights into the type of goal motives is essential. Therefore, the objective of this study was to create and validate a Spanish version of the Goal Motives Questionnaire to assess goal motives in athletes.

2. Method

2.1. Participants

The study involved 332 athletes aged between 17 and 32 ($M = 20.55$; $SD = 2.44$) who competed in university teams in Valencia (Spain). Women constituted 51% of the sample. All the participants completed a set of questionnaires measuring the variables of interest.

2.2. Instruments

The Goal Motives Questionnaire was created to assess personal goal motives in the Spanish population. Based on the idiographic methodology proposed by Sheldon and Elliot (1999), items were translated into Spanish and were adapted for use with athletes, following previous procedures in the field of sport (e.g., Smith et al., 2011; Smith & Ntoumanis, 2014). Moreover, the questionnaire was adapted to cater for males and females, in order to comply with the demands of Spanish language. The result was an 8 item questionnaire that assessed the reasons why athletes pursue their most important sporting goal for autonomous (4 items; e.g., "Because of the fun and enjoyment the goal provides me") or controlled (4 items; e.g., "Because someone else wants me to") goal motives. All responses range on a 7-point Likert scale from 1 (*not at all true*) to 7 (*very true*).

The Spanish version (Castillo et al., 2017) of the Subjective Vitality Scale (SVS; Ryan and Frederick, 1997) was used to assess athletes' subjective experience of being full of energy and alive, which has traditionally been employed as an indicator of eudemonic well-being (Ryan and Deci, 2001). The instrument has 6 items (e.g., "I feel alive and vital") and responses were provided on a Likert scale from 1 (*not at all*) to 7 (*very much so*).

As an ill-being indicator, athletes' physical and emotional exhaustion was assessed through a subscale of the Athlete Burnout Questionnaire (ABQ; Raedeke & Smith, 2001), which has been adapted to the field of Spanish sport (Balaguer et al., 2011). This subscale is composed of 5 items (e.g. "I am exhausted by the mental and physical demands of sport") and responses were provided on a Likert scale from 1 (*almost never*) to 7 (*almost always*).

2.3. Data analysis

Descriptive analyses, reliability coefficients and bivariate correlations were conducted using the IBM SPSS Statistics 25 software. Confirmatory factor analyses (CFA) were carried out using Mplus (Version 7; Muthén and Muthén, 2012) to test a two-factor model, in which one factor was autonomous goal motives and the other was controlled goal motives. Several indices were used to evaluate the model fit: the raw and relative chi-square statistics, the comparative fit index (CFI), the Tucker-Lewis index (TLI) and the root mean square error of approximation (RMSEA). The model was considered to have an adequate fit if the relative chi-square was less than 3. Moreover, a CFI and TLI greater than .95 suggested an excellent fit of the data (Hu and Bentler, 1999), while values of RMSEA equal to or lower than 0.08 were optimal (Cole and Maxwell, 1985).

3. Results

3.1. Preliminary analyses

The analysis of the normality distribution (see Table 1) revealed moderate skewness and kurtosis in some items on the scale. The Kolmogorov-Smirnov test results also showed that item scores were not normally distributed. Specifically, item 8 had higher values of skewness and kurtosis, as well as high mean and lower standard deviation, indicating that most athletes pursue their goals because they want to.

Table 1. Descriptive statistics of the Spanish version of the Goal Motives Questionnaire in athletes

Item	*M*	*SD*	Ske	Kur	$Z_{K\text{-}S}$
1. Porque alguien más quiere que lo haga [Because someone else wants me to]	2.58	1.81	.76	−.70	.27**
2. Porque me sentiría avergonzado/a, culpable o ansioso/a si no lo hiciera [Because I would feel ashamed, guilty, or anxious if I didn't]	2.45	1.62	.87	−.32	.23**
3. Porque personalmente pienso que es una meta importante [Because I personally believe it's an important goal to have]	6.24	.89	−1.04	.41	.29**
4. Por la diversión y el disfrute que la meta me proporciona [Because of the fun and enjoyment the goal provides me]	6.41	.84	−1.50	1.98	.35**
5. Porque me siento presionado/a por los demás para hacerlo [Because I feel pressure from other people to do it]	1.95	1.31	1.31	.70	.31**
6. Porque me sentiría fracasado/a si no lo hiciera [Because I would feel like a failure if I quit]	2.80	1.72	.58	−.80	.19**
7. Porque me enseña autodisciplina [Because it teaches me self-discipline]	5.09	1.65	−.83	.16	.18**
8. Porque me gusta [Because I like it]	6.62	.71	−2.10	4.31	.42**

Note: The items were preceded by the stem "Persigo esta meta…" [I pursue this goal…]

***p <.01.*

3.2. Confirmatory factor analysis

The CFA results revealed that a model of two correlated factors provided the best fit to the data (χ 2 = 20.32, p < .01, χ 2/df = 2.54, RMSEA = .07, CFI = .98, TLI = .97) and Cronbach's alpha coefficients showed acceptable levels of internal reliability for autonomous and controlled goal motives (α = 70 and 67, respectively). During the analyses, two items (6 and 7) were removed, following the suggestions of both CFA fit indices and Cronbach's alpha coefficients. Specifically, the reliability analysis suggested removing item 7, which implied an increased in the Cronbach's alpha coefficient from .60 to .70 in autonomous goal motives. Moreover, the CFA suggested removing item 6 in order to improve fit indices.

As can be seen in Figure 1, the standardized factor loadings had moderate to strong values, ranging from .55 to .86.

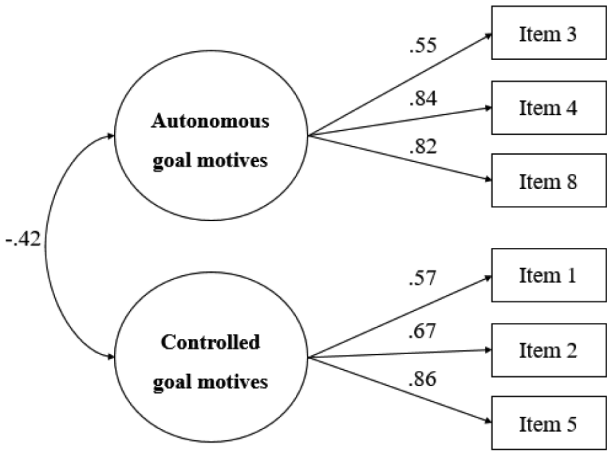


Figure 1. Two-dimensional model
Note. All parameter estimates were standardized, and all were statistically significant (p< .01)

3.3. Relationships between goal motives and well- and ill- being indicators

The correlation with other variables (see Table 2) showed that autonomous goal motives significantly and positively correlated with subjective vitality and negatively with physical and emotional exhaustion, whereas controlled goal motives correlated significantly and negatively with subjective vitality and positively with physical and emotional exhaustion.

Table 2. Descriptive statistics, Cronbach's alphas, and correlations between goal motives and well- and ill- being indicators

	1	2	3	4
Autonomous goal motives	.70			
Controlled goal motives	−.24**	.67		
Subjective vitality	.16**	−.11*	.87	
Physical and emotional exhaustion	−.26**	.25**	−.33**	.89
Mean	6.42	2.33	4.90	2.16
SD	.65	1.23	1.13	.81

Note. Internal reliability coefficients are shown on the diagonal.

**p < .05. **p <.01.*

4. Conclusions

The results of this study suggest that the Spanish version of this questionnaire is adequate and valid for use in the context of sport. Moreover, the associations between the two subscales and adaptive (i.e. subjective vitality) and maladaptive (i.e. physical and emotional exhaustion) variables, were consistent with the Self-Concordance Model's main assumptions.

On a practical level, the availability of a validated goal motive instrument for use in the field of Spanish sport will provide advances in knowledge of how the goals that athletes pursue can promote adaptive patterns when they do sport.

References

Balaguer, I., Castillo, I., Duda, J. L., Quested, E., & Morales, V. (2011). Predictores socio-contextuales y motivacionales de la intención de continuar participando: un análisis desde la SDT en danza [Social-contextual and motivational predictors of intentions to continue participation: A test of SDT in dance]. *RICYDE, 7*(25), 305–319. https://doi.org/10.5232/ricyde2011.02505

Castillo, I., Tomás, I., & Balaguer, I. (2017). The Spanish-version of the subjective vitality scale: psychometric properties and evidence of validity. *The Spanish Journal of Psychology, 20*:E26. https://doi.org/10.1017/sjp.2017.22

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research, 20*(4), 389–417. https://doi.org/10.1207/s15327906mbr2004_3

Gaudreau, P., & Braaten, A. (2016). Achievement goals and their underlying goal motivation: Does it matter why sport participants pursue their goals? *Psychologica Belgica, 56*(3), 244–268. https://doi.org/10.5334/pb.266

Healy, L. C., Ntoumanis, N., van Zanten, J. J. V., & Paine, N. (2014). Goal striving and well-being in sport: The role of contextual and personal motivation. *Journal of Sport and Exercise Psychology, 36*(5), 446–459. https://doi.org/10.1123/jsep.2013-0261

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Muthén, L. K., & Muthén, B. O. (2012). Mplus Version 7 user's Guide. Muthén and Muthén.

Raedeke, T. D., & Smith, A. L. (2001). Development and Preliminary Validation of an Athlete Burnout Measure. *Journal of Sport and Exercise Psychology, 23*(4), 281–306. https://doi.org/10.1123/jsep.23.4.281

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52*(1), 141–166. https://doi.org/10.1146/annurev.psych.52.1.141

Ryan, R. M., & Frederick, C. (1997). On energy, personality, and health: subjective vitality as a dynamic reflection of well-being. *Journal of Personality, 65*(3), 529–565. https://doi.org/10.1111/j.1467-6494.1997.tb00326.x

Sheldon, K. M., & Elliot, A. J. (1999). Goal striving, need satisfaction, and longitudinal well-being: the self-concordance model. *Journal of Personality and Social Psychology, 76*(3), 482–497. https://doi.org/10.1037/0022-3514.76.3.482

Smith, A. L., & Ntoumanis, N. (2014). An examination of goal motives and athletes' self-regulatory responses to unattainable goals. *International Journal of Sport Psychology, 45*(6), 538–558. https://doi.org/10.7352/IJSP2014.45.538

Smith, A. L., Ntoumanis, N., & Duda, J. L. (2007). Goal striving, goal attainment, and well-being: Adapting and testing the self-concordance model in sport. *Journal of Sport and Exercise Psychology, 29*(6), 763–782. https://doi.org/10.1123/jsep.29.6.763

Smith, A. L., Ntoumanis, N., Duda, J. L., & Vansteenkiste, M. (2011). Goal striving, coping, and well-being: a prospective investigation of the selfconcordance model in sport. *Journal of Sport and Exercise Psychology, 33*(1), 124–145. https://doi.org/10.1123/jsep.33.1.124

App for Android and iOS for hypothesis testing: the relationships between two variables

Gaspar Berbel[1], Emili Álvarez[2]

[1]*Department of Economics, University of Girona (EUMediterrani), Spain,*

[2] *Computer consulting LEULIT, s.l., Spain*

Abstract

ESTATest is an application developed in Flutter for Android and iOS. It shows which statistical test to perform for each combination of two variables (categorical of two categories or more, and numerical or metric). ESTATest is an intuitive, easy-to-use mobile application, designed for any student, teacher, researcher, or data analyst who needs to relate variables (perform hypothesis testing, bivariate relationship tests). The app not only proposes the test to be performed, but it also allows the user to watch a short video (1-2 minutes) explaining how the statistical procedure is performed and how it is interpreted, with the IBM-SPSS statistical package.

Keywords: Research design applications, Inferential statistics, Bivariate relationships.

E-mail: bel@aptabel.com

1. Introduction

The creation of ESTATest is motivated by the need to understand what is behind the systematics and the logic of parametric and non-parametric hypothesis tests, as they are currently applied.

Before making this app, we did some research to see if there were any similar apps. To our surprise, there was nothing similar on the market, which encouraged us even more to design it.

It is part of a learning system based on the PLE (Personal Learning Environment), gathered in the statistics manual "Paola aprende estadística. Desde un PLE" (Berbel, 2020).

Its purpose is to enable students or researchers—in a maximum of 4 clicks—to reach the proper test of a relationship between two variables, showing how it is done and how the test is interpreted and finished, in APA style, with IBM-SPSS.

2. Method

The design of the app began with a flow diagram in which the statistical procedures that resulted from the possible combinations of variables—categories of 2 or more categories and metrics—were collected. These procedures were parametric and non-parametric (when the metric variable does not follow normal law). See Table 1.

From the flow diagram we went to an application developed in Flutter, an SDK (Software Development Kit) developed by Google to create mobile applications for both Android and iOS (Apple).

2.1. Flutter. Google's UI toolkit

Flutter is Google's UI toolkit for building beautiful, natively compiled applications for mobile, web, desktop, and embedded devices from a single codebase.

The software was initially developed for internal use in the company. The huge potential of the tool led Google to launch it as an open source project. It is currently one of the fastest growing mobile application development projects.

2.2. App navigation

First, users choose the language (Spanish, Catalan, French or English). Then, they select the combination of variables and the application conditions are determined. Finally, the app shows them the sequence of each procedure in different menus. It also displays a video tutorial with voice-off that explains and shows how each statistical procedure is performed, interpreted and finished following APA standards.

3. Results

Firstly, the user is asked to indicate which variables that they want to relate, and their type (categorical or metrical). Once the combination has been input -see Table 1-, the app –see figure 1- guides them through the application conditions (effective, normal law). When the filters are applied, the app indicates which test the user should apply for hypotheses or procedures to obtain a possible association between these variables. The possible tests that the app will suggest are: Chi square, t-test independent measures, t-test related measures, anova, correlation or parametrical tests. If the metric variable does not follow the normal law – non parametrical tests – the application will suggest: Mann-Whitney U test, Wilcoxon test, Kruskal-Wallis test, or the Spearman's Rho correlation.

Table 1. Combination of variables.

COMBINATIONS		2 k	More of 2 k	C
Categorical variable (K)	**2 k**	Chi Square		t-test (MI) NOPAR: Mann-Whitney U test
	More of 2 k			ANOVA (MI) NOPAR: Kruskal-Wallis test
Metric variable (C)				1) Comparison: t-test (MR) NOPAR: Wilcoxon test 2) Association: r correlation NOPAR: Spearman's Rho

Notes: k=categories, C=metric or numerical, NOPAR=Non-parametric, MR=Paired samples, MI=Independent samples

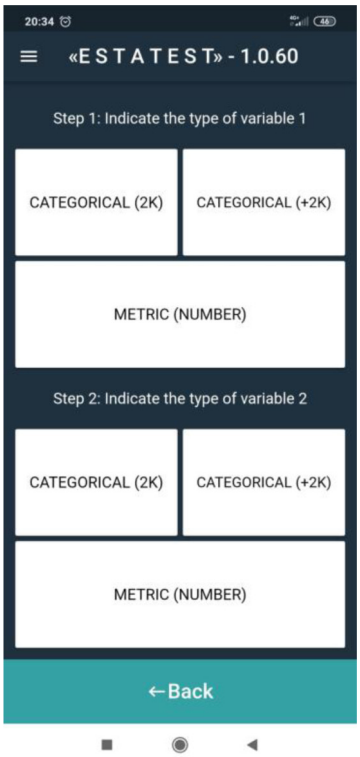


Figure 1. Main Estatest app menu

4. Conclusions

The ESTATEST app enables the user to ascertain which possible analyses are suitable for testing for hypotheses about the relationships between two variables in a simple and intuitive way. It is an excellent didactic tool for any student or researcher who needs to relate variables.

References

Berbel, G. (2020). *Paola aprende estadística. Dentro de un entorno personal de aprendizaje*. Barcelona: Aptabel grup.

Flutter (2021, July 25). Flutter documentation. California, EE.UU. https://flutter.dev/docs

IBM Corp. (2013, 2015, 2016). *IBM SPSS Statistics para Windows*, versión 22.0. Armonk, Nueva York: IBM Corp.

Multilevel Single-Trial Analysis of Event-Related Potentials: A Case Study

Juan C. Oliver-Rodríguez[1]

[1]*Department of Psychology, Universitat Jaume I, Spain*

Abstract

Repeated measures ANOVA and MANOVA are frequently used procedures in the analysis of Event-related Brain Potentials (ERPs). They are typically applied to averaged repeated stimulus trial responses to increase the reliability of the electroencephalogram signal. However, trial averaging can lead to information loss concerning the covariances of random individual differences in treatment or time effects. Previous more general studies, as well as a simulation based on a single electrode have shown that lack of an adequate specification of the covariance matrix can lead to inference errors, such as increases in Type I error rates. It would be expected that such biases would also occur in situations with multiple electrode records. The objective of the present study is to conduct preliminary comparative analyses of a case facial perception experiment with responses from three electrodes and gender-role measures as covariates. The multilevel analysis on single trials was more sensitive to detect a curvature apparent in the empirical face effects on mean Late Positive Component amplitudes, as well as differential gender role response profiles across electrodes. Only the ANOVA on averaged-trial responses detected an interaction between face slope effects and femininity. A new simulation study may be instrumental to disentangle potential inferential biases and effect sizes and to study the more general properties of the two procedures in the multiple electrode situation.

Keywords: Event-related potentials; multilevel models; late positive component

1. Introduction

Event-related Brain Potentials (ERPs) are oscillations in brain electroencephalogram (EEG) measures that occur as a result of a stimulus presented to the subject, such as a picture or a sound. Since raw EEG responses have low reliabilities, multiple repeated trials under each experimental condition are presented to each participant. A common approach is to average them subsequently to reduce random noise and increase the detectability of the signal of interest. Results are then quantified, typically in a spatial and temporal region that reflects the psychological process under study, such as affective perception, memory or language (Luck, 2014). A typical ERP statistical analysis consists of applying General Linear Model (GLM) procedures to the resulting data set with the purpose of testing *population-average* effects. As a consequence, individual difference parameters remain hidden in the error term. This produces information loss on subject-specific treatment effects which could be of substantive or clinical interest. GLM Anova or Manova methods also produce information loss in a second way, since records containing missing observations are automatically deleted prior to the analysis (Kristjansson, Kircher & Webb, 2007).

Alternatively, analyses with mixed multilevel models are performed on single unaveraged trial responses as nested within participants. Observations obtained from the same person are represented as clusters. In doing so, the multilevel model allows for a richer, more flexible specification of the covariance matrix, makes individual difference parameters explicit, and allows for testing *subject-specific* effects. They use maximum likelihood estimation procedures with desirable properties like consistency and minimum variance that additionally make more efficient use of observations with missing data in comparison to the Anova approach (Hox, Moerbeek & van de Schout, 2018; Kristjansson, Kircher & Webb, 2007).

In terms of analytical considerations (Goldstein, 2011), ignoring dependencies between observations in the Anova /Manova analyses of averaged responses could lead to inference errors. This is what was found in a simulation study using parametric values from an ERP case study on facial perception (Oliver-Rodríguez & Moerbeek, 2018, 2019) (Fig. 1). In the simulation conditions where random individual differences in slope effects were generated, Anova analyses on trial-averaged responses produced a negative standard error bias. As a consequence, a statistically detected increase in Type I Error rate was observed, which could lead to a lack of replicability of the experimental effects. This inference bias was not observed in the multilevel analysis of single-trial responses.

The above simulation was performed using a simplified situation from results recorded on a single parietal electrode. In practice, however, analyses typically include multiple electrodes, and the data have a more complex temporal and spatial covariance structure. Since Anova analysis based on trial-averaged responses is still a common procedure in psychophysiological literature (Brace & Sussman, 2021; Coch & Mahoney, 2021; Keil et al., 2014; Hülsemann & Rasch, 2021; Mitra & Koch, 2018; Wu et al., 2018), a new simulation study using multiple electrodes would be desirable. The purpose of the present communication is to present a preliminary comparative case-study analysis of the inferential results obtained by the two procedures.

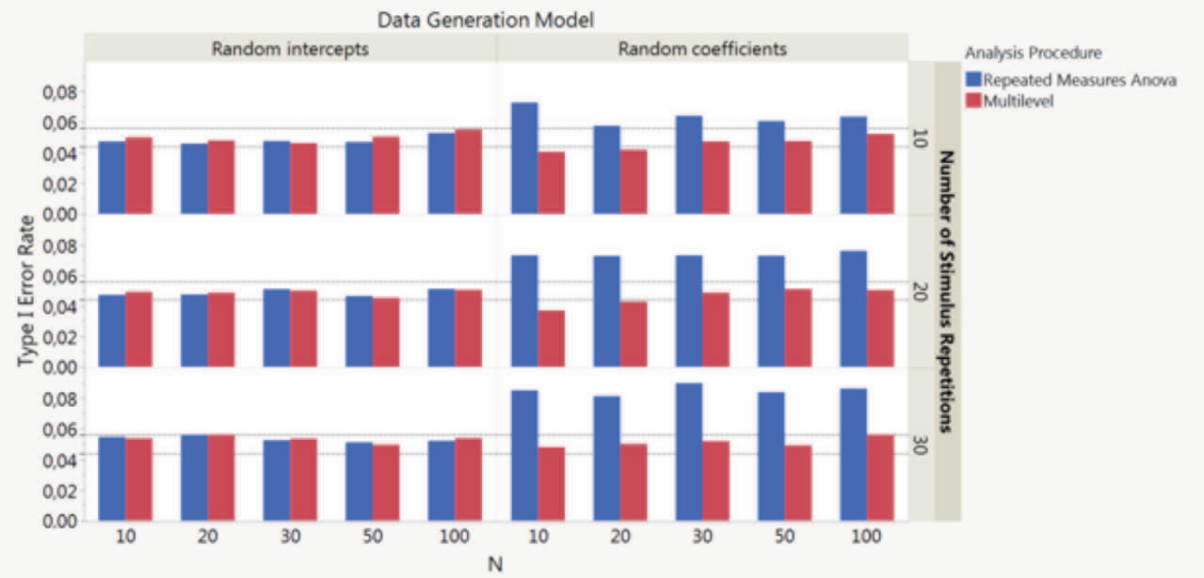


Figure 1. Mean Type I Error rates as a function of sample size, number of stimulus repetitions, data generation, and data analysis model. Dashed lines indicate statistical significance bounds for the binomial test. Results are based on 5000 simulation runs.

2. Method

2.1. Case Experimental Design

The purpose of the ERP study was to study affective and attractiveness perception to female facial characteristics by means of the Late Positive Component (LPC) in the ERP (Oliver-Rodríguez, Scarbrough & Johnston, 2016). The stimuli consisted of five faces ranging on a continuum from less feminine to more feminine physical characteristics. They were extracted as equally spaced images from a movie constructed on the basis of standard facial features and attractiveness ratings. The association of LPC component amplitude with affective and attractiveness perception processes has been replicated in a number of studies (Oliver-Rodríguez, Guan & Johnston, 1999; Marzi & Viggiano, 2010; Werheid, Schacht & Sommer, 2007; Zhang & Deng, 2012). LPC amplitude was measured on twenty-five electrode sites, but only results from three sites (Fz, Cz and Pz) were used in the present comparison. Fifty-seven male students from the Universitat Jaume I participated in the experiment, during which they were exposed to six blocks of trials. Each block contained one presentation of each stimulus face in a random order. LPC amplitude was hypothesized to increase monotonically with more feminine facial characteristics, as well as to be modulated by participants' gender-roles as measured by the Bem Sex Role Inventory (Bem, 1981).

2.2. Case Analysis Comparisons

The multilevel analysis model in single trials included all sources of variation from the original design with Electrode Site crossed with Trials, which were nested within the Face by Participant combinations. The Anova model included Electrode Site, Face and Participant as completely crossed factors, since single trials are typically averaged out for each stimulus and participant combination. Masculinity and Femininity Bem scores were used as covariates in both analyses. An $\alpha = 0.05$ criterion was used for the detectability of statistical effects.

3. Results

The multilevel model selection steps for the variance components are described in Table 1. Each successive step was statistically significant according to the likelihood ratio test. The final model included a total of ten covariance parameters with random individual intercepts and slopes, heterogeneous electrode site variances and a residual first order autocorrelation. Plots of random individual differences in face intercepts and slope estimates and individual response profiles are shown in Figs. 2 and 3.

Table 1. Cumulative Multilevel Model Selection Steps

Model	Covariance Parameters	Model Fit -2RLL[a]	AIC
Random Intercepts	**2**	**132514.5**	**132518.5**
Random Coefficients	**3**	**132440.2**	**132448.2**
Heterogenous Electrode Site Covariances	**9**	**123079.6**	**123097.6**
First Order Autocorrelation	**10**	**122957.4**	**122977.4**

a. -2 Residual Log Likelihood

Comparative results between the multilevel model in single trials and the repeated measure Anova on trial averages using the Geisser-Greenhouse correction are shown in Table 2. Both the multilevel and the Anova model detected a linear increase in LPC amplitude as a function of face, but the quadratic linear trend apparent in the empirical means (Fig. 4) was only obtained with the multilevel model. The main Bem masculinity effects were detected by both procedures, but only the multilevel analysis detected two-way and three-way interactions of Bem masculinity and femininity with an electrode site, the latter being represented in Fig. 5. On the other hand, only the repeated measure Anova on average responses detected an interaction between Bem femininity and the quadratic face effects, due to decreasing curvature with increasing femininity levels.

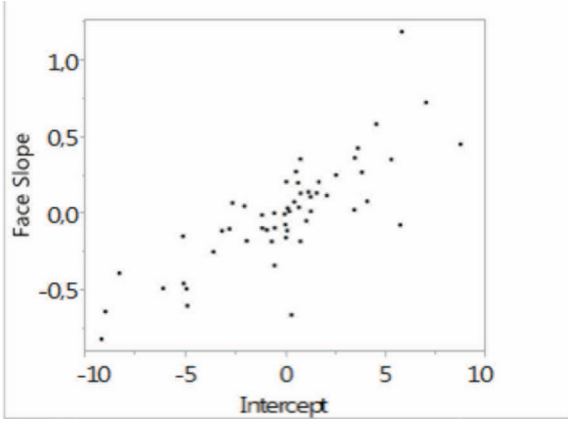


Fig. 2. Random differences in individual intercepts and slopes for face effects.

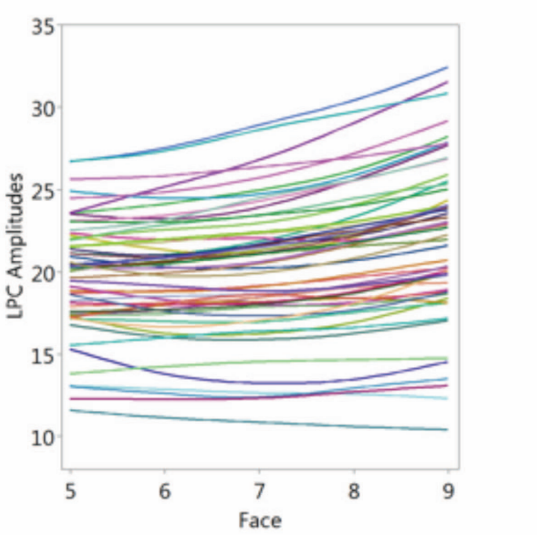


Fig. 3. Random differences in predicted individual response profiles

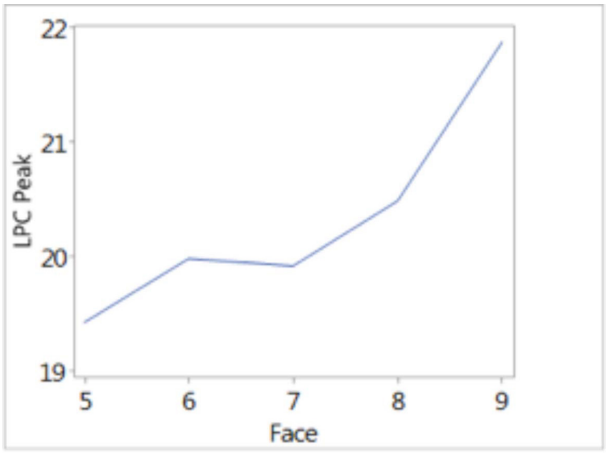


Fig. 4. Empirical LPC Means as a function of face.

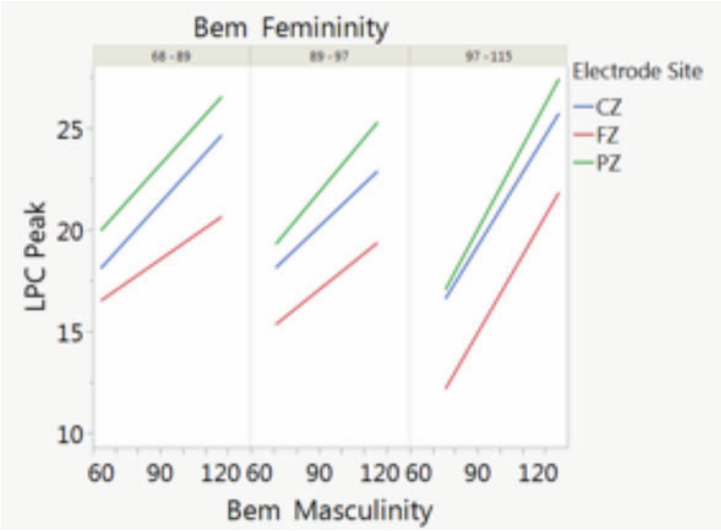


Fig. 5. Representation of the three way interaction between Bem gender role measures and Electrode Site on mean Late Positive Component Amplitude.

Table 2. Model Effect Comparisons

Effects	Multilevel p value	Repeated Measures Anova p value	Anova $\hat{\eta}^2_{partial}$
Face (Linear)	< 0.0001***	< 0.0001***	0.31
Face (Quadratic)	0.0381*	0.143	0.04
Bem Masculinity	0.0151*	0.011*	0.12
Bem Masc. × Esite	0.0085**	>.10	0.02
Bem Fem. × Esite	0.049*	>.10	< 0.01
Bem Masc × Bem Fem × Esite	0.049*	>.10	<0.01
Face (Linear) × Bem Masc.	0.0682	0.0812	0.06
Face (Quadratic) × Bem Fem.	0.0632	0.046*	0.076

4. Conclusions

The purpose of the study was to compare performances of multilevel single trials vs repeated measure Anova on averaged-trial responses of an ERP case study in facial perception. The multilevel model was more sensitive in detecting a quadratic face effect, and a changing modulation of gender-role measures on LPC means (or centered intercepts) across electrode sites. This could be due to increased precision and efficiency occurring as the mixed model incorporated a fuller description of the covariance components (Rao, 1973).

Modulation of the face slope coefficient as a function of gender role measures only appeared with the Anova on averaged-trial responses. Although the previous simulation study with one electrode brings some inferential caution for the interpretation of the Anova results, a case comparative analysis cannot be used to determine the degree to which the latter effect may be influenced by an underestimate of the standard error bias or an increase in Type I Error rate. A new simulation study may be instrumental in disentangling potential standard error bias and effect sizes and in studying more general properties of the two procedures in the multiple electrode situation.

Simulation methods are also increasingly being used in the sample size planning of multilevel data (Moerbeek & Teerenstra, 2016), and may be instrumental in ensuring good use of resources in ERP research. Stricter levels of evidence may be required in psychophysiological studies that take into account a fuller and more realistic description of their variance components. These considerations for sample size planning could however be an asset for improving replicability.

References

Brace, K. M. & Sussman, E. S. (2021). The role of attention and explicit knowledge in perceiving bistable auditory input. *Psychophysiology;58*:e13875. https://doi.org/10.1111/psyp.13875

Bem, S. (1981). Bem Sex-Role Inventory professional manual. Palo Alto CA: Consulting Psychology Press.

Coch D, Mahoney M. R. (2021). When two vowels go walking: An ERP study of the vowel team rule. Psychophysiology;58:e13870. https://doi.org/10.1111/psyp.13870

Goldstein, H. (2011). Multilevel statistical models. London: Wiley.

Hox, J. J., Moerbeek, M., van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). London: Routledge.

Hülsemann MJ, Rasch B. (2021) Embodiment of sleep- related words: Evidence from event-related potentials. *Psychophysiology*;58:e13824. https://doi.org/10.1111/psyp.13824

Kristjansson, Kircher & Webb (2007). Multilevel models for repeated measures research designs in psychophysiology: An introduction to growth curve modeling. *Psychophysiology, 44*, 728–736.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press.

Marzi, T. & Viggiano, M. P. (2010). When memory meets beauty: Insights from event-related potentials. *Biological Psychology, 84*, 192–205.

Mitra P, Coch D. (2019) An ERP study of cross-modal rhyming: Influences of phonology and orthography. *Psychophysiology*.;56:e13311. https://doi.org/10.1111/psyp.13311

Moerbeek, M. & Teerenstra, S. (2016). *Power analysis of trials with multilevel data*. London: CRC Press.

Oliver-Rodríguez, J. C., Guan, Z. & Johnston, V. (1999). Gender differences in late positive components evoked by human faces. *Psychophysiology*, *36*, 176–185.

Oliver-Rodriguez, J. C. y Moerbeek, M. (2018). Comparative performance of single-trial multilevel analysis of event-related brain potentials. Communication presented in the 8th European Congress of Methodology. Jena Universität (Germany), July 25–27.

Oliver-Rodríguez, J. C. & Moerbeek, M. (2019). A critical perspective on trial-averaged analyses of Event-Related Potentials. Communication presented at the 16th Congress of Methodology for Health and Social Sciences. Universidad Autónoma de Madrid, July 8–10.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.

Werheid, K., Schacht, A. & Sommer, W. (2007). Facial attractiveness modulates early and late event-related brain potentials. *Biological Psychology, 76*, 100–108.

Wu L, Müller HJ, Zhou X, Wei P.(2019) Differential modulations of reward expectation on implicit facial emotion processing: ERP evidence. *Psychophysiology*;56:e13304. https://doi.org/10.1111/psyp.13304

Zhang, Z. & Deng, Z. (2012). Gender, facial attractiveness, and early and late event-related potential components. *Journal of Integrative Neuroscience, 11* (4), 477–487. doi: 10.1142/S0219635212500306.

SYMPOSIA

Gender issues in methodology: A discussion on its effects on research and teaching

Amparo Oliver[1], Inés Tomás[1,2]

[1]*Department of Methodology for Behavioral Sciences, University of Valencia, Spain,*

[2]*IDOCAL, University of Valencia, Spain*

1. State of the art

In recent decades, there has been a paradigm shift in science regarding how gender is approached. First, the need to overcome the binary perspective of gender identity and to consider gender diversity has been reivindicated. In many ways, this also affects our research and teaching practices as methodologists. Second, one important question has been raised, which is whether the measurement tools that are assumed to perform equally for men and women could be biased by gender. Third, in some research fields, it could be relevant to adopt a gender-sensitive approach and analyze whether there are different design modality preferences based on gender. Finally, we have to face the fact that traditionally more men than women study careers tagged as STEM disciplines (science, technology, engineering, and mathematics). This reveals a preference that we must question as methodology professionals, and should lead us to make some self-criticism... Do we have something to do with this?

2. New perspectives and contributions

Five contributions were presented in this symposium, even though only two of them appear as extended summaries in this Proceedings Book. The first contribution ("The inclusion of gender diversity in psychological research" presented by Alicia Tamarit), discussed how the integration of knowledge on gender and sexual diversity facilitates the adequate representation of all realities, and thus guarantees scientific rigor in the use of psychometric measures in research. The second ("Measurement invariance by sex in sensitive constructs: first evidence for Ambivalent Sexism Inventory" presented by Irene Fernández) and third ("Do gender role stereotypes still prevail? Measurement invariance of work centrality and job meaningfulness over gender" presented by Marija Davcheva) contributions, provided new evidence on measurement invariance in validated scales approaching "sensitive" topics, like sexism in youth or the role of women at work. The fourth contribution ("A pilot study on paper-based visual analogue scale items, focus on gender" presented by Klemens Weigl), represented an example of a study on paper-based visual analogue scale items from a gender-sensitive approach. The last contribution ("Statistics anxiety and gender: A systematic literature review and research agenda" presented by Sara Martínez-Gregorio), addressed the topic of statistics anxiety by gender and how it connects with many domains affecting our research and teaching sphere.

3. Research and practical implications

The contributions to the symposium tried to provide insightful and updated information on the state of the art about "gender issues in methodology" to facilitate discussion among attendees. The main implications of the contributions shared in this symposium were the following: to become aware of gender issues occurring in our daily practice as methodologists, to share the tools we already have to answer this challenge and to exchange ideas on further initiatives.

Keywords: Gender issues in methodology; design and gender issues; statistical analysis bias; invariance.

E-mails: Amparo.Oliver@uv.es; Ines.Tomas@uv.es

The inclusion of gender diversity in psychological research

Alicia Tamarit[1], Estefanía Mónaco[1], Marta Cañero[1], Inmaculada Montoya Castilla[1]

[1]*Department of Personality, Assessment and Psychological Treatment, Universitat de València, Spain*

Abstract

Purpose: In recent times, there has been a paradigm shift in psychology regarding how gender is studied. Traditionally, assessment instruments included the variable of sex, referring to the distinction according to sex chromosomes (XX or XY), or the variable of gender, meaning whether the person identified as a man or a woman. Today, we know this binary perspective of gender identity is a misrepresentation of the reality of the population, and it became imperative to include gender diversity in research design and methodology. The aim of this study is to identify the variables related to gender and sexuality that could be included when designing research and developing new measuring instruments in psychology, or adapting those that already exist. *Results*: The essential aspects to be considered in the development of psychometric measures are 1) gender, understood as a non-binary spectrum that includes cisgender and transgender people, which covers, among others, transgender, genderfluid and non-binary people, 2) sexual orientation, not only including traditional labels such as heterosexual, homosexual or bisexual, but also the gradient that exists between alosexuality and asexuality, 3) non-sexist language, including the gender perspective in scientific texts, and 4) the rupture of classic schemas related to affective bonds and sexuality, such as the assumption of monogamy or traditional gender roles. *Conclusions*: The integration of this knowledge on gender and sexuality facilitates the adequate representation of all realities, and thus guarantees scientific rigor in the use of psychometric measures in psychological research.

Keywords: gender issues in methodology, sexuality, gender and sexual diversity.

E-mail: alicia.tamarit@uv.es

1. Introduction

Psychological research currently includes gender assessment in most socio-demographic evaluation instruments, serving as a method for participant categorization into distinct gender categories (Luyt, 2013). Despite its ubiquity in most studies in psychology, there has been little evolution in this primary approach to gender, which is incongruous with the social and cultural progress in understanding this construct (Holdcroft, 2007). Thus, contemporary methods of evaluation tend to ignore diverse realities in their target populations and disregard the current policies that urge researchers to assess gender as a non-binary construct (American Psychological Association, 2015).

1.1. Biological sex and gender identity

Scientific understanding of gender stems from the binary categorization of biological sex (Hyde et al., 2019). Up until this day, newborns are assigned to one of the two main categories of the gender binary (namely, a girl or a boy) according to their established biological sex (i.e., female or male), and this becomes the basis for the political, social and cultural position they later occupy in society (Darwin, 2020). Interestingly, biological sex is a complex phenomenon that might significantly differ from the assigned category at birth (Morrison et al., 2021).

Biological sex is observed and measured according to two different manifestations: genotypic sex and phenotypic sex (Karkazis, 2019). Genotypic sex is determined by the sex chromosomes, which involves a wide array of combinations of X and Y (beyond the XX and XY dichotomy), further translating into varied gene expressions (Carpenter, 2018). This phenomenon in itself challenges the sex binary, which could be better understood as a spectrum: these combinations lead to a myriad of different hormonal patters that change within individuals and across sex labels, resulting in diverse sexual organs (Karkazis, 2019). Phenotypical sex is observed by the formation of these sexual characteristics, which seldom adjust to two strict categories, rather than a complex spectrum that goes beyond the sex binary as has been traditionally studied (Štrkalj & Pather, 2021). Intersex individuals, who make up around 2% of the population (rough estimate, based on a mostly white, Western population), have contributed to this shift in the paradigm of sexual categorization (Hyde et al., 2019).

Sex assigned at birth is, therefore, a diagnostic label based on the genitals of the newborn, which do not necessarily match the multiple biological characteristics that are assumed of the baby's assigned sex category (i.e. female or male) (Karkazis, 2019). Furthermore, based on this binary label, the gender identity of the baby is inferred, thus starting the socialization process that will ensure they adapt to the gendered roles that are expected for them to perform (Morrison et al., 2021).

However, on many occasions, the gender identity of the individual does not match the one that was assigned at birth, since it was assigned according to one phenotypical characteristic (external genitalia) that holds no correlation with their personality, identity or behavior. Therefore, research in psychology is focused on assessing *gender* rather than sex (Nguyen et al., 2018; Olson et al., 2015). *Cisgender* people are those whose gender identity matches their gender assigned at birth, and *transgender*, or *trans* people are those whose gender identity is different from their gender assigned at birth (Darwin, 2020).

As with biological sex, research and political activism has challenged the gender binary and sees gender as a spectrum rather than two distinct labels (Hyde et al., 2019). This spectrum is considered bimodal: there are two poles (i.e. women and men) and a wide umbrella of non-binary identities in between: non-binary, genderqueer, genderfluid, gender non-conforming,

pangender, androgyne, and more: agender, genderless or gender neutral people who do not identify with any gender, or identify with having no gender (Bertrand, 2020).

1.2. Measuring gender

Despite the overwhelming literature on this multiplicity of gender identities, researchers often fail to properly assess this construct and abide by the American Psychological Association guidelines (2015) for gender measurement in psychology. According to Cameron and Stinson (2019), there are serious implications for gender mismeasurement: (1) it is inherently transphobic, so it violates the ethical principles of harm avoidance in health research, (2) misclassifications due to the impossibility of non-binary and trans participants to accurately answer these questions and (3) the threat to the validity of the research that this mismeasurement entails. Thus, updating the traditional, inaccurate gender measurement approaches is fundamental for preserving the studies' psychometric properties, and more importantly, for respecting and protecting the participants' human rights (Bauer et al., 2017).

1.3. The current study

In the light of this research, it would be of interest to review the main approaches to gender measurement in research and to ascertain the structure of assessment tools that account for gender diversity. The aim of this study was to conduct a bibliography review of the general tendencies in the assessment of gender, including the literature recommendations for the inclusion of gender diversity in research, and conducting a comprehensive proposal for inclusive gender assessment in psychology and health sciences.

2. Method

The methodology used in this study consisted of an analysis of the existing scientific literature on this subject. A bibliographic search was conducted to identify and select the articles, books and scientific reports that were reviewed. Preliminary searches included articles published in the last 10 years, later adding older, fundamental references to account for the origins of the topic of gender assessment. Keywords for the bibliographic search were "gender spectrum", "gender assessment", "gender measurement", "gender binary" and "non-binary gender assessment". The scientific databases used were Proquest, Psycinfo, Pubmed, Dialnet, JCR and Google Scholar.

3. Results

In the beginning of psychology studies, most research on human behavior did not take gender into account: following the biomedical approach, the first experimental studies relied on using male, cisgender, white, young participants as the "norm" or "default", rendering gender assessment redundant or unnecessary (Miller & Hay, 2004; Richardson et al., 2015)the US National Institutes of Health (NIH. As a consequence of social and political activism, and scientific difficulties derived from using only male participants, female individuals with similar socio-demographic characteristics were included, therefore stating the need for universal gender assessment (Criado Perez, 2019; Fine, 2010).

The traditional structure of gender assessment tools has been based on close-ended, ad hoc questionnaires that conceived gender as a binary concept with two distinct categories: female and male (Fine, 2010). Moreover, this model uses the concept "sex" and "gender" indistinctively, neglecting the fundamental differences between the medical category of sex assigned at

birth and the nuanced, psychological and social foundations of gender identity (Gannon et al., 1992; Hyde et al., 2019).

From the 1990s until today, social awareness of non-binary identities has been reflected in scientific studies on human psychology (Aparicio-García et al., 2018). The default ad-hoc questionnaires assessing gender started changing the two-category model with the categories (a) female or woman and (b) male or man, to a three-alternative question that included the option (c) unspecified/other (Ho & Mussap, 2019). This new approach allows people who do not conform to the gender binary to disclose their identity as not belonging to this traditional model—however, this *othering* of individuals may result as invalidating of their realities, as it implicitly suggests that only cisgender, binary individuals are allowed to express their actual identity and all non-binary genders are discarded to this separate category (Morrison et al., 2021). This is solved by substituting the "other" alternative with an open-ended question (Paveltchuk et al., 2019); this option helps collecting the details of the identities that fall under this category, however it might hinder the automatization of data processing and analysis.

An inclusive approach is proposed by Cameron and Stinson (2019), which includes the option for transgender and non-binary individuals, besides the three closed and open-ended alternatives mentioned above. However, this poses a different difficulty for data interpreting, since it fails to account for gender differences *within* transgender identities. Bauer et al. (2017) suggest a succinct but thorough ad-hoc item that includes the assessment of both cisgender and transgender women and men, including genderqueer or non-binary identities.

According to the American Psychological Association guidelines (2015), research in psychology is required to assess gender as a non-binary construct, therefore accounting for the diversity in gender identities in the default socio-demographic evaluation. Based on these guidelines and the current evidence on gender identities and how to accurately evaluate them (Bauer et al., 2017)government and research organizations are increasingly expanding measures of sex/gender to be trans inclusive. Options suggested for trans community surveys, such as expansive check-all-that-apply gender identity lists and write-in options that offer maximum flexibility, are generally not appropriate for broad population surveys. These require limited questions and a small number of categories for analysis. Limited evaluation has been undertaken of trans-inclusive population survey measures for sex/gender, including those currently in use. Using an internet survey and follow-up of 311 participants, and cognitive interviews from a maximum-diversity sub-sample (n = 79, a basic structure of an ad-hoc gender assessment item is proposed in Table 1.

Table 1. Examples of gender measurement in increasing the order of diversity inclusion.

Gannon et al. (1992)	Ho & Mussap (2019)	Paveltchuk et al. (2019)	Cameron & Stinson (2019)	Proposal based on APA (2015) and Bauer et al. (2017)
□ Woman □ Man	□ Woman □ Man □ Other	□ Woman □ Man □ I identify my gender as _____	□ Woman □ Man □ Transgender □ Nonbinary □ I identify my gender as _____	□ Woman □ Man □ Transgender woman □ Transgender man □ Nonbinary □ Nonconforming □ Genderqueer □ I identify my gender as _____ □ I prefer not to say

4. Conclusions

The issue of gender assessment that accounts for diversity is still unresolved today despite the ample research conducted on this matter, which not only provides comprehensive, inclusive assessment tools but also highlights the social and cultural importance of evaluating identities outside the normative cisgender binary.

Literature on psychology and political activism have raised awareness of this wide diversity of realities in the gender spectrum, which have been neglected in and removed from scientific research. Researchers today hold a responsibility to acknowledge these identities and finally include them in scientific literature, towards an inclusive, ethical, and conscious approach to gender diversity.

References

American Psychological Association. (2015). *Guidelines for Psychological Practice With Transgender and Gender Nonconforming People*. https://doi.org/10.1037/a0039906

Aparicio-García, M. E., Díaz-Ramiro, E. M., Rubio-Valdehita, S., López-Núñez, M. I., & García-Nieto, I. (2018). Health and well-being of cisgender, transgender and non-binary young people. *International Journal of Environmental Research and Public Health*, *15*(10), 2133. https://doi.org/10.3390/ijerph15102133

Bauer, G. R., Braimoh, J., Scheim, A. I., & Dharma, C. (2017). Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PLOS ONE*, *12*(5), e0178043. https://doi.org/10.1371/JOURNAL.PONE.0178043

Bertrand, M. (2020). Gender in the Twenty-First Century. *AEA Papers and Proceedings*, *110*, 1–24.

Carpenter, M. (2018). The "Normalization" of Intersex Bodies and "Othering" of Intersex Identities in Australia. *Journal of Bioethical Inquiry 2018 15:4*, *15*(4), 487–495. https://doi.org/10.1007/S11673-018-9855-8

Criado Perez, C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. Abrams Press. https://books.google.es/books?hl=es&lr=&id=MKZYDwAAQBAJ&oi=fnd&pg=PT8&dq=invisible+women+caroline+criado&ots=PqTLVWX0Lm&sig=k4CWjGpl6wDTvxc7weRh92GfEFo#v=onepage&q=invisible women caroline criado&f=false

Darwin, H. (2020). Challenging the Cisgender/ transgender Binary nonbinary People and the transgender label. *GENDER & SOCIETY*, *34*(3), 357–380. https://doi.org/10.1177/0891243220912256

Fine, C. (2010). *Delusions of Gender: How Our Minds, Society, and Neurosexism Create Difference*. W. W. Norton & Company. https://books.google.es/books?hl=es&lr=&id=s2ZtdAx83yMC&oi=fnd&pg=PR15&dq=delusions+of+gender&ots=GXcDLICJUW&sig=rt5s2CkXDuRz6Kf1Fj934FhhI7s#v=onepage&q=delusions of gender&f=false

Gannon, L., Luchetta, T., Rhodes, K., Pardie, L., & Segrist, D. (1992). Sex bias in psychological research: Progress or complacency? *American Psychologist*, *47*(3), 389–396. https://doi.org/10.1037/0003-066X.47.3.389

Ho, F., & Mussap, A. J. (2019). *The Gender Identity Scale: Adapting the Gender Unicorn to Measure Gender Identity*. https://doi.org/10.1037/sgd0000322

Holdcroft, A. (2007). Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, *100*(1), 2–3. https://doi.org/10.1177/014107680710000102

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

Karkazis, K. (2019). The misuses of "biological sex." *The Lancet*, *394*(10212), 1898–1899. https://doi.org/10.1016/S0140-6736(19)32764-3

Luyt, R. (2013). Beyond traditional understanding of gender measurement: the gender (re)presentation approach. *Http://Dx.Doi.Org/10.1080/09589236.2013.824378*, *24*(2), 207–226. https://doi.org/10.1080/09589236.2013.824378

Miller, V., & Hay, M. (2004). *Principles of Sex-based Differences in Physiology*. Gulf Professional Publishing. https://books.google.es/books?hl=es&lr=&id=wH1Pt5hbqAQC&oi=fnd&pg=PR5&dq=principles+of+sex-based+differences+in+physiology&ots=k9iNRjn_jx&sig=_lmkEY9bmFxjr-FAOYwbnm-hUho#v=onepage&q=principles of sex-based differences in physiology&f=false

Morrison, T., Dinno, A., & Salmon, T. (2021). The Erasure of Intersex, Transgender, Nonbinary, and Agender Experiences by Misusing Sex and Gender in Health Research. *American Journal of Epidemiology*. https://doi.org/10.1093/AJE/KWAB221

Nguyen, H. B., Loughead, J., Lipner, E., Hantsoo, L., Kornfield, S. L., & Epperson, C. N. (2018). What has sex got to do with it? The role of hormones in the transgender brain. *Neuropsychopharmacology 2018 44:1*, *44*(1), 22–37. https://doi.org/10.1038/s41386-018-0140-7

Olson, K. R., Key, A. C., & Eaton, N. R. (2015). Gender Cognition in Transgender Children. *Psychological Science*, *26*(4), 467–474. https://doi.org/10.1177/0956797614568156

Paveltchuk, F. de O., Damásio, B. F., & Borsa, J. C. (2019). Impact of sexual orientation, social support and family support on minority stress in lgb people. *Trends in Psychology*, *27*(3), 735–748. https://doi.org/10.9788/TP2019.3-10

Richardson, S. S., Reiches, M., Shattuck-Heidorn, H., Labonte, M. L., & Consoli, T. (2015). Opinion: Focus on preclinical sex differences will not address women's and men's health disparities. *PNAS*, *112*. https://doi.org/10.1073/pnas.1516958112

Štrkalj, G., & Pather, N. (2021). Beyond the Sex Binary: Toward the Inclusive Anatomical Sciences Education. *Anatomical Sciences Education*, *14*(4), 513–518. https://doi.org/10.1002/ase.2002

Do gender role stereotypes still prevail? Measurement invariance of work centrality and job meaningfulness across gender

Marija Davcheva[1], Inés Tomás[1], Vicente González-Romá[1], Ana Hernández[1]

[1]*IDOCAL, University of Valencia, Spain*

Abstract

Measurement invariance analyses across gender groups are a prerequisite for meaningful gender group comparisons. However, measurement invariance studies are typically carried out without a priori hypotheses about expected differences in item responses, which would contribute to understanding the way different gender groups interpret the constructs. The aim of this study is to test measurement invariance in two important work-related constructs, work centrality and job meaningfulness, across gender. Based on Social Role theory and gender stereotype literature, we expect to find differences in the way men and women respond to work centrality and job meaningfulness items (items would be more salient or discriminative for men, whereas women would be less likely to agree with them). Hypotheses were tested in a sample of 704 employees. The results provide support for strong invariance in both scales. Thus, our hypotheses were not supported. Additional comparison of latent means shows that men and women do not significantly differ on any of the constructs. Our study shows that gender stereotypes do not have an impact on the way men and women respond to items of work centrality and job meaningfulness or on their mean levels. These findings illustrate the current societal change in gender roles from more traditional gender role beliefs to more egalitarian ones.

Keywords: Measurement invariance; gender stereotypes; work centrality; job meaningfulness.

Funding: Grant PSI2017-86882-R funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe.

E-mail: Marija.Davcheva@uv.es

1. Introduction

Gender is a salient variable in psychology research, especially in the field of work psychology. A long line of research has focused on gender differences in work-related behaviors, experiences, and perceptions by performing comparisons across gender groups (Cropley & Cropley, 2017). Measurement invariance (MI) is a prerequisite for carrying out meaningful comparisons across groups (including gender groups) in psychology (Tsaousis & Kazi, 2013). However, this practice is often omitted, which can lead to biased results and conclusions (Steyn & De Bruin, 2020). In addition, one of the criticisms of gender measurement invariance studies is that these analyses are sometimes conducted without *a priori* hypotheses about expected differences in item responses (González-Romá et al., 2005; Hatlevik et al., 2017).

One of the prevalent gender theories is Social Role theory (Eagly, 1987), which explains gender differences in behaviors and perceptions due to socially constructed gender roles and their respective stereotypes. These gender stereotypes can produce a variation in item responses between men and women due to a motivation to respond to a measurement instrument according to a specific stereotype (Libbrecht et al., 2014).

Because gender role stereotypes are an important factor in all life domains, including work, in this study we explore whether these stereotypes lead to different conceptualizations and item responses by men and women on work-related constructs. In particular, we focus on job meaningfulness and work centrality as variables influenced by gender stereotypes based on notions about the ideal worker and meaningful paid work (Hofmeister, 2019). The objective of this study is to understand whether gender stereotypes still prevail in two crucial work-related constructs: work centrality and job meaningfulness, by testing their measurement invariance and global differences across groups.

1.1 Gender roles and work centrality

Work centrality refers to the importance that an individual assigns to work in comparison with other life spheres, such as leisure, family, and religion. Gender ideology and social role theory explain how men and women differ in the extent to which work roles are important to them (Davis & Greenstein, 2009). Specifically, life domains that provide resources and status, such as work, are more fundamental to men's identity, whereas life domains that revolve around nurturance and compassion, such as the family and relations, are more central to women's identity (Cinamon & Rich, 2002).

1.2 Gender roles and job meaningfulness

Job meaningfulness refers to the subjective evaluation of the work people do in their jobs as significant, worthwhile, and having positive meaning (Tims et al., 2016). Scholars suggest that job meaningfulness is socially constructed, and that self-concept and identities play a pivotal role in the sense-making process that shapes people's experiences of work meaningfulness. Work can be distinguished as paid and non-paid work. In particular, people's gender identity affects their sense of meaningfulness of different types of work (Baldry et al., 2007, Rosso et al., 2010). However, in the context of job meaningfulness, we refer to finding significance in paid work activities. Hofmeister (2019) suggested that gender stereotypes have typified paid work as being meaningful for men, and emotional work and caregiving as being meaningful for women.

Considering that gender role stereotypes make these two work-related constructs more salient for men, we can expect that items on work centrality and job meaningfulness will be more strongly related to the intended underlying latent traits for men than for women. Thus, items will discriminate better (i.e. have higher factor loadings) in the male group than in the female group. In addition, we expect that the gender stereotypes and expectations for the male role initiate a tendency for men to score higher on these constructs despite having the same latent score (i.e. having higher item intercepts) as women.

Considering these arguments, we hypothesize the following:

H1. *The items on work centrality will be more salient or discriminative for men and less likely to be agreed with (lower "item difficulty or location") by women.*

H2. *The items on job meaningfulness will be more salient or discriminative for men and less likely to be agreed with (lower "item difficulty or location") by women.*

2. Method

2.1. Participants and procedure

We hired the services of a market research company that managed a respondent panel. Employed panel members were invited to participate in the study as long as they were not self-employed. The sample consisted of 704 employees in Spain, 49.6% female, aged between 21 and 59 years (M = 35.4, SD = 9.8 years).

2.2. Measures

Work centrality was measured with Kanungo's (1982) 3-item scale, rated on a 6-point Likert scale (1. Strongly Disagree, 6. Strongly Agree). An example of an item on this scale is "The most important things that are happening to me are related to my work". Cronbach's alpha was .80.

Job Meaningfulness was measured with May et al.'s (2004) 6 -item scale, rated on a 6-point Likert scale (1. Strongly Disagree, 6. Strongly Agree). An example of an item on this scale is "My job activities are personally meaningful to me". Cronbach's alpha was .95.

2.3. Analysis

A Multi-Group Confirmatory Factor Analysis with Latent Mean and Covariance structure (MACS) was conducted. Configural invariance, weak or metric invariance, and strong or scalar invariance were examined in both scales. Configural invariance means that there is a qualitatively invariant measurement pattern of latent constructs across groups. Weak or metric invariance means that item loadings of the invariant configural model (i.e. the item discrimination parameter) are the same across groups. Strong or scalar invariance means that the items' intercepts (i.e., the item locations when the latent mean is at 0) are also equivalent across groups. The MACS analyses were carried out by means of the Mplus 8 software (Muthén & Muthén, 2017).

3. Results

The results of the gender measurement invariance analysis for both scales, work centrality and job meaningfulness, are shown in Table 1.

Table 1. Gender measurement invariance

	χ2	df	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
Work centrality								
Configural invariance	0	0	0	1	1	-	-	-
Weak invariance	0.63	2	0	1	1.01	0	0	.01
Strong invariance	2.43	4	0	1	1.01	0	0	.01
Job Meaningfulness								
Configural invariance	70.93	16	.09	.99	.98	-	-	-
Weak invariance	81.60	21	.09	.99	.98	0	0	0
Strong invariance	85.40	26	.08	.99	.99	.01	0	.01

Note. χ^2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; ΔRMSEA = change in the root mean square error of approximation; ΔCFI = change in the comparative fit index; ΔTLI = change in the Tucker-Lewis index.

Our results provided support for strong or scalar invariance (for which configural and weak invariance have to be met) for both scales. Specifically, the differences in the goodness of fit statistics between the [4]baseline model of configural invariance and the strong invariance model were never larger than .01 (ΔCFI < .001, ΔTLI < .001, ΔRMSEA < .01), and the differences in χ^2 were not significant (work centrality $\Delta\chi^2 = 2.43$, $\Delta df = 4$, $p > .05$; job meaningfulness $\Delta\chi^2 = 14.47$, $\Delta df = 10$, $p > .05$). Thus, our hypotheses were not supported: there were no significant differences between men and women in the factor loadings or item intercepts.

Additionally, a comparison of latent means showed that men and women did not significantly differ on any of the constructs. The work centrality latent mean difference was -.07, $p = .33$, whereas the job meaningfulness latent mean difference was .17, $p = .05$.

4. Conclusions

Our results show measurement invariance across genders for both the work centrality and job meaningfulness scales. Contrary to our expectations, our study results show that gender stereotypes do not have an impact on the way men and women respond to items for two key work-related constructs, work centrality and job meaningfulness, or on their mean levels of these two work-related constructs. Moreover, our study supports the tendency toward a change from traditional gender role beliefs to egalitarian gender role beliefs that has been recently noted in the literature (Cotter et al., 2011).

Our study has some limitations. The sample consists only of Spanish employees, and so it presents limited generalizability. This is especially important because we are studying gender differences in work-related constructs, which are culture sensitive (Neculăesei, 2015). Future studies should test the gender measurement invariance of work centrality and job meaningfulness in samples from different cultures.

Our study adds value to the gender measurement invariance literature and emphasizes the need to test for gender measurement invariance when studying differences between men and women in work-related constructs.

References

Baldry, C., Brain, P., Taylor, P., Hyman, J., Scholarios, D., Marks, A., Bunzel, D. (2007). *The meaning of work in the new economy*. Houndmills, Basingstoke, Hampshire, New York, NY: Palgrave Macmillan

Cinamon, R. G., Rich, Y. (2002). Gender differences in the importance of work and family roles: Implications for work-family conflict. *Sex Roles, 47*, 531–541. https://doi.org/10.1023/A:1022021804846

Cotter, D., Hermsen, J. M., & Vanneman, R. (2011). The end of the gender revolution? Gender role attitudes from 1977 to 2008. *American Journal of Sociology, 117*(1), 259–289. https://doi.org/10.1086/658853.

Cropley, D., & Cropley, A. (2017). Innovation capacity, organisational culture and gender. *European Journal of Innovation Management, 20*(3), 493–510. https://doi.org/10.1108/EJIM-12-2016-0120

Davis, S. N., Greenstein, T. N. (2009). Gender ideology: Components, predictors, and consequences. *Annual Review of Sociology, 35*, 87–105.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Lawrence Erlbaum Associates, Inc

Gonzalez-Roma, V., Tomas, I., Ferreres, D., & Hernandez, A. (2005). Do items that measure self-perceived physical appearance function differentially across gender groups? An application of the MACS model. *Structural Equation Modeling, 12(1),* 148–162. https://doi.org/10.1207/s15328007sem1201_8

Hatlevik, O. E., Scherer, R., & Christophersen, K.-A. (2017). Moving beyond the study of gender differences: An analysis of measurement invariance and differential item functioning of an ICT literacy scale. *Computers & Education, 113*, 280–293. https://doi.org/10.1016/j.compedu.2017.06.003

Hofmeister H. (2019). *Work Through a Gender Lens: More Work and More Sources of Meaningfulness*. In: Yeoman R., Bailey C., Madden A., Thomspon M. The Oxford Handbook of Meaningful Work. (pp.302-325).

Libbrecht, N., De Beuckelaer, A., Lievens, F., & Rockstuhl, T. (2014). Measurement invariance of the Wong and Law Emotional Intelligence Scale scores: Does the measurement structure hold across far Eastern and European countries? *Applied Psychology, 63*(2), 223–237. https://doi.org/10.1111/j.1464-0597.2012.00513.x

May, D. R., Gilson, R. L., & Harter, L. M. (2004). The psychological conditions of meaningfulness, safety and availability and the engagement of the human spirit at work. *Journal of Occupational and Organizational Psychology*, *77*(1), 11-37. https://doi.org/10.1348/096317904322915892

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8)*. Los Angeles, CA: Authors.

Neculăesei, A. N. (2015). Culture And Gender Role Differences. *Cross Cultural Management Journal*, *17(1)*, 31-35.

Kanungo, R. N. (1982). Measurement of job and work involvement. *Journal of Applied Psychology, 67*(3), 341. https://doi.org/10.1037/0021-9010.67.3.341

Rosso, B. D., Dekas, K. H., & Wrzesniewski, A. (2010). On the meaning of work: A theoretical integration and review. *Research in Organizational Behavior, 30*, 91–127. https://doi.org/10.1016/j.riob.2010.09.001

Steyn, R., & De Bruin, G.P. (2020). An investigation of gender-based differences in assessment instruments: A test of measurement invariance. *SA Journal of Industrial Psychology, 46*(0), a1699. https://doi. org/10.4102/sajip.v46i0.1699

Tims, M., Derks, D., Bakker, A. B. (2016). Job crafting and its relationship with person-job fit and meaningfulness: A three-wave study. *Journal of Vocational Behavior, 92*, 44-53. https://doi.org/10.1016/j.jvb.2015.11.007

Tsaousis, I., & Kazi, S. (2013). Factorial invariance and latent mean differences of scores on trait emotional intelligence across gender and age. *Personality and Individual Differences, 54*(2), 169–173. https://doi.org/10.1016/j.paid.2012.08.016

Empirical research in observational methodology (1): Sport and physical activity (I)

Daniel Lapresa[1], M. Teresa Anguera[2]

[1]*University of La Rioja, Logroño, Spain,*

[2]*University of Barcelona, Barcelona, Spain*

1. State of the art

Systematic observation, essentially characterized by focusing on the scientific study of spontaneous or habitual behavior in natural contexts, has not only been consolidated in the last few decades, but the scope of application has been considerably expanded, revealing itself as flexible, useful, and of great rigor, characteristics that constitute its fundamental virtues. Its nature as a scientific method makes it suitable for psychologists in a wide spectrum of research and professional areas.

2. New perspectives and contributions

In this Symposium, four papers are presented, which refer to the field of physical activity and sport, and methodologically special emphasis is given to: (1) *mixed methods*, from questionnaires in an educational context (2) *T-Patterns*, reviewing how over the last two decades they have had increasing applicability, and especially in studies in the field of sport, and (3) analysis of generalizability, while the substantive aspects are soccer, applied to the Spanish team which won the 2012 UEFA European Championship, and chess, where it is clearly novel.

3. Research and practical implications

Observational methodology is increasingly focusing on specific aspects, such as quantitizing, generalizability, coding in indirect observation, *T-Patterns* analysis, stability of sequential analysis, such as polar coordinate analysis, among others, and as a consequence, a large number of works that use observational methodology have been published in journals with a high impact factor. Undoubtedly, the culture of systematic observation is progressively intensifying, being the only possible methodology in a large number of situations, whenever an interest exists in studying spontaneous or habitual behavior, in a non-artificial context, and ensuring that there is visual and/or auditory perceptivity. Furthermore, in this online 9th *European Congress of Methodology,* we are interested in highlighting that we are working within the framework of *mixed methods*, which are currently in a phase of constant growth throughout the world, and we emphasize that observational methodology, according to its profile, can be considered as a *mixed method* in itself, taking into account the QUAL-QUAN-QUAL transition in its successive stages. This consideration opens up a relevant space for increased interest in quantitizing

within observational methodology, leading to a wide spectrum of practical implications in many substantive areas.

Keywords: T-Patterns analysis; generalizability analysis; observational instruments; direct observation; *mixed methods*.

E-mails: daniel.lapresa@unirioja.es; tanguera@ub.edu

Enhancing learner motivation and classroom social climate: A mixed methods approach

Oleguer Camerino[1,2], Alfonso Valero[3], David Manzano-Sánchez[3], Queralt Prat[1], Marta Castañer[1,2]

[1] *National Institute of Physical Education of Catalonia (INEFC), University of Lleida (UdL), Spain,*

[2] *Lleida Institute for Biomedical Research Dr. Pifarré Foundation (IRBLLEIDA), Lleida, Spain,*

[3] *Faculty of Sciences of the Sport, University of Murcia, Spain*

Abstract

The aim of this study was to analyze how motivation and classroom social climate was enhanced in a teaching–learning context through a Pedagogical Model of Personal and Social Responsibility (TPSR). The Observational System of Teaching Oriented Responsibility (OSTOR) revealed how students' behavior patterns shifted during the interventions. The results confirmed have shown an improvement of the TPSR implementation in student responsibility and satisfaction and the social climate of the classroom.

Keywords: teaching strategies; motivational mechanisms; observational analysis

Funding: We gratefully acknowledge the support of the National Institute of Physical Education of Catalonia (INEFC) and the support of a Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* (PGC2018-098742-B- C31) (Ministerio de Economía y Competitividad, Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema I+D+i), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) (SPGC201800X098742CV0), and the support of the Generalitat de Catalunya Research Group, Research group and innovation in designs (GRID). Technology and multimedia and digital application to observational designs (Grant No. 2017 SGR 1405).

E-mails: avalero@um.es; ocamerino@inefc.udl.cat; david.manzano@um.es; qprat@inefc.es; mcastaner@inefc.es

1. Introduction

The term classroom social climate (CSC) refers to how students and teachers perceive the quality of their experiences in the classroom. How they ultimately feel determines their behavior in this setting (Hoy & Miskel, 1996). The term classroom social climate refers to how students and teachers perceive the quality of their experiences in the classroom. Classroom climates depend on a complex ecological framework that includes self-efficacy and the different socioemotional factors that influence this construct (Givens, 2012). In addition to improving student motivation (Martin, et, al., 2016; Makara & Madjar, 2015) and basic psychological needs (autonomy, competence, and relatedness) (Deci & Ryan, 2016), a positive classroom climate stimulates learning and improves academic outcomes (Ainley & Ainley, 2011; Parra, 2010). The Teaching Personal and Social responsibility (TPSR) model improves prosocial behavior and classroom climates (Sanchez Alcaraz, et, al., 2019; Caballero, 2015) and is one of the most powerful tools for promoting self-autonomy in students (Camerino, et.al., 2019).

The model has five levels, which are well described in the literature (Hellison, 2011). The idea is that students gradually work their way up through the levels, but within a flexible set-up that allows them to return to a previous level when necessary. Students at level 0, irresponsibility, for example, do not take responsibility for their acts and blame others. When they reach level 1, respect for the rights and feelings of others, they use negotiation and dialogue to resolve conflicts and disagreements and show respect for other people's qualities and characteristics. At level 2, participation and effort, they are willing to make an effort to achieve goals and they show interest in activities, regardless or not of whether these are aligned with their preferences. At level 3, self-direction, they show autonomy and take ownership of their learning, without instruction from their teachers, while at level 4, caring and leadership, they show empathy and commitment to others without expecting anything in return. Finally, at level 5, transfer, they are capable of transferring what they have learned at previous levels to contexts outside the classroom (e.g., friends, family, and formal and non-formal educational activities).

The aim of this study was to determine the effects of a primary and secondary school intervention aimed at increasing learner autonomy and personal and social responsibility in classroom social climate and its determinants, motivation, basic psychological needs (autonomy, competence, and relatedness), and levels of violence.

2. Materials and Methods

2.1. Participants

An educational program was applied in a primary and a secondary school in a total of 44 sessions -11 for each teacher- during an academic year. Participants were students and their teachers, and they were both selected by accessibility and convenience. There were a total of four teachers with a level of experience between 5 and 10-years teaching in their subjects. They were video-recorded and analyzed in 44 sessions (11 sessions per teacher with a duration of 55 minutes). Two were Physical Education (PE) teachers, one was a History teacher, and another was a Spanish teacher. One of the PE teachers was an experienced teacher in TPSR while the rest were inexperienced teachers in TPSR.

The students: the group was composed of 54 students, between 11 and 16 years old ($M = 13.41$ years, $SD = 1.73$) One class was randomly selected out of all the ones each teacher had. For student age selection, as a point of interest we included the first stage of secondary education, defined according to current legislation in Spain (LOMCE, 2013). Both, the informed parental consent related to the students and the signed consent form from teachers were

obtained in writing. Furthermore, they were informed in accordance with the Declaration of Helsinki and were accepted and verified by the Ethics Committee of the University of Murcia, Spain (ID 1685/2017).

Table 1. Observational System of Teaching Oriented Responsibility (OSTOR) (Camerino et al., 2019).

Criterion	Category	Code	Description
Expectations	Objective of Session	OBS	Prospects and aims of the session
	Objective of Task	OBT	Prospects and aims of the task
Explanations	Imposition Instructions	IMP	Without the possibility to include changes
	Shared	SHA	Proposals are allowed to be decided in common
Organization	Established	EST	Spaces and materials are mandated
	Distribution of Function	DIS	Functions and roles are allocated
	Suggested	SUG	Teachers give opportunities to pupil interventions
Task adjustments	Negative Evaluation	NEG	Rebuke for students
	Redirect	RED	Correct student responses
	Positive Evaluation	POS	Encourage and motivate the students
	Proposals	PRO	Formulate new options to be successful
Student's responses	Reproduction	REP	Replicate tasks or situations
	Unbalances	UNB	Disarranged or disordered responses
	Autonomy and Leadership	AUT	Drive initiatives
	Self-Assessment	SAS	Students evaluate their own performance
Session summary	Guided Summary	GUS	The teacher summarizes the session
	Shared Summary	SHU	The students take part in the session summary
	Nonexistent Summary	NSU	The sessions end without being summarized

2.1. Observational System of Teaching Oriented Responsibility

The OSTOR [23] (Table 1) comprised six criteria. The first four criteria were related to teacher actions: (1) (Expectations); (2) (Explanations); (3) (Organization); (4) (Task adjustments). The fifth criterion was related to the student: (5) (Student's responses). The last criterion was related to how the session concluded: (6) (Session summary). Each criterion was expanded to build an exhaustive and mutually exclusive observation total of 18 categories

2.2. Recording Instrument software LINCE PLUS

The teaching behavior sequences, session by session, were coded using the free instrument software LINCE PLUS (Soto et al., 2019). This software has been designed to facilitate the systematic observation of spontaneous behaviors in any situation or habitual context. It is highly practical and easy to use, and integrates a wide range of functions: coding, recording and enabling data export to several data analysis applications. Besides, LINCE PLUS enables the

obtention of data quality optimizing the data and the verification of the quality of the data between observers. The data obtained was automatically exported to the programs for data analysis THEME software package (Magnusson et al., 2016) for T-pattern detection.

2.3. Procedure

To familiarize students and avoid reactivity behaviors about being observed, a camera for video recording was installed in the classroom several sessions before the beginning of the educational program. The total of the 44 sessions-11 for each teacher- were recorded from the moment the teacher effectively started the educational activities of the session. Two experts, PhDs in physical education, coded the categories of the OSTOR system via LINCE PLUS, which was also used to obtain the reliability between the observers. The function resulted in a kappa statistic of 0.95 for inter-observer and 1.98 for intra-observer analysis.

3. Results

3.1. Evolution behavior from T-Patterns Detected

The implementation of the TPSR of the four teachers caused an evolution of the levels of responsibility of the students. In figure 2, we represent the t-patterns that show this progress; in the first level (A) of the TPRS, student answers are still based on reproduction tasks before an imposition or instruction task; in the second level (B) and the third (C) the behaviors show the corrections and suggestions of the teacher, more autonomous students and major participation and effort; in the fourth level (D) teachers use positive evaluations and shared proposals, showing total autonomy in all the cases.

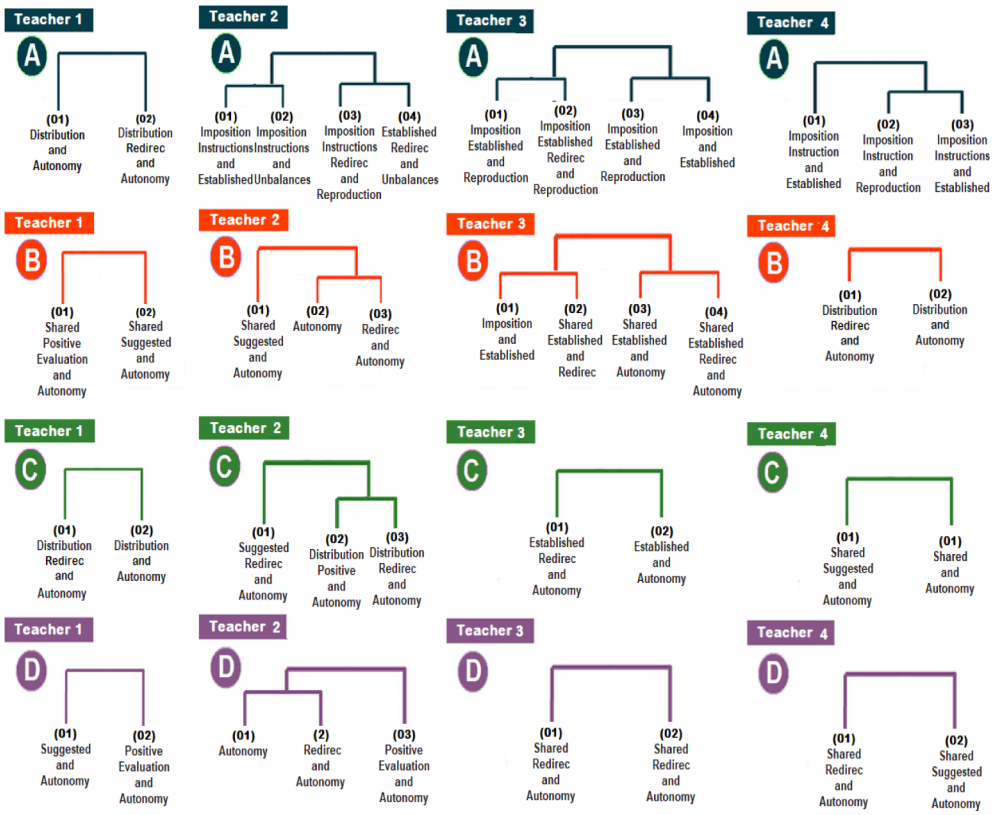


Figure 2. The t-patterns of the four teachers (1, 2, 3 and 4) throughout the four levels (A, B, C and D) of TPSR implementation

4. Discussion

The main objective of this study was to verify how the application of an educational program based on increasing responsibility levels through a TPSR improved CSC, as well as the factors that determine it; motivation, basic psychological needs, and education of violence levels. The findings obtained fit with different studies that reveal the advantages that can be achieved with a TPSR in an educational context related to the improvement of the classroom climate and personal and social positive values, such as responsibility, autonomy and competence (Pozo et al., 2018); (Escartí et al., 2018); (Camerino et al., 2019).

Despite the fact that there are still few studies that have implemented a TPSR in different subjects, these do report improvements in some of the student motivation dimensions, specifically, improvements in intrinsic and introjected motivation (Manzano and Valero-Valenzuela, 2019) or in the self-determination index and reduction in the amotivation values (Manzano-Sánchez & Valero-Valenzuela, 2019). However, in the present work, no significant changes over time have been achieved in students who participated in this program, also reporting a low effect size in all its dimensions, which is therefore not attributable to the sample size either.

5. Conclusions

The mixed methods approach used to analyze changes over time in a group of primary and secondary students was effective in identifying improvements in personal and social responsibility, in the students' autonomy levels and in the classroom social climate.

Finally, assuming that the selection of the sample was made by accessibility and convenience, an unavoidable condition on most occasions when working with natural groups of students in schools, a proposal for improvement in new research is to have a control group to compare the results obtained by students who have participated in the TPSR with those who maintained the teaching used previously. It will be particularly interesting to identify the methodological strategies applied by teachers in the classroom according to the methodology they are using, and to observe possible differences in student behavior patterns.

References

Hoy, W.; Miskel, C. (1996). *Educational administration. Theory, research and practice*; 5th, ed, McGraw-Hil, London.

Givens, R. (2012). Synthesizing the evidence on classroom goal structures in middle and secondary schools: A meta-analysis and narrative review. *Review of Educational Research*, *82*, 396–435.

Martin, A. J.; Papworth, B.; Ginns, P.; Malmberg, L. E. (2016). Motivation, engagement, and social climate: An international study of boarding schools. *Journal of Educational Psychology*, *108*, 772–787. https://doi.org/10.1037/edu0000086

Makara, K. A.; Madjar, N. (2015). The role of goal structures and peer climate in trajectories of social achievement goals during high school. *Developmental Psychology*, *51*, 473–488. https://doi.org/10.1037/a0038801

Deci, E. L.; Ryan, R. M. (2016). Optimizing Students´Motivation in the Era of Testing and Pressur: A Self-Determination Theory Perspective. In *Building Autonomous Learners*, 9–29, Springer, Singapore. https://doi.org/10.1007/978-981-287-630-0_2

Ainley, M.; Ainley, J. (2011). Student engagement with science in early adolescence: the contribution of enjoyment to students' continuing interest in learning about science. *Contemporary Educational Psychology*, *36*, 12.

Parra, P. (2010). Relación entre el nivel de Engagement y el rendimiento académico teórico/práctico. *Revista de Educación en Ciencias de la Salud, 7,* 57–63.

Sánchez-Alcaraz, B. J.; Cañadas, Mª.; Valero, A.; Gómez, A.; Funes, A. (2019). Results, Difficulties and Improvements in the Model of Personal and Social Responsibility. *Apunts. Educación Física y Deportes*, *136*, 62–82. https://doi.org/10.5672/apunts.2014-0983.es.(2019/2).136.05

Caballero, P.(2015). Percepción del alumnado de formación profesional sobre los efectos de un programa de desarrollo positivo (modelo de responsabilidad de Hellison). *Journal of Sport and Health Research*, *7*, 113–126.

Camerino, O.; Valero-Valenzuela, A.; Prat, Q.; Manzano Sánchez, D.; Castañer, M. (2019). Optimizing Education: A Mixed Methods Approach Oriented to Teaching Personal and Social Responsibility (TPSR). *Frontiers in Psychology*, *10*, 1439. https://doi.org/10.3389/fpsyg.2019.01439

Hellison, D. (2011). Teaching personal and social responsibility through physical activity. Human Kinetics, Champaign, IL.

LOMCE. (2013). Ley Orgánica Para la Mejora de la Calidad Educativa. (Organic Law for the Improvement of Educational Quality), *BOE*, España, 8/2013. https://www.boe.es/eli/es/lo/2013/12/09/8/com

Soto, A., Camerino, O., Iglesias, X., Anguera, M. T., Castañer, M. (2019). LINCE PLUS: Research Software for Behavior Video Analysis. *Apunts. Educación Física y Deportes*, *137*,149–153. https://doi.org/10.5672/apunts.2014-0983.es.(2019/3).137.11

Magnusson, M.S.; Burgoon, J.K.; Casarrubea, M. (2016). *Discovering hidden temporal patterns in behavior and interaction. Neuromethods*; Springer: New York, NY. https://doi.org/10.1007/978-1-4939-3249-8

Pozo, P.; Grao-Cruces, A; Pérez-Ordás, R. (2018). Teaching Personal and Social Responsibility Model-Based Programmes in Physical Education: A Systematic Review. *European Physical Education Review*, *24*, 1, 56–75. https://doi.org/10.1177/1356336X16664749

Escartí, A.; Llopis-Goig, R.; Wright, P.M. (2018). Assessing the Implementation Fidelity of a School- Based Teaching Personal and Social Responsibility Program in Physical Education and Other Subject Areas. *Journal of Teaching in Physical Education*, *37*, 1, 12–23. https://doi.org/10.1123/jtpe.2016-0200

Manzano, D.; Valero-Valenzuela, A. (2019). The personal and social responsibility model (TPSR) in the different subjects of primary education and its impact on responsibility, autonomy, motivatoin, self-concept and social climate. *Journal of Sport and Health Research, 11*, 3, 273–288.

An overview of TPA/T-Pattern analysis in sports science over the past 20 years

Gudberg K. Jonsson[1]

[1]*University of Iceland. Reykjavík, Iceland*

Abstract

Purpose: The behavior of all living beings consists of hidden patterns in time; consequently, its nature and its underlying dynamics are intrinsically difficult to perceive and detect by the unaided observer. *Method*: By using a powerful technique known as T-pattern detection and analysis (TPA) it is possible to unveil hidden relationships between behavioral events over time. The technique is built on a unique algorithm that searches for hidden repeated patterns in behavior and interactions, based on a model of the temporal organization of behavior. *Results*: This review will focus on its application in the field of sport, and provide an overview of research carried out in different areas over the past 20 years, i.e., a temporal pattern analysis and its applicability in soccer, boxing, tennis, motor skills, dance and body movement, martial arts, basketball and swimming. *Conclusions*: Over the past two decades there has been a significant increase in the use of TPA/T-Pattern analysis in sport and movement science, both as a single instrument approach or in combination with other methods, i.e., polar coordinates analysis. This increase is also reflected in the number of different group and individual sports that the TPA/T-Pattern analysis is applied to.

Keywords: T-Patterns, TPA, Theme, Polar Coordinates Analysis, Sport-Research.

Funding: This study has been supported by Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mail: gjonsson@hi.is

1. Introduction

The behavior of all living beings consists of hidden patterns in time; consequently, its nature and its underlying dynamics are intrinsically difficult to perceive and detect by the unaided observer. Such a scientific challenge in all behavioral sciences calls for improved means of detection, data handling and analysis. Within the sport coaching process great emphasis is placed on coaches' ability to observe and recall all the critical, discrete incidents in sporting performance. However, it has been shown that coaches cannot accurately observe and recall all of the detailed information that is required for a complete understanding or interpretation of performance (Franks and Miller, 1986). Traditional analysis methods have used frequency of event occurrence as their index of performance. For example, the analyst recorded the number of passes made from particular playing zones or how many times a team/individual makes an unforced error. However, if one accepts the argument that sporting performance consists of a complex series of interrelationships between a wide array of performance variables, then simple frequency data can only provide a relatively superficial view of performance. The challenge for the performance analyst is to find data analysis methods or techniques that can generate more complete, and therefore more complex, quantitative representations of performance (Borrie et al., 2002).

2. Method

By using a powerful technique known as T-pattern detection and analysis (TPA) it is possible to unveil hidden relationships between behavioral events over time (Magnusson, 2000). The technique is built on a unique algorithm that searches for hidden repeated patterns in behavior and interactions, based on a model of the temporal organization of behavior. It considers both the order and the time distances between behavioral event types as well as hierarchical organization. The basic assumption of this methodological approach, embedded in the Theme software, is that the temporal structure of a complex behavioral system is largely unknown, but may involve a set of a particular type of repeated temporal patterns (T-patterns) composed of simpler directly distinguishable event types, which are coded in terms of their beginning and end points (such as "boy begins talking" or "girl ends walking"). The kind of behavior record (as a set of time point series or occurrence time series) that results from this coding of behavior within a particular observation period (here called T-data) constitutes the input to the T-pattern definition and detection algorithms. Essentially, within a given observation period, if two actions, A and B, occur repeatedly in that order or concurrently, are said to form a minimal T-pattern (AB) if they are found more often than expected by chance, assuming as h0 independent distributions for A and B, there is approximately the same time distance (called critical interval, CI) between them. Instances of A and B related by that approximate distance then constitute an occurrence of the (AB) T-pattern and its occurrence times are added to the original data. More complex T-patterns are consequently gradually detected as patterns of simpler already detected patterns through a hierarchical bottom-up detection procedure (see a simple example in Fig. 1). Pairs (patterns) of pairs may thus be detected, for example, ((AB)(CD)), (A(KN))(RP)), etc. Special algorithms deal with potential combinatorial explosions due to redundant and partial detection of the same patterns using an evolution algorithm (completeness competition), which compares all detected patterns and lets only the most complete patterns survive. As any basic time unit may be used, T-patterns are in principle scale-independent, while only a limited range of basic unit size is relevant in each concrete study.

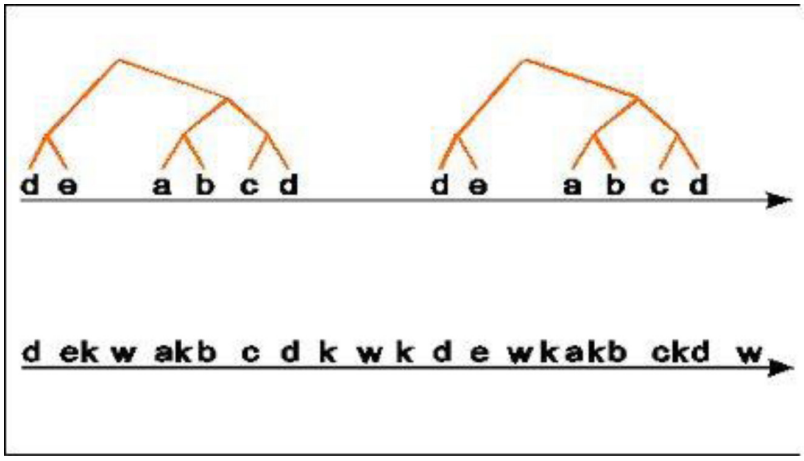


Figure 1. The lower part of this figure shows a simple real-time behavior record containing a few occurrences of several event types, a, b, c, d, & e, indicating their respective instances within the observation period. The upper line is identical to the lower one, except that occurrences of k and w have been removed. A simple t-pattern (abcd) then appears, which was difficult to see when the other events were present.

3. Results

TPA has been successfully applied in the study of various aspects of human or animal behavior, such as behavioral modifications in neuro-psychiatric diseases, interaction between human subjects and animals, artificial agents and sporting and physical activities. This review will focus on its application in the field of sport, and provide an overview of research carried out in different areas over the past 20 years, i.e., a temporal pattern analysis and its applicability in soccer, boxing, tennis, motor skills, dance and body movement, martial arts, basketball and swimming (Borrie et al., 2002; Amatria et al,, 2017; Castañer et al., 2017; Ibánez et al., 2018). The example in figure 2 shows a detected pattern occurring three times during the first half of

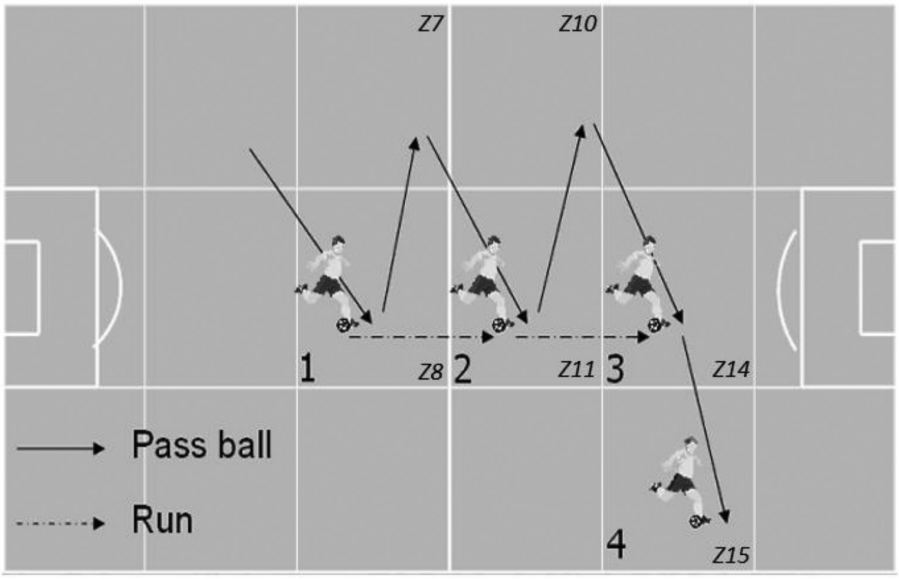


Figure 2. Schematic representation of the pattern, 1) Player A receives the ball in Zone 8, passes the ball to a team mate in Zone 7and runs forward. 2) Player A receives the ball in Zone 11, passes the ball to a team mate in Zone 10 and runs forward. 3) Player A receives the ball in Zone 14, passes the ball to a team mate in Zone 15.

a national game between Iceland and France in 1998, with the same order of events occurring with a significant similar time interval between each event on each occasion.

4. Conclusions

The number, frequency and complexity of detected patterns in different sporting and physical activities indicates that behavior is even more synchronized than the human eye can detect. This synchrony was found to exist on different levels, with highly complex time structures that extended over considerable time spans where many of the patterns occurred in a cyclical fashion.

Traditional frequency analyses of performance have provided, and continue to provide, valuable information that coaches and performers use to enhance the coaching process. It is not the assertion of this paper that an analysis of temporal structure is better than other analysis approaches, merely that an analysis of temporal structure provides an additional, fresh perspective for performance analysis to consider and use.

Over the past two decades there has been a significant increase in the use of TPA/T-Pattern analysis in sport and movement science, both as a single instrument approach or in combination with other methods, i.e., polar coordinates analysis. This increase is also reflected in the number of different group and individual sports that the TPA/T-Pattern analysis is applied to.

References

Amatria, M., Lapresa, D., Arana, J., Anguera, M.T., & Jonsson, G.K. (2017). Detection and selection of behavioral patterns using Theme: a concrete example in grassroots soccer. Sports, 5, 20; doi:10.3390/sports5010020.

Borrie, A.; Jonsson, G.K.; Magnusson, M.S. (2002). Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. Journal of Sports Sciences, 20, 845–852.

Castañer M, Barreira D, Camerino O, Anguera MT, Fernandes T and Hileno R (2017) Mastery in Goal Scoring, T-Pattern Detection, and Polar Coordinate Analysis of Motor Skills Used by Lionel Messi and Cristiano Ronaldo. Front. Psychol. 8:741. doi: 10.3389/fpsyg.2017.00741

Franks, I.M. and Miller, G. (1986). Eyewitness testimony in sport. Journal of Sport Behavior, 9, 39–45.

Ibánez R, Lapresa D, Arana J, Camerino O, Anguera M (2018) Observational analysis of the technical-tactical peformance of elite karate contestants. CCD 13(14): 61–70.

Magnusson, M.S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. Behavior, Research Methods, Instruments & Computers, 32, 93–110.

Successful behaviors in a high-performance champion football team: detection of T-patterns

Rubén Maneiro[1], Mario Amatria[1], and M. Teresa Anguera[2]

[1]*Pontifical University of Salamanca, Salamanca,*

[2]*University of Barcelona, Barcelona*

Abstract

Sports performance analysis is an area of study that attempts to describe or predict successful behaviors in high-performance football. In this study, we performed an in-depth analysis of play by the Spanish football team during the 2012 UEFA European Championship, where it was crowned champion. Observational methodology was used (Anguera and Blanco-Villaseñor, 2003), since it is the one that best fits the evaluation of sports behavior. The T-patterns detection statistic was used to identify the hidden regularity patterns made by the team. Following Anguera, Blanco-Villaseñor and Losada, (2001), an ideographic, punctual and multidimensional design was used. We identified hidden patterns of play that ended in a goal for the Spanish team. A generalizability coefficient (e2) of 0.986 demonstrated that the offensive patterns detected were robust and highly generalizable. These patterns were formed by technical actions consisting of ball control and passes, with alternations between short and long passes, in the central area of the rival's pitch, with the use of both wings to achieve width of play and prioritization of width over depth of play.

Keywords: football; t-patterns; high performance, observational methodology

Funding: The authors gratefully acknowledge the support of a Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] and *Mixed method approach on performance analysis (in training and competition) in elite and academy sport* [PGC2018-098742-B-C33] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which are part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mail: rmaneiro@upsa.es

1. Introduction

When studying performance in soccer, one could assume that by analyzing success or failure indicators related to the performance of individual players or the team as a whole, an accurate picture of the match as a whole could be obtained. However, as pointed out by Lago (2005), there is always an element of chance and unpredictability in team sports. Players, trainers, and fans largely agree that chance is sometimes important for understanding the result of a match.

T-pattern detection has enormous potential in applied research and in interdisciplinary areas such as sport, where researchers are interested not only in quantifying performance indicators, such as goals, passes, or shots, but also in qualifying the steps that lead up to these actions. T-pattern analysis can detect the structures that trigger what can be termed as a successful action in soccer. Numerous studies have used T-pattern analysis to identify these invisible structures that underlie all dimensions of soccer through algorithmic computations and have demonstrated that the results can have important practical implications (Barlett, Button, Robins, Dutt-Mazunder, & Kennedy, 2012).

The aim of this study was to analyze offensive play by the Spanish soccer team during the 2012 UEFA European Championship (UEFA Euro 2012) through the detection of T-patterns reflecting intrinsic patterns of play established during the spontaneous course of the game.

2. Method

We used observational methodology (Anguera, 1979) and applied the observational design I/P/M, which stands for Idiographic/Point/Multidimensional.

2.1 Observation instrument

The observation instrument proposed by Maneiro and Amatria (2018) was used, for consultation about the instrument's criteria and categories.

2.2 Software tools

Data were annotated using the free software tool Lince (v. 1.2.1; Gabin, Camerino, Anguera, & Castañer, 2012)

2.3 Procedure

The observation sample for the offensive actions by the Spanish national team during the UEFA EURO 2012 contained 6861 events, corresponding to 5005 technical actions and 746 offensive sequences. The data were type IV data, which means they are concurrent and time-based (Bakeman, 1978).

3. Results

3.1. T-pattern detection

A total of 1465 T-patterns that met the search criteria were detected in the full dataset of offensive play by the Spanish national team during the UEFA Euro 2012. There were 987 two-cluster patterns, 387 three-cluster patterns, 72 four-cluster patterns, 16 five-cluster patterns, and 3 six-cluster patterns.

The results presented below were generated by the application of the automatic (quantitative) sort settings (Amatria, Lapresa, Arana, Anguera & Jonsson, 2017) in THEME, v. Edu.

They show the T-patterns with the highest number of occurrences, the highest number of clusters, and the longest duration (Table 1, Figure 1).

Table 1. T-patterns detected using the automatic sort settings in THEME, v. Edu.

Setting	Code	String-like pattern	O / L/ D	Mean Internal Interval (in frames)
Occurrences (O)	O.1	(zi61,zf61,c1 zi61,zf61,c1)	35 / 2 / 3388	95.80
Length (L)	L.1	((((zi61,zf61,c1 zi61,zf61,c1)(zi61,zf71,c2 zi71,zf71,c2)) zi110,zf110,c1) zi110,zf110,ioc)	3 / 6 / 1718	20.00 /103.67 /126.00 / 300.00 / 22.00
	L.2	((zi61,zf61,c1 zi61,zf61,c1)((zi61,zf71,c2 zi71,zf71,c2)(zi110,zf110,c1 zi110,zf110,c1)))	3 / 6 / 1707	20.00 / 103. 67 / 126.00 / 300.00 / 18.33
	L.3	(zi71,zf71,c2 ((((zi71,zf71,c1 zi71,zf71,c1) zi71,zf71,c2) zi71,zf120,c2) zi71,zf71,c2))	3 / 6 / 2984	65.67 / 47.67 / 158.33 / 265.67 / 456.33
Duration	D.1	(zi71,zf71,c2 (zi71,zf71,c1 zi71,zf61,c2))	10 / 3 / 4371	245.30 / 190.80

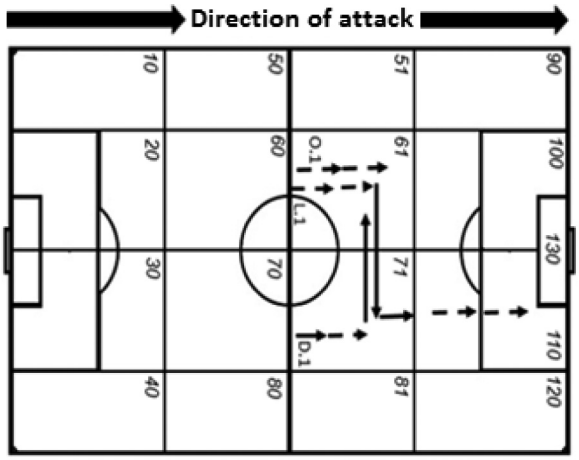


Figure 1. T-patterns following application of automatic sort settings = C1 and = C2.

In the next section, we present the T-patterns detected using the qualitative filters (Amatria et al., 2017) applied to answer the questions posed in this study. These were patterns related to both the depth of play (i.e., movement of the ball from one sector of the pitch up to another one) and the width of play (i.e., movement of the ball from one side of the pitch to the other). We used four qualitative filters, or levels of success, to analyze offensive performance in relation to depth of play. These levels of success, which are the equivalent of optimal targets (Hugues & Bartlett, 2002) were defined as follows: a) sequences of play ending in the definition sector

(Level IV), b) sequences of play ending in a pass to the goal area (Level III), c) sequences of play ending in a shot on goal (Level II), and d) sequences of play ending in a goal, which is the ultimate measure of success. These four success levels followed a hierarchy ranging from the least complex (Level IV) to the most complex (Level I).

It should be noted that only T-patterns that do not appear at lower levels are shown for a given level. For example, although patterns detected at Level I are also present at Levels II, III, and IV, they are shown only at the top level.

Success Level IV shows the T-patterns ending in the definition sector (zones 90, 100, 110, 120, and 130). This success level is relevant, because it shows the progression that takes place while the team is building an attack. Just one T-pattern was detected in this case (Table 2 and Figure 2).

Table 2. T-patterns ending in the definition sector detected using pre-established sort settings.

Code	String-like pattern	O / L/ D	Media Ii
D.1	(zi61,zf90,c3 zi90,zf90,c2)	7 / 2 / 1052	149,29

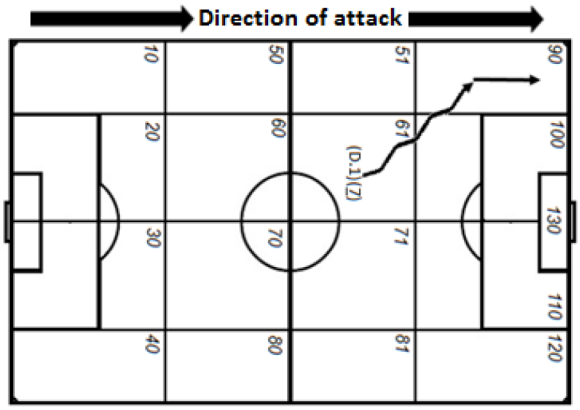


Figure 2. Graphic representation of T-patterns detected, where ● = corner kick, P = loss of possession ➖➡ = C1, ➡ = C2, ↝ = C3 and ✖ = shot on goal against team being observed.

Level III contains T-patterns corresponding to sequences of play ending in a pass to zones 100, 110, and 130 (goal area). These patterns are obviously valuable, as they can show the actions that lead up to a ball being delivered to the immediate goal area. We detected 12 T-patterns at this level. Ten of these corresponded to sequences of play in the central areas of the pitch, and two to sequences in the lateral areas (Table 3 and Figure 3).

Table 3. T-patterns ending with delivery of the ball to the goal area detected using pre-established sort settings.

Code	String-like pattern	O / L/ D	Media Ii
A.1	(zi61,zf100,c3 zi100,zf100,p)	11 / 2 / 1561	140.91
A.2	(zi120,zf120,ffse zi120,zf110,c1)	11 / 2 / 25	1.27
A.3	(zi61,zf61,c2 zi61,zf100,c3)	9 / 2 / 1491	164.67
A.4	(zi61,zf100,c2 zi100,zf100,p)	8 / 2 / 807	99.88

Code	String-like pattern	O / L/ D	Media Ii
A.5	(zi61,zf61,c2 zi61,zf100,c2)	7 / 2 / 981	139.14
A.6	(zi71,zf110,c3 zi110,zf110,p)	7 / 2 / 747	105.71
A.7	(zi61,zf71,c2 zi71,zf110,c3)	7 / 2 / 1259	178.86
A.8	(zi61,zf90,c2 zi100,zf100,p)	7 / 2 / 1937	275.71
A.9	(zi70,zf61,c2 zi100,zf100,p)	7 / 2 / 2879	410.29
A.10	(zi71,zf110,c3 zi110,zf110,c1)	7 / 2 / 634	89.57
A.11	(zi71,zf71,c2 zi71,zf110,c3)	7 / 2 / 1176	167.00
A.12	(zi120,zf120,ffse zi130,zf130,cfff)	7 / 2 / 886	125.57

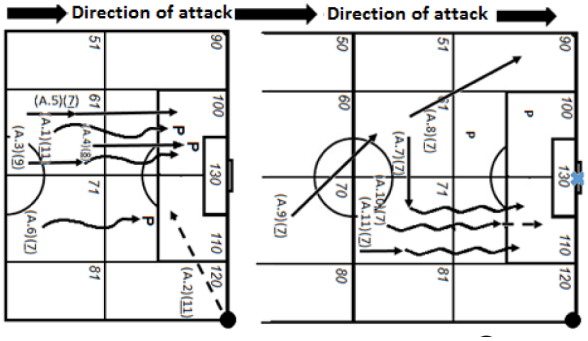


Figure 3. Graphic representation of T-patterns detected, where ● = corner kick, P = loss of possession ▬ ➔ = C1, ➔ = C2, ↝ = C3 and ✖ = shot on goal against team being observed.

Level II shows T-patterns corresponding to sequences of play that contain at least one shot on goal, regardless of whether this was successful or not. Again, these patterns are important, as they reflect the occurrence of actions aimed at scoring a goal. The majority of T-patterns detected at Level II occurred in zone 130, the rival goal area (Table 4 and Figure 4).

Table 4. T-patterns detected using pre-established search settings that contain an end move.

Code	String-like pattern	O / L/ D	Mean Ii
F.1	(zi110,zf130,f zi130,zf130,cfff)	12 / 2 / 788	64.67
F.2	(zi100,zf130,c1 zi100,zf130,f)	11 / 2 / 44	3.00
F.3	(zi110,zf130,c1 zi110,zf130,f)	9 / 2 / 28	2.11
F.4	(zi120,zf120,ffse zi110,zf130,f)	7 / 2 / 680	96.14
F.5	(zi110,zf110,f zi110,zf110,ioc)	7 / 2 / 68	8.71
F.6	(zi110,zf130,f zi130,zf130,p)	7 / 2 / 136	18.43
F.7	(zi61,zf71,c2 zi110,zf130,f)	7 / 2 / 1753	249.43

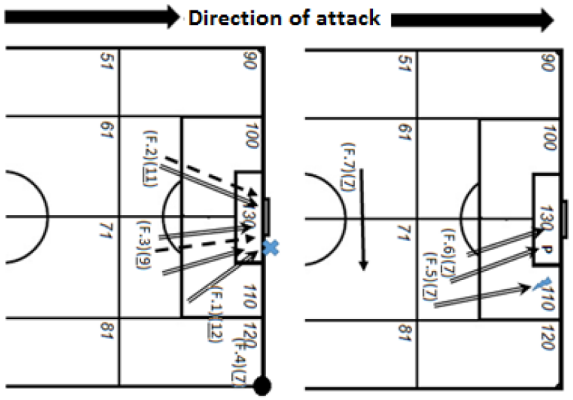


Figure 4. Graphic representation of T-patterns detected, where ● = corner kick, P = loss of possession ➖➡ = C1, ➡ = C2, ↝ = C3 and ✖ = shot on goal against team being observed.

Level 1 shows T-patterns corresponding to sequences of play that end in a goal, the ultimate measure of success in soccer. Five T-patterns were detected at this level (Table 5 and Figure 5).

Table 5. T-patterns ending in a goal detected using pre-established search settings.

Code	String-like pattern	O / L/ D	Media Ii
G.1	(zi100,zf130,f zi130,zf130,gf)	6 / 2 / 262	42.67
G.2	(zi110,zf130,f zi130,zf130,gf)	6 / 2 / 149	23.83
G.3	((zi100,zf130,c1 zi100,zf130,f) zi130,zf130,gf)	4 / 3 / 99	1.25 / 22.50
G.4	(zi61,zf61,r (zi100,zf130,f zi130,zf130,gf))	3 / 3 / 826	216.67 / 57.67
G.5	(zi51,zf50,c1 zi130,zf130,gf)	3 / 2/ 1761	586.00

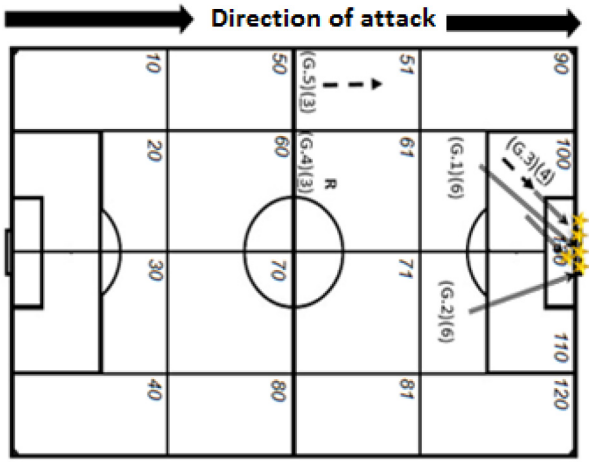


Figure 5. Graph showing T-patterns ending in a goal, where ➖➡ = C1, ⇒ = shot, and ★ = goal.

4. Discussion

The aim of this study was to apply T-pattern analysis to identify strings of events that occur intrinsically and spontaneously during the course of a soccer match but remain invisible to the naked eye.

In relation to the most frequent T-patterns corresponding to offensive sequences involving use of the two outer corridors, we observed that changes of direction were achieved by both passing and dribbling. Such strategies are designed to achieve a greater width of play. Crossing the ball from one side of the pitch to the other is not an easy task (Garganta, 1997). Our results in this respect differ from those of Castellano (2000). Identifying T-patterns of this type is important, as they describe effective sequences of play in which the attacking team avoids the more crowded central corridor (Anguera, 2004; Garganta, 1997).

5. Conclusions

We have used T-pattern detection to identify and describe aspects of the successful attacking style of the champions of the UEFA Euro 2012. Apart from shedding light on some of the secrets of the Spanish team's success, our results also serve to build on previous findings and contribute to a better understanding of what occurs within the deeper layers of a soccer match.

Our results can be summed up in two main points:

a) To make a shot on goal and score, the Spanish national team simultaneously makes good use of the width and depth of the pitch to create space through team and individual actions.

b) The Spanish team prioritizes width of play over depth of play to find space in which to build its attack and make a shot on goal or score.

References

Amatria, M., Lapresa, D., Arana, J., Anguera, M.T., & Jonsson, G.K. (2017). Detection and selection of behavioral patterns using Theme: a concrete example in grassroots soccer. *Sports, 5*, 20; doi:10.3390/sports5010020.

Anguera, M.T. (1979) Observational Typology. *Quality & Quantity. European-American Journal of Methodology, 13*(6), 449–484.

Anguera, M.T., Arnau, J., Ato, M., Martínez, R., Pascual, J., y Vallejo, G. (1995). *Métodos de investigación en psicología* [Research methods in psychology]. Madrid: Síntesis.

Bakeman, R. (1978). Untangling streams of behavior: Sequential analysis of observation data. In G.P. Sackett (Ed.), *Observing behavior, Vol. 2: Data collection and analysis methods* (pp. 63–78). Baltimore: University of Park Press.

Barlett, R., Button, C., Robins, M., Dutt-Mazunder, A., & Kennedy, G. (2012). Analysing team coordination patterns from player movement trajectories in soccer: Methodological considerations. *International Journal of Performance Analysis in Sport, 12*, 398–424.

Castellano, J. (2000). *Observación y análisis de la acción de juego en el fútbol*. Universidad del País Vasco: Tesis doctoral.

Gabin, B., Camerino, O., Anguera, M.T., Castañer, M. (2012). Lince: Multiplatform sport analysis software. *Procedia. Social and Behavioral Sciences, 46*, 4692–4694.

Garganta, J. (1997). *Modelaçao táctica do jogo de futebol. Estudo da organizaçao da fase ofensiva em equipas de alto rendimento*. Universidade do Oporto: Tesis doctoral.

Lago, C. (2005). Ganar o perder en el fútbol de alto nivel. ¿Una cuestión de suerte? [Winning or losing in elite soccer. A question of luck?]. *Motricidad. European Journal of Human Movement, 14*, 137–152.

Maneiro, R., y Amatria, M. (2018) Polar Coordinate Analysis of Relationships With Teammates, Areas of the Pitch, y Dynamic Play in Soccer: A Study of Xabi Alonso. *Frontiers in Psychology, 9*:389. doi: 10.3389/fpsyg.2018.00389

Observational analysis of illegal movements in chess initiation

Jorge Miranda[1], Daniel Lapresa[1], Javier Arana[1], M. Teresa Anguera[2]

[1]*Department of Educational Sciences, University of La Rioja, Logroño, Spain,*

[2]*Faculty of Psychology, University of Barcelona, Barcelona, Spain*

Abstract

Purpose: The present work has two objectives. The first is the creation of an observation system that analyzes illegal movements in chess initiation. The second objective aims to analyze illegal movements in the initiation of chess. *Method*: Based on a detailed analysis of the regulation, an *ad hoc* observation instrument was prepared, guaranteeing the reliability of the observation system -in the form of concordance-; the validity of the observation instrument in the theoretical framework of the theory of generalizability; and the generalizability of the results obtained with the illegal movements registered. *Results*: The results obtained in the analysis of illegal movements reveal the difficulties that children (under 12 years of age) find in understanding and playing chess. *Conclusions*: The second objective, which aims to analyze illegal movements in the initiation of chess, enabled the categorization of the types of illegal actions committed by chess players in Primary Education.

Keywords: Observational methodology, chess, illegal movements, reliability, generalizability, adjusted residual analysis.

Funding: The authors gratefully acknowledge the support of a Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0]. In addition, authors thanks the support of the Generalitat de Catalunya Research Group, *GRUP DE RECERCA I INNOVACIÓ EN DISSENYS (GRID). Tecnología i aplicació multimedia i digital als dissenys observacionals* [Grant number 2017 SGR 1405].

E-mails: jorge_7_95@hotmail.com; daniel.lapresa@unirioja.es; javier-sabino.arana@unirioja.es; tanguera@ub.edu

1. Introduction

Playing chess among children develops skills such as attention, concentration, identification and resolution of problems, the development of planning strategies, creativity, and empathy with the rival, etc. (Christiaen and Verholfstadt, 1978; Storey, 2000; Trinchero, 2013).

Illegal movements show us the difficulties that children face when approaching the game of chess, being a clear indicator of the complexity of the game while being initiated and of the barriers that children encounter in understanding and playing the game.

According to article 3.10.2 of the FIDE Chess Laws (2017), a move is legal when all the relevant requirements of articles 3.1 to 3.9 have been met. Not every violation of the rules of chess involves an "illegal" movement or action; only those that are stated as an illegal action, the rest being "irregularities". Illegal movements entail the corresponding sanctions -article 7.5.3 of the FIDE Chess Laws (2017)-.

The objectives of this work are: a) to present an observation system to analyze illegal movements in initiation chess; b) analyze the type of illegal movements in chess players in Primary Education.

2. Method

The present work has been developed within the use of observational methodology (Anguera, 1979). The observational design, according to Anguera, Blanco-Villaseñor, Hernández-Mendo and Losada (2011), was nomothetic -39 chess players in the sub-12 category who did not have an Elo-FIDE reference score (https://ratings .fide.com)-, punctual -it was not intended to track the illegal movements committed by the participants, but to accumulate illegal games- intra-session monitoring -throughout each of the 101 registered games-; and multidimensional -reflected in the different criteria of the observation instrument-.

In the present work an exceptional enclave was produced in which direct and indirect observation were contemplated in an integrated way, since direct observation was required (Sánchez-Algarra and Anguera, 2013) through visual perception and indirect observation (Anguera, Portell, Chacón-Moscoso, and Sanduvete-Chaves, 2018) from the reasoning and strategy that each player has cognitively created taking into account -or not, as in illegal movements- the rules that regulate chess.

2.1. Observation instrument

The observational instrument was developed *ad hoc*, as a combination of field format and category systems (Anguera, Magnusson and Jonsson, 2007). Based on the information contained in the aforementioned articles of the FIDE Chess Laws (2017), the types of illegal movements dimension was configured: 1) Castling 2) Movement of the piece pinned by the king 3) Movement of the king to a threatened square 4) Incorrect movement of the piece 5) Promotion of the pawn 6) Problems of square occupation 7) Movement of a piece that does not remove the check. The complete observation instrument can be found in free access in Miranda et al. (2019), https://revistas.um.es/cpd/article/view/370871.

2.2. Registration and Coding

The game record was moved from the spreadsheet completed by the players to the chess program *Chessbase*, version 14. For registration with the *software* Lince (Gabin, Camerino, Anguera, and Castañer, 2012), video cuts were generated from the captures corresponding to

each illegal movement. According to Bakeman and Quera (1996), multi-event data was handled in the present work.

3. Results

The reliability of the observation system was guaranteed from the results of Cohen's Kappa coefficient that show an *almost-perfect* consideration of the agreement according to the classic reference values established by Landis and Koch (1977).

Within the generalizability theory, the generalizability of the results obtained with the 101 illegal movements was endorsed, obtaining a relative generalizability coefficient, corresponding to the measure plan [Categories] / [Illegal], of e^2=0.96. On the other hand, the relative generalizability coefficient (e^2= 0.00) resulting from the [Illegal] / [Categories] measurement plan guarantees the validity of the observation instrument constructed *ad hoc*, in the theoretical framework of the theory of generalizability (Blanco-Villaseñor, Castellano, Hernández-Mendo, Sánchez-López and Usabiaga, 2014; García-García, Hernández-Mendo, Serrano and Morales-Sánchez, 2013).

To find out whether the data of the illegal movements dimension was proportionally distributed among the seven categories that make up this dimension (castling=17.8%; pinned piece=23.7%; king to threatened square=30.6%; movement not corresponding to the piece=5.9%; promotion of the pawn=1.9%; problems of occupation of squares=0.9%; movement that does not remove the check=18.8%) the goodness of fit test χ^2 was used, which compares the observed frequencies with the expected ones, under the hypothesis that the data is evenly distributed among the different categories that make up the variable. The registered categories that made up the illegal move type dimension were not distributed proportionally (χ^2=55.842; p<0.001).

The adjusted residual analysis (Allison and Liker, 1982) carried out in two levels of analysis is found in Miranda et al. (2019), https://revistas.um.es/cpd/article/view/370871. The first level corresponds to the dimensions: type of illegal, side, phase, piece, adequacy, involvement and pressure. The second level of analysis deepens the relationship between: a) the "type of illegal" and the dimensions: material gain, threatened piece, type of castling, no right to castling, not right and inappropriate movement, causes inappropriate movement and promotion of pawn-; b) the received "pressure" and the dimensions: area in which the offending side does not receive pressure, first piece that exerts pressure on the non-offending side, second piece, distance from the first piece that exerts pressure, second distance, distance between pinned piece and king, area of the board where the first piece that exerts pressure is located, area of the board where the second piece that exerts pressure is located, area of the board where the king of the offending side is located; and area of the piece that commits the illegal movement.

4. Conclusions

The results show the difficulty for children of this age to remember the spatial relationships that occur at all times with the king (types of illegal: "king to a threatened square" and "pinned piece"). In addition, the difficulties that children face in the act of castling when having to assess different facets of the situation: different target positions in short and long castling; and the right to castling. The "movement not corresponding to the piece", which incorporates spatial relationships relative to a single piece, has less presence, but is very relevant when pointing out the figures with which children have greater difficulties. The illegal type, "pawn promotion" is less frequent as it corresponds to its realization, generally in the final phase of the game. The type of illegal "box occupation problems" shows a residual presence.

The results obtained from the analysis of each of the types of illegal movements characterized will facilitate the optimization of chess teaching programs, stressing the difficulties detected.

References

Anguera, M.T. (1979). Observational Typology. *Quality & Quantity. European-American Journal of Methodology, 13*(6), 449–484.

Anguera, M.T., Blanco-Villaseñor, A., Hernández-Mendo, A. y Losada, J.L. (2011). Diseños observacionales: Ajuste y aplicación en psicología del deporte. *Cuadernos de Psicología del Deporte, 11*(2), 63–76.

Anguera, M.T., Magnusson, M.S. y Jonsson, G.K. (2007). *Instrumentos no estandar: planteamiento, desarrollo y posibilidades*. Avances en medición, 5, 63–82.

Anguera, M.T., Portell, M., Chacón-Moscoso, S., y Sanduvete-Chaves, S. (2018). Indirect observation in everyday contexts: Concepts and methodological guidelines within a mixed methods framework. *Frontiers in Psychology, 9*:13. https://doi.org/10.3389/fpsyg.2018.00013

Bakeman, R. y Quera, V. (1996). *Análisis de la interacción. Análisis secuencial con SDIS y GSEQ*. Madrid: Ra-Ma.

Blanco-Villaseñor, A., Castellano, J., Hernández-Mendo, A., Sánchez-López, C.R. y Usabaiaga, O. (2014). Aplicación de la TG en el deporte para el estudio de la fiabilidad, validez y estimación de la muestra. *Revista de Psicología del Deporte, 23*(1), 131–137.

Christiaen, J. y Verholfstadt, D. C. (1978). Chess and cognitive development. *Nederlandse Tydschrift voorde Psychologie en haar Grensegebieten*, 36, 561–582.

Gabin, B., Camerino, O., Anguera, M. T. y Castañer, M. (2012). *Lince: multiplatform sport analysis software*. Procedia-Social and Behavioral Sciences, 46, 4692–4694. https://doi.org/10.1016/j.sbspro.2012.06.320

García-García, O., Hernández-Mendo, A., Serrano, V. y Morales-Sánchez, V. (2013). Aplicación de la teoría generalizabilidad a un análisis de tensiomiografía en ciclistas profesionales de ruta. *Revista de Psicología del Deporte, 22*(1), 53–60.

Landis, J.R. y Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. https://doi.org/10.2307/2529310

Miranda, J., Lapresa, D., Arana, J., Iza, A. y Anguera, M.T. (2019). Análisis observacional de los movimientos ilegales en la iniciación al ajedrez: identificando dificultades en el entendimiento del juego. Cuadernos de Psicología del Deporte, 19(3), 90–101, https://doi.org/10.6018/cpd.370871

Storey, K. (2000). Teaching beginning chess skills to students with disabilities. *Preventing School Failure: Alternative Education for Children and Youth*, *44*(2), 45–49.

Trinchero, R. (2013). *Can Chess Training Improve Pisa Scores in Mathematics? An Experiment in Italian Primary School*. Paris: Kasparov Chess Foundation Europe.

Empirical research in observational methodology (2): Sport and physical activity (II)

Marta Castañer[1], M. Teresa Anguera[2]

[1]*University of Lleida, Lleida, Spain,*

[2]*University of Barcelona, Barcelona, Spain*

1. State of the art

Systematic observation, essentially characterized by focusing on the scientific study of spontaneous or habitual behavior in natural contexts, has not only been consolidated in the last few decades, but the scope of application has been considerably expanded, revealing itself as flexible, useful, and of great rigor, characteristics that constitute its fundamental virtues. Its nature as a scientific method makes it suitable for psychologists in a wide spectrum of research and professional areas.

2. New perspectives and contributions

In this Symposium, four papers are presented, centering on the field of physical activity and sport, and specifically concerning substantive aspects of fencing, judo, women's soccer, and futsal. From a methodological side, special emphasis is given to: (1) decision trees, applied to the effectiveness of combat actions (2) *T-Pattern* analysis, applied to technical-tactical actions in judo (3) univariate, bivariate and multivariate analysis, in a study on elite women's soccer, and (4) a systematic review, carried out giving special relevance to the methodological quality of the primary documents.

3. Research and practical implications

Observational methodology is increasingly focusing on specific aspects, such as quantitizing, generalizability, coding in indirect observation, *T-Patterns* analysis, stability of sequential analysis, or polar coordinate analysis, among others, and as a consequence, a large number of works that use observational methodology have been published in journals with a high impact factor. Undoubtedly, the culture of systematic observation is progressively intensifying, being the only possible methodology in a large number of situations, whenever an interest exists in studying spontaneous or habitual behavior, in a non-artificial context, and ensuring that there is visual and/or auditory perceptivity. Furthermore, in this online 9th *European Congress of Methodology,* we are interested in highlighting that we are working within the framework of *mixed methods*, which are currently in a phase of constant growth throughout the world, and we emphasize that observational methodology, according to its profile, can be considered as a *mixed method* in itself, taking into account the QUAL-QUAN-QUAL transition in its successive stages. This consideration opens up a relevant space for increased interest in quantitizing

within observational methodology, leading to a wide spectrum of practical implications in many substantive areas.

Keywords: Decision trees; T-Patterns analysis; systematic review; direct observation; *mixed methods*.

E-mails: mcastaner@inefc.es; tanguera@ub.edu

Pattern recognition in fencing strategy using decision trees: Elite foil

Xavier Iglesias[1], Rafael Tarragó[1], Laura Ruiz-Sanchis[2]

[1]*INEFC – Universitat de Barcelona, Barcelona, Spain,*

[2]*Universidad Católica de Valencia, San Vicente Mártir, Valencia, Spain*

Abstract

The aim was to determine the effectiveness of elite foil fencers based on the application of decision-tree analysis. A nomothetic, punctual & multi-dimensional design was used. 13 male foil (MF) and 12 female foil (FF) combats were recorded. ESGRIMOBS and Lince were utilized as observational and recording instruments. The fencer who made the first attack was "A", and their rival "B". "A or B" pressure was analyzed. The piste zones were: End_A, End_B and centre. A decision-tree model was applied. The differences in distribution were checked with a chi-square.1509 actions were analyzed. 67.1% were Pres_A, 13.5% Pres_B. "A" won 25.6% and "B" won 14.6%. There was no relationship between pressure, piste and effectiveness (n.s.). In FF (n=677), Pres_A (68.7%) got 23.9% A_Touch and 16.8% B_Touch. Pres_B (11.7%) got 30.4% A_Touch and 16.5% B_Touch. In MF, Pres_A (65.7%) got 23.4% A_Touch and 14.4% B_Touch. Pres_B (15.0%) got 31.2% A_Touch and 7.2% B_Touch. Combat conventions could determine different effectiveness actions; the combination of pressure factors and piste did not determine effectiveness. No decision trees were detected in relation to efficacy, analyzing pressure and piste.

Keywords: Observational designs, response behavior, fencing, decision tree

Funding: This study has been supported by Spanish government subproject Mixed method approach on performance analysis (in training and competition) in elite and academy sport [PGC2018-098742-B-C33] (Ministerio de Ciencia, Innovación y Universidades, Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema I+D+i), which is part of the coordinated project New approach of research in physical activity and sport from mixed methods perspective (NARPAS_MM) [SPGC201800X098742CV0].

E-mails: xiglesias@gencat.cat; rtarragog@gencat.cat; laura.ruiz@ucv.es

1. Introduction

Technical and tactical components are decisive for success in fencing. Many fencing masters structure their training lessons according to tactical thinking processes designed by the Hungarian master Szabó (1977). In elite fencing it is important to analyze tactical actions that affect sporting performance.

Decision trees have not been used in the specific fencing research literature. It is a methodology used for decision analysis, to help identify a strategy that is more likely to achieve a goal. For this reason, we have decided to use it. We only found one study that analyzed the system supporting decision-making in fencing training, based on a Bayesian network (Liu & Cui 2009). Furthermore, they have a very similar structure to Szabo's tactical thinking scheme (Szabó, 1977). On previous occasions, we have used this tree format by analyzing actions and their effectiveness, but without applying the specific decision tree methodology, as we have done in this work.

The aim was to determine the effectiveness of elite foil fencers based on the application of the decision-tree analysis for initiative attacks, pressure and the piste area.

2. Method

A nomothetic, punctual & multi-dimensional observational methodology design was used (Anguera, Blanco-Villaseñor, Hernández-Mendo, & Losada, 2011). The study was nomothetic because we observed different fencers in different assaults; punctual because we studied the set of assaults of the competition as a unit; and multidimensional because we studied different dimensions of behavior by the observation instrument.

2.1. Participants

13 male foil (MF) and 12 female foil (FF) combats were recorded during the celebration of the final stages of the 2014 World Fencing Championships.

2.2. Instruments

ESGRIMOBS (Tarragó & Iglesias, 2016) and Lince (Soto, Camerino, Iglesias, Anguera, & Castañer, 2019) were utilized as observational and recording instruments.

The fencer who made the first attack was "A" and their rival was "B". "A" pressure (Pres_A), "B" pressure (Pres_B) or no pressure (N_Pres) was analyzed. The piste zones were: End_A, End_B and centre. The effectiveness was determined: A_Touch, B_Touch or no touch.

2.3. A decision-tree model

We applied a decision-tree model. This works as a statistical resource that utilizes a tree-type decision-making model. It is a method of showing an algorithm that has conditional control statements.

A decision tree is essentially a flowchart structure. The internal nodes represent the features, a decision rule is shown as a branch, and each outcome is represented by a leaf node. The root node is on top. Having a flowchart structure helps decision-making. It is a visual way of displaying the human thinking process. Decision trees are useful because they are easy to comprehend visually and are popular because they are non-parametric or distribution-free methods, which don't need assumptions of distribution or profitability (Navlani, 2018).

A chi-square was used to check the differences in distribution.

2.4. Quality of the data

Validity

ESGRIMOBS was used as an observation instrument (Tarragó et al., 2016).

Intra-observer reliability

An observer analyzed 3 assaults twice, in a total of 45 records based on 29 observation criteria.

Values were obtained criterion by criterion by Fleiss' Kappa (1971) between 0.49 and 1. The lowest values corresponded to criteria with very few records (n = 4). The average value of all the Kappa coefficients of the 29 criteria was 0.797. The general level of agreement in all records was determined through the Iota coefficient (Janson & Olsson, 2001) in the set of agreements of both observers, obtaining a value of 0.806.

Inter-observer reliability

Two different observers analyzed 3 assaults twice, in a total of 45 records based on 29 observation criteria.

Values were obtained criterion by criterion by Fleiss' Kappa (1971) between 0.48 and 1. The lowest values corresponded to criteria with very few records (n = 4). The average value of all the Kappa coefficients of the 29 criteria was 0.714. The general level of agreement in all records was determined through the Iota coefficient (Janson & Olsson, 2001) in the set of agreements of both observers, obtaining a value of 0.794.

The quality of the data was calculated using the RStudio v. 1.2.5033 (© 2009-2019 RStudio, Inc.).

3. Results

1509 actions were analyzed. 67.1% were Pres_A, 13.5% Pres_B, while 19.4% N_Pres. "A" won 25.6% and "B" won 14.6%. There was no relationship between pressure, piste and effectiveness (n.s.) in total records. In FF (n=677), Pres_A (68.7%) got 23.9% A_Touch, 16.8% B_Touch and 59.4% no touch. Pres_B (11.7%) got 30.4% A_Touch and 16.5% B_Touch. N_Pres got 27.1% A_Touch, 12.8% B_Touch (n.s.). In MF, Pres_A (65.7%) got 23.4% A_Touch and 14.4% B_Touch. Pres_B (15.0%) got 31.2% A_Touch and 7.2% B_Touch. N_Pres got 30.0% A_Touch and 15.6% B_Touch (p=.049).

Figure 1 presents the distribution of the different actions registered in the assaults and their effectiveness depending on the pressure and area of the piste where they took place.

In figure 2 we can see the decision trees found in the analysis of the male foil actions, in figure 3 those of female foil actions, whilst figure 4 represents the set of relationships for all the subjects (both male and female sets).

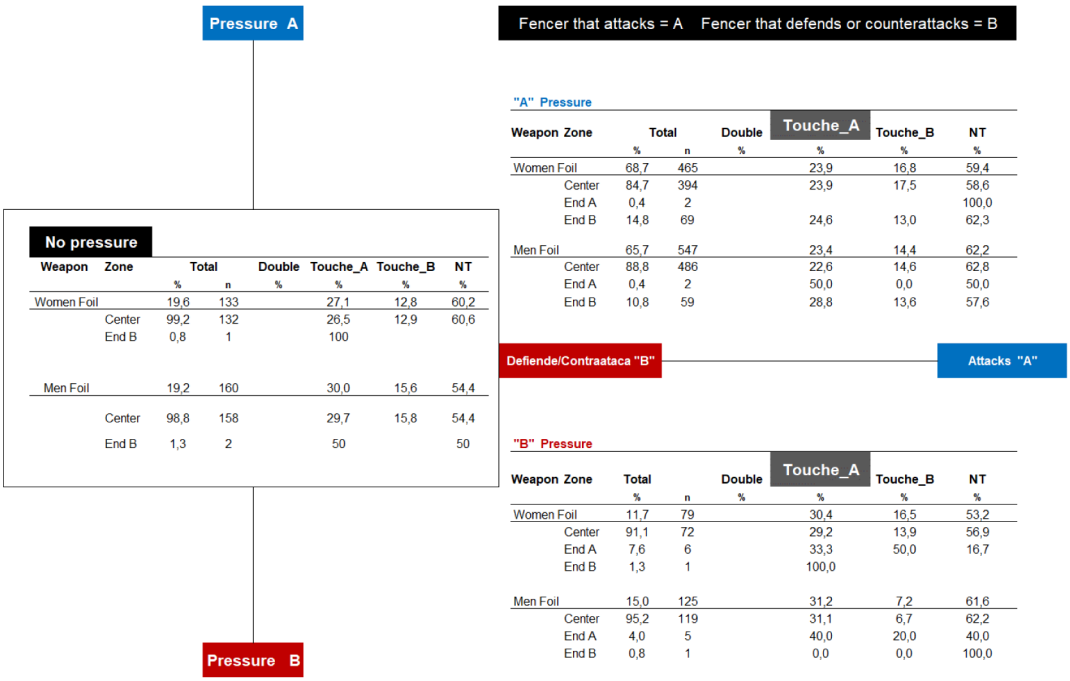


Figure 1. Actions' frequency distribution and their effectiveness

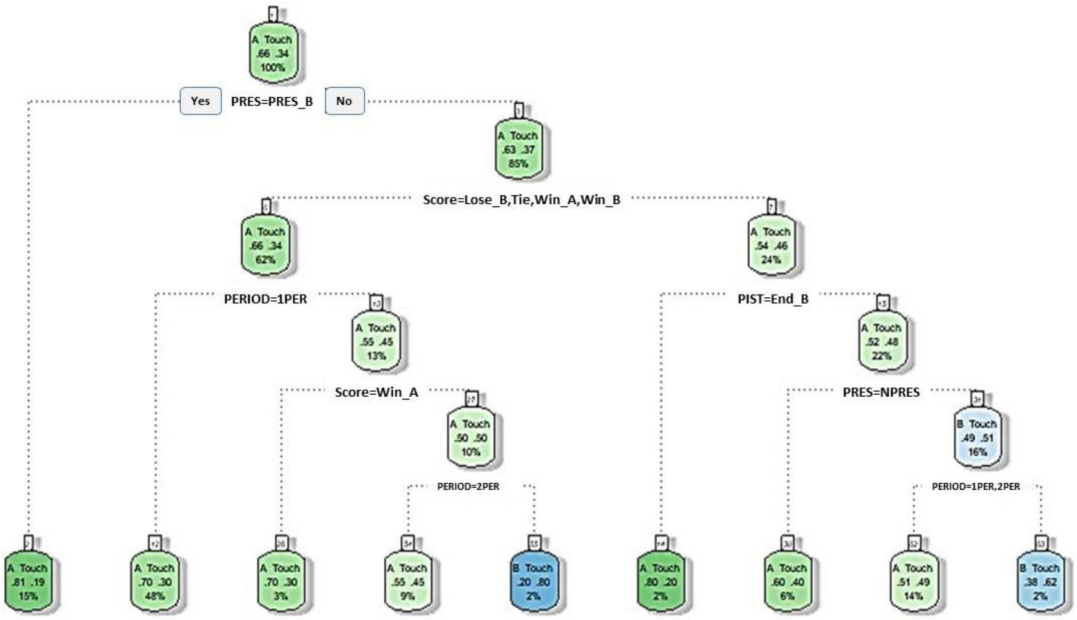


Figure 2. Decision-tree model: Men foil.

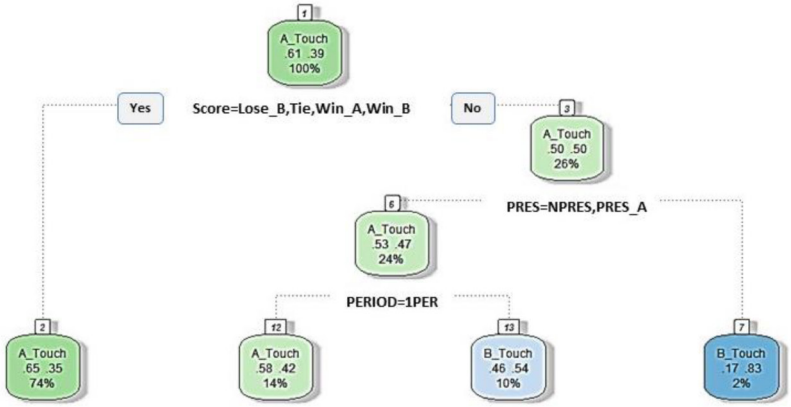


Figure 3. Decision-tree model: Women foil.

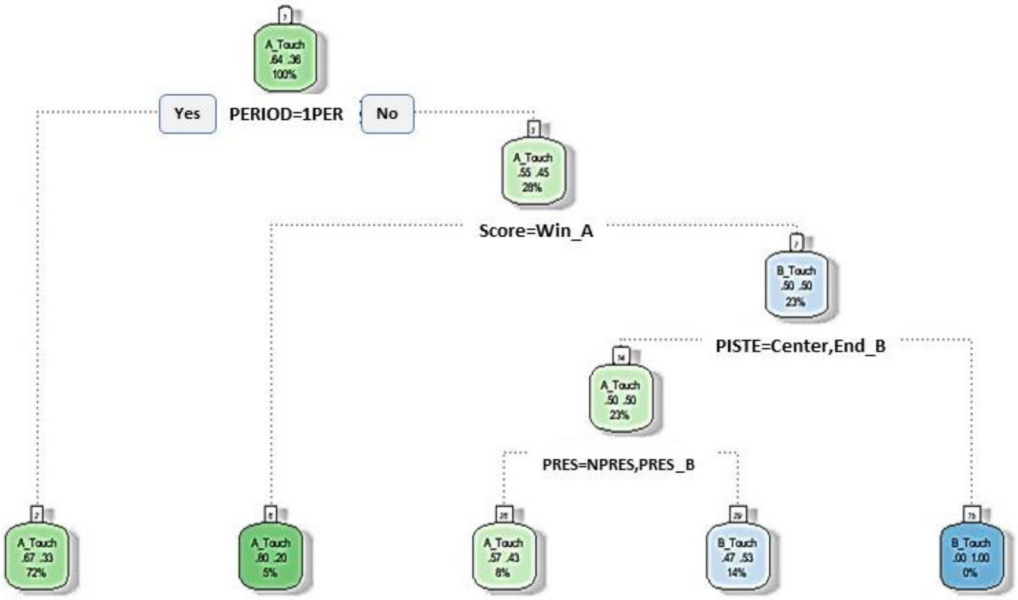


Figure 4. Decision-tree model: Foil (men and women).

4. Conclusions.

Combat conventions could determine different effectiveness actions; the combination of pressure factors and piste did not determine effectiveness. No decision tree was detected in relation to efficacy, analyzing pressure and piste. Decision trees appeared when we considered only the touch A or B, the assault period and the marker.

References

Anguera, M.T., Blanco-Villaseñor, A., Hernández-Mendo, A. & Losada, J.L. (2011). Diseños observacionales: ajuste y aplicación en psicología del deporte. *Cuadernos de psicología del deporte, 11*(2), 63–76.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61*(2), 277–289.

Liu, F., & Cui, X. (2009, December). Fencing Training Decision Support System Based on Bayesian Network. In 2009 *International Conference on Computational Intelligence and Software Engineering* (pp. 1–4). IEEE.

Navlani, A. (2018, 28 diciembre). *Decision Tree Classification in Python. DataCamp Community*. https://www.datacamp.com/community/tutorials/decision-tree-classification-python.

Soto, A., Camerino, O., Iglesias, X., Anguera, M. T., & Castañer, M. (2019). LINCE PLUS: Research software for behavior video analysis. *Apunts. Educación fisica y deportes, 3*(137), 149–153.

Szabó, L. (1977). *Fencing and the master*. Budapest: Corvina Kiado.

Tarragó, R., & Iglesias, X. (2016). Eficacia de las acciones técnicas y tácticas de la espada masculina de élite según su distribución espacial y temporal. *Apunts. Educación Física y Deportes, 125*, 79–89.

Observational analysis of judo combats: from highly structured records to the selection of T-patterns by specific dimensions

David Soriano[1], Rafael Tarragó[1], Xavier Iglesias[1], Daniel Lapresa[2], M. Teresa Anguera[3]

[1]*INEFC, Universitat de Barcelona, Barcelona, Spain,*

[2]*Department of Educational Sciences, University of La Rioja, Logroño, Spain,*

[3]*Faculty of Psychology, University of Barcelona, Barcelona, Spain*

Abstract

Purpose: We present an observational tool that records, analyzes and interprets technical-tactical performance in judo matches. *Method*: Five bouts were recorded and analyzed -two semifinals, two bouts for the bronze medal and the final- of three female weight categories (-48kg, -63kg and +78kg) and three male categories (-60kg, -81kg and +100kg). *Results*: The consistent structure of the observational tool designed implies that the records made were a faithful record of the behavior performed by the judokas in combat, and endow it with a high interest for its use not only by scientists, but also by coaches and competitors of this sport. However, the consequent variability of each event (row of the record or multievent in GSEQ terminology) that entails the complexity of the observation instrument indicated the convenience of reducing the number of dimensions to be incorporated in the detection of T-patterns with the THEME software so that regular behavior structures can be obtained. *Conclusions*: We present a specific example of a targeted process of selecting T-patterns through the tool provided by THEME to incorporate dimensions into the search process which, in addition, was enriched with the subsequent application of qualitative and quantitative filters.

Keywords: Observational Methodology; judo; T-patterns.

Funding: The authors gratefully acknowledge the support of a Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed*

methods perspective (NARPAS_MM) [SPGC201800X098742CV0]. In addition, the authors thank the support of the Generalitat de Catalunya Research Group, *GRUP DE RECERCA I INNOVACIÓ EN DISSENYS (GRID). Tecnología i aplicació multimedia i digital als dissenys observacionals* [Grant number 2017 SGR 1405].

E-mails: sorianodavid22@hotmail.comrtarragog@gencat.cat; xiglesias@gencat.cat; daniel.lapresa@unirioja.es; tanguera@ub.edu

1. Introduction

Judo is a combat sport about which there is abundant scientific literature, which collects information as disparate as that referring to the quantification of the effectiveness of judo techniques, the energy demands of judo fighting, injuries most common among its practitioners and the temporal structure of fighting. All the information provided by the different investigations is of great value to specialist judo technicians, who will have a better chance of success in their work if, in addition to their own experience, they base their decisions as a coach on the scientific evidence provided by these studies. At a technical and tactical level, it is not enough to analyze the judoka's behavior in isolation, the interaction that takes place between both opponents must be studied, since one's behavior is conditioned by the actions of their rival. For this reason, the objective of this research is to develop an observation system *ad hoc* so we can interpret the exchange of actions by athletes during judo matches, in order to subsequently design training sessions with solid scientific support.

The present work aims to demonstrate the operation of the observational tool designed, based on a) the information contained in the records -data packages- and b) of the regular behavior structures (T-patterns) detected. In the analysis of the behavior records obtained by means of observation instruments constructed *ad hoc*, THEME generates an output of T-patterns, based on more or less restrictive search parameters. Those that are most relevant to the study objectives need to be selected. The profusion of T-patterns is usually a fact when handling large samplings even when very restrictive search parameters are included (see Lapresa, Arana, Anguera, and Garzón, 2013); but the opposite can also be the case, mainly due to short samplings or highly structured records, which makes it difficult to detect regular behavior structures. In this work, we present a specific example of a targeted process of selecting T-patterns in a highly structured register, through the tool provided by THEME to incorporate dimensions to the search process which, in addition, is enriched with the subsequent application of qualitative and quantitative filters.

2. Method

2.1. Observational design

According to Anguera et al. (2011), this study was punctual (records of a single competition but without an individualized follow-up of the judokas that competed in it), nomothetic (36 judokas) and multidimensional (taking into account different co-occurrence behaviors in the same registry which corresponded to the different criteria that made up the observational instrument). The observation was active and non-participant (Anguera, 1990).

2.2. Participants

The sample of this investigation was made up of the judokas (n = 36) who participated in the combats (n = 30) corresponding to the semifinals (n = 12), finals (n = 6) and dispute for third to fifth place (n = 12) in the six Olympic categories at the 2016 Rio Olympic Games (less than 48 kg, less than 63 kg and more than 78 kg in the women's category, and less than 60 kg, less than 81 kg and more than 100 kg in male category). The project in which it was included was approved (0099S/2912/2010 2607/LA) by the clinical research ethics committee of the Catalonian sports administration (2005).

2.2. Observation and recording instruments

To obtain all the information regarding the techniques and tactics used by judokas during the judo matches to be analyzed, an observation instrument was designed *ad hoc*, which was called

JUDOBS. This is a combination of field format and category systems, made up of 52 criteria and 555 categories. There is a wide range of analysis options provided by the designed observation instrument and its adaptations used in this study. Table 1 presents the adaptation that was made in JUDOBS to analyze the behavior patterns of the Olympic champions, but an adaptation was also made to determine the existing T-patterns in the exchange of actions between judokas, not focusing on a specific judoka but on the one who performed the first action in each of the recorded events. In this way, it was possible to analyze whether any pattern of behavior exists when a specific judo technique is performed in certain circumstances, regardless of the judoka who performs it. The data was recorded and encoded using the software Lince, version 1.2.1 (Gabín et al., 2012).

2.3. Data quality

The panel of experts involved in the validation of JUDOBS was made up of a total of 28 judo coaches. The statistic used to validate this instrument was the calculation of the percentage of positive matches, which was 0.809 as there were 8258 matches out of a total of 10206 possible matches/discrepancies. To obtain the confidence intervals, we assumed the binomial model taking into account 8258 successes out of 10206 possible options, applying the binom.test function of R to obtain 95% confidence intervals starting from the number of successes (coincidences) in the total number of trials (possible matches/discrepancies). The resulting values returned a 95% confidence interval (CI), between 0.801 and 0.817, for the proportion of positive matches of a 0.809 probability.

To assess intraobserver reliability, one observer analyzed 5 combats twice, with values obtained criterion by criterion by Fleiss's Kappa (1971) of between 0.911 to 1 (with an average of 0.988 in all Kappa values). The Iota coefficient (Janson & Olsson, 2001) of the set of agreements of both observers in all criteria was 0.977. Interobserver reliability was determined by calculating the agreement between the records of three observers who analyzed these same combats, with values obtained criterion by criterion by Alpha Krippendorff (2018) of between 0.664 to 1 (with an average of 0.941 in all Alpha values). The Iota coefficient (Janson & Olsson, 2001) of the set of agreements of both observers in all criteria was 0.932.

Table 1. Adaptation of JUDOBS that contains the criteria and categories that were activated for the detection of regular behavior structures of the six champions of the 2016 Rio Olympic Games.

Criteria	Codes and categories
Scoreboard situation	EMPATE: matched score; GANANDOC: champion ahead on the scoreboard; GANANDOO: opponent is ahead on the scoreboard
Duration	HAJ: hajime; MAT: matte
Combat time	TI: first minute; TII: second minute; TIII: third minute; TIV: fourth minute; TV: fifth minute; GS: golden score
Stand or floor action	TW: tachi w aza; NW: ne w aza
Tatami placement of w ho initiates the 1st action	CCT / OCT: champion / opponent on the center of the tatami; CLTE / OLTE: champion / opponent on the line and not facing it; CLTC / OLTC: champion / opponent on the line and facing it
Scoring	NC / NO: nothing for the champion / opponent; YC / YO: yuko for the champion / opponent; WC / WO: w aza ari for the champion / opponent; IC / IO: ippon for champion / opponent
Kumi kata of the champion previous to the 1st action	KCIMDS / KCIMIS: one hand_right / left on flap; KCIMDM / KCIMIM: one hand left / right on sleeve; KCSK: no kumi kata; KCIIMC: classic_sleeve and flap; KCIIMA: tw o-handed_high; KCIIM: on tw o sleeves; KCIIS: on tw o flaps; KCCS: flap cross; KCCA: high cross; KCCMDU: crossed uke right sleeve; KCCMIU: crossed uke left sleeve; KCCIIM: one hand w aist and the other on sleeve; KCCIM: one hand w aist; KCAIM: high on one hand
Kumi kata of the opponent previous to the 1st action	Same as above but w ith the opponent's kumi kata (KOIMDS; KOIMIS; KOIMDM; KOIMIM; KOIIMC; KOIIMA; KOIIM; KOIIS; KOCS; KOCA; KOCMDU; KOCMIU; KOSK; KOCIIM; KOCIM; KOAIM)
Action 1	IPSNC / IPSNO: ippon seoi nage from the champion / oponnent; ISNC / ISNO: seoi nage from the champion / oponnent; ISOC / ISOO: seoi otoshi from the champion / oponnent; ISNRC / ISNRO: seoi nage reverse from the champion / oponnent; IKGC / IKGO: kata guruma from the champion / oponnent; ITOC / ITOO: tai otoshi from the champion / oponnent; IUMSC / IUMSO: uchi mata sukashi from the champion / oponnent; IOGC / IOGO: o goshi from the champion / oponnent; IKOGC / IKOGO: koshi guruma from the champion / oponnent; IHRGC / IHRGO: harai goshi from the champion / oponnent; IUGC / IUGO: ushiro goshi from the champion / oponnent; ISTGC / ISTGO: sode tsurikomi goshi from the champion / oponnent; IDABC / IDABO: de ashi barai from the champion / oponnent; IHIGC / IHIGO: hiza guruma from the champion / oponnent; IOSGC / IOSGO: o soto gari from the champion / oponnent; IOUGC / IOUGO: o uchi gari from the champion / oponnent; IUMC / IUMO: uchi mata from the champion / oponnent; ISTAC / ISTAO: sasae tsurikomi ashi from the champion / oponnent; IKUGC / IKUGO: ko uchi gari from the champion / oponnent; IOSGAC / IOSGAO: o soto gaeshi from the champion / oponnent; IOUGAEC / IOUGAEO: o uchi gaeshi from the champion / oponnent; ITNC / ITNO: tomo nage from the champion / oponnent; ISGC / ISGO: sumi gaeshi from the champion / oponnent; IUNC / IUNO: ura nage from the champion / oponnent; ISMC / ISMO: soto makikomi from the champion / oponnent; IYGC / IYGO: yoko guruma from the champion / oponnent; IKUMC / IKUMO: ko uchi makikomi from the champion / oponnent; IKEGC / IKEGO: kesa gatame from the champion / oponnent; IKSGC / IKSGO: kami shiho gatame from the champion / oponnent; IYSGC / IYSGO: yoko shiho gatame from the champion / oponnent; ITSGC / ITSGO: tate shiho gatame from the champion /

Criteria	Codes and categories
	oponnent; INJJC / INJJO: nami juji jime from the champion / oponnent; IHJC / IHJO: hadaka jime from the champion / oponnent; IOEJC / IOEJO: okuri eri jime from the champion / oponnent; IKJC / IKHO: kataha jime from the champion / oponnent; IKTJC / IKTJO: kata te jime from the champion / oponnent; ISJC / ISJO: sankaku jime from the champion / oponnent; IGJJC / IGJJO: gyaku juji jime from the champion / oponnent; IKJGC / IKJGO: kata juji gatame from the champion / oponnent; IUGRC / IUGRO: ude garami from the champion / oponnent; IJGC / IJGO: juji gatame from the champion / oponnent; IUGAC / IUGAO: ude gatame from the champion / oponnent; IHGC / IHGO: hiza
Actions 2 to 4	The categories of action 1 are repeated but for actions 2 to 4 (replacing the initial "I" of each criterion "II", "III" or "IV")
Winner of the combat	GC / GO: Combat won by champion / opponent

2.4. Data analysis with Theme

The detection of regular behavior structures (*T-patterns*) was performed using the *software* THEME (version 6.Edu) (Magnusson, 1996). The selected search parameters, which guaranteed that the detected T-pattern was not a product of chance, were the following: a) a frequency of occurrence equal to or greater than 2 was set; b) a significance level of 0.005 was used, which means that the percentage of accepting a critical interval due to chance is 0.5%; c) Redundancy reduction was set so that the *T-pattern* was not incorporated into the output of Theme when more than 90% of the occurrences of a new *T-pattern* start and end coincided with the critical interval relationships of patterns already detected; d) *fast requirement* was deactivated at all levels, selecting the critical interval mode *Free*. Taking into account the work of Lapresa, Arana et al., (2013), the detection of T-patterns was carried out under the order parameter, assigning a constant duration to each unit of behavior -row of the registry-, which deduced whether the behaviors reflected in the T-pattern were consecutive or if there were intercalated behaviors between the detected multi-events.

3. Results

As an example of the operation of the observation system built, Figure 1 presents: a) the data package corresponding to the record of the combat of the semifinals in the under 48 kg female category, disputed by the judokas Pareto (Argentina) and Kondo (Japan) and b) one of the T-patterns, including the performance of technical-tactical actions, reflected regular behavior patterns of the judoka Pareto and only from a data package corresponding to a single combat.

Specifically, the *T-pattern* ((ganandoc, haj, tw, gc (ganandoc, tw, nc, kciimc, koiimc, iai, iougo, gc (ganandoc, haj, tw, gc ganandoc, tw, nc, kcimis, kosk , idd, ipsnc, gc))) ganandoc, haj, tw, gc) was made up of five events -Theme terminology; multi-events in GSEQ terminology- which were repeated twice. The first occurrence was made up of rows 5-7-20-21-23 from the registry; and the second occurrence of rows was 23-24-27-28-30. The information contained in the T-pattern shows us that during these sequences of actions the Olympic Champion (Pareto) was always winning on the scoreboard. The following behavior pattern was repeated significantly during combat in the interaction between both judokas: classic sleeve grip and flap of both judokas, Kondo made attacks using O Uchi Gari on the right, without obtaining a result, while Pareto made attacks using Ippon Seoi Nage on the right, also without scoring. A reading and interpretation of this behavior patterns reveals that that the Japanese judoka tries to make attacks backwards, probably due to the defensive posture of her winning rival. And the Argentine judoka takes advantage of this circumstance to make attacks forwards thanks to Kondo's risky attitude, in an attempt to get ahead on the scoreboard. The usefulness of these results is high, as it shows the offensive profile of the Japanese judoka at times when she must take risks to get a result. At the same time, we know what Pareto does to contain that situation without being sanctioned for passivity and how the attitude and direction of rival attacks are used.

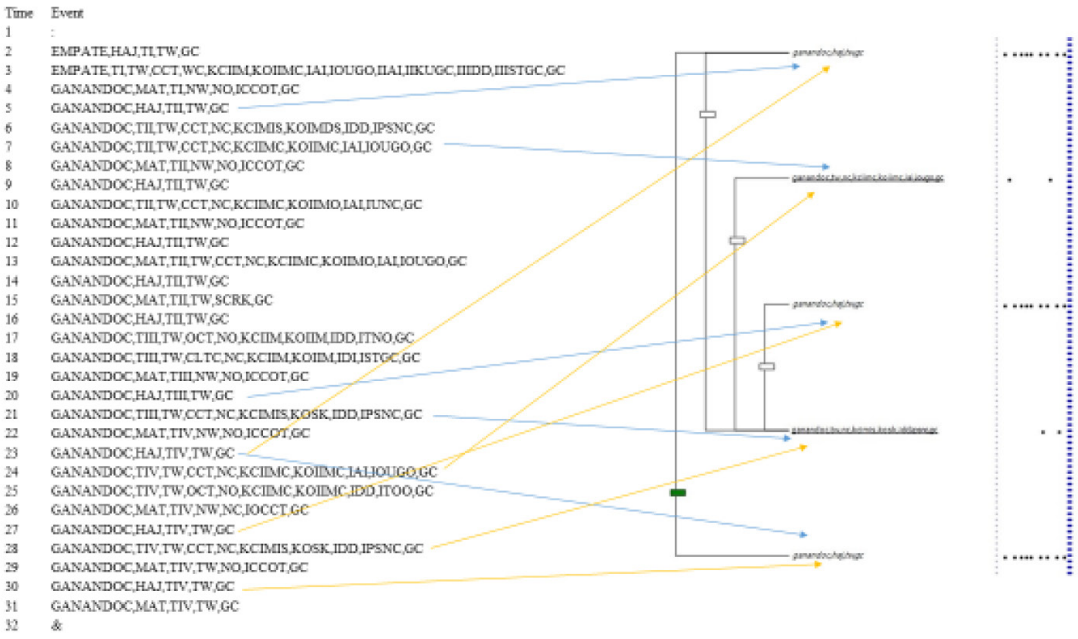


Figure 1. Data pack corresponding to the Pareto vs Kondo semifinal match of the female category of under 48 kg in the 2016 Rio Olympic Games and the dendogram of the selected T-pattern to demonstrate the operativity of the observation instrument. The arrows highlight the events that made up the T-pattern in each of the two occurrences.

3.1. T-patterns

At this point, and in the same way that the researchers who use THEME for their corresponding works, we had to select the T-patterns to present out of all of the T-patterns detected in all thirty combats (five per category studied: 48F, 1-5; 63F, 6-10; 78F, 11-15; 60M, 16-20; 81M, 21-25; 100M, 26-30).

Of the T-patterns detected in the analysis of the 30 combats, specifically, in the exchange of actions between judokas -when the focus was on the one who took the initiative and performed the first action, called "p" (and their rival "q")- including in the search all the variable criteria of the observation instrument, the detected T-pattern with the greatest scope was presented: ((pdqz, tie, haj, ti, tw, gp pdqz, tie, haj, ti, tw, gp)((pdqz, tie, haj, tii, tw, gp pdqz, tie, haj, tii, tw, gp) ((pdqz, tie, haj, tiii, tw, gp pdqz, tie, haj, tiii, tw, gp) (pdqz, tie, haj, tiv, tw, gp pdqz, tie, haj, tiv, tw, gp)))), with occurrence in combats 18 (category 60M) and 21 (category 81M), and with an average of the internal intervals between its multi-events constitutive of: 3,50-3,50-2-2,50-2,50-2,50-2. The T-pattern that occurs in a greater number of combats such as (pdqz, tie, haj, ti, tw, gq pzqd, tie, haj, tii, tw, gp), with occurrences in combats 5-9-20-23-25-29; and with an average of its internal intervals of 6-5-7-2-3-5 was also interesting.

Now, if we look at the information that these *T-patterns* contain, we will see that in all the multi-events the row of the record corresponding to the action of the combat by the referee is reflected. So, as may be usual, in this case the most frequent multi-events are those that are reflected in the *T-patterns*. But, on the other hand, the consequent variability of each event (row of the record or multievent in GSEQ terminology) that entails the complexity of the observation instrument indicates the convenience of reducing the number of dimensions to be incorporated

in the detection of *T-patterns* with the THEME software so that regular behavior structures can be obtained.

At this point and from the activation tool dimension presented by THEME, we proceeded to detect regular behavior structures activating the punctuation, kumikata prior to the 1st action of both judokas, the type of techniques performed within the exchange of actions, and the duration dimensions. As the categories that made up this last dimension, hajime -activation of the combat by the referee-, and mate -interruption of the combat by the referee- were inherent to the structure of the registry, we decided to select for presentation the T-patterns that provided more information than that contained in said rows of the record.

Furthermore, we may be interested in: a) T-patterns that are detected in the same category (weight / masc-fem), such as the T-pattern (((mat, nq haj)(mat, nq haj))((mat, np haj) np, kpimis, kqsk, ipsnp)), which occurred in combats 2 and 5, category 48F, with the average of the internal intervals 1-2-1-8-1-1; b) orthe T-patterns that are detected in various categories, such as the T-pattern ((haj (mat ha j)) mat, np, kpiim, kqiim, istgp), which takes place in fights 1, 5 and 26 (40F and 100M), with a mean of internal intervals=6; c) or, perhaps, regular structures of conduct that are repeated intra-combat, such as the T-pattern (mat, np, kpcs, kqsk, iosgp mat, np, kpcs, kqsk, iosgp), which takes place in combat 22 (category 81M) with an average of internal intervals =4 (that is, with three rows of the register interspersed).

But it could also be interesting to study the technical-tactical performance of the competitors without considering their effectiveness; for this, the Point criterion could be deactivated from the search; and respecting the three decisions made in the previous paragraph, we find: a) T-patterns that are detected in the same category (weight / male-fem), such as the T-pattern (mat, kpiima, kqiimc, iougp (kpiim, kqiim, istgp (mat haj))) that occurred in combats 1 and 2, category 48F, with the average of the internal intervals 6-1-1; b) T-patterns that are detected in various categories, such as the T-pattern (kpiim, kqiimc, idabp kpiim, kqiimc, istgp), which took place in combats 4-17 (40F and 60M), with an average of internal intervals of 2-1-1.67. c) T-patterns that occurred intra-combat (kpiimc, kqiimc, idabp, iiump mat, kpiimc, kqiimc, idabp, iiump), which occurred twice in combat 8 (category 63F) with an average of internal intervals =3 (that is, with two rows of the record interspersed).

4. Conclusions

The operativity of the observation system was supported by the data package presented as an example of the registry (one of the female semifinals in the under 48 kg category), and the informative power of the *T-patterns* showed regular structures in the behavior displayed in a single match by the judoka who ended up becoming the Olympic champion in the category of under 48 kg. The record that configured each data package allowed us to represent what happens in the course of a judo match, based on the structure that underpins the observation instrument. This representation facilitates the understanding of the behavior developed by judokas in combat. In addition, the detection of regular behavior structures (T-patterns) has been used through the software Theme (version 6.Edu) (Magnusson, 1996) within the data package corresponding to the semifinal match of less than 48 kg between Pareto and Kondo. The presented *T-pattern* provides detailed information (synchronous and diachronic) on regular guidelines in the technical-tactical performance of the Olympic champion and her interaction with the rival, and in the course of a single combat. The consequent variability of each event that entails the complexity of the observation instrument indicates the convenience of reducing the number of dimensions to be incorporated in the detection of *T-patterns* with the THEME software so

that regular behavior structures can be obtained. In this work, a specific example of a targeted process of selecting *T-patterns* through the tool provided by THEME has been presented to incorporate dimensions to the search process which, in addition, has been enriched with the subsequent application of qualitative and quantitative filters.

References

Anguera, M.T. (1990). observacional. En J.Arnau, M.T. Anguera y J. Gómez, (Eds.), *Metodología de la investigación en ciencias del comportamiento* (pp. 125–238). Murcia, España: Universidad de Murcia.

Anguera, M.T., Blanco-Villaseñor, A., Hernández-Mendo, A., & Losada, J.L. (2011). Diseños observacionales: ajuste y aplicacion en psicología del deporte. *Cuadernos de Psicología del Deporte, 11*(2), 63–76.

Bakeman, R. (1978). Untangling streams of behavior: sequential analysis of observation data. En G.P. Sackett (Ed.) *Observing Behaviour, Vol. II: Data Collection and Analysis Methods* (pp. 63–78). Baltimore: University Park Press.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382.

Gabin, B., Camerino, O., Anguera, M.T., y Castañer, M. (2012). Lince: multiplatform sport analysis software. *Procedia-Social and Behavioral Sciences*, 46, 4692–4694. https://doi.org/10.1016/j.sbspro.2012.06.320

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*(2), 277–289.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Beverly Hills, Sage publications.

Lapresa, D., Arana, J., Anguera, M.T. y Garzón, B. (2013). Comparative analysis of the sequentiality using SDIS-GSEQ and THEME: a concrete example in soccer. *Journal of Sports Sciences*, 31(15), 1687–1695. https://doi.org/10.1080/02640414.2013.796061

Magnusson, M.S. (1996). Hidden real-time patterns in intra- and inter-individual behavior. *European Journal of Psychological Assessment*, 12(2), 112-123. https://doi.org/10.1027/1015-5759.12.2.112

Effect of a goalkeeper's distribution on the outcome of play in Women's Elite Football. Iberdrola League

José L. Losada[1], Claudio A. Casal[2], Rubén Maneiro[3]

[1]*Department of Methodology of Behavioral Sciences, University of Barcelona, Spain,*

[2] *Department of Science of Physical Activity and Sport, Catholic University of Valencia "San Vte Mártir", Spain,*

[3]*Department of Science of Physical Activity and Sport, Pontifical University of Salamanca, Spain*

Abstract

In today's football, the goalkeeper's role is not limited to the defensive phase of the game, as the ultimate player responsible for preventing a goal. Currently, the goalkeeper must also assume an important function during the offensive phase, as its initiator or continuator. *Purpose***:** To analyze the goalkeeper's distribution in the offensive phase and, whether this distribution influences the offensive performance of the team. *Method***:** The sample consisted of the matches corresponding to the 2018/2019 season of the Iberdrola League. The performance indicators were distribution zone and type, distribution, number of passes, outcome, pitch zone of first pass to outfield, pitch zone by goalkeeper, pitch zone of outcome and defensive pressure. Univariate and multivariate analyses were performed (Chi-square test, $p<0.05$). *Results***:** There were significant differences between the analyzed indicators and the outcome ($p=0.000$). Specifically, in most of the goals scored in an offensive attack in which the goalkeeper participated, the type of distribution was indirect, with possession of more than 6 passes, the goalkeeper sending a pass to the middle zone and without pressure from the opposing team.

Keywords: elite football; women; goalkeeper; offensive play; match analysis.

Funding: The authors gratefully acknowledge the support of a Spanish government subprojects *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] and *Mixed method approach on performance analysis (in training and competition) in elite and academy sport* [PGC2018-098742-B-C33] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach*

of research in physical activity and sport from mixed methods perspective (NARPAS_MM) [SPGC201800X098742CV0]. In addition, the first author thanks the support of the Generalitat de Catalunya Research Group, *GRUP DE RECERCA I INNOVACIÓ EN DISSENYS (GRID). Tecnología i aplicació multimedia i digital als dissenys observacionals* [Grant number 2017 SGR 1405].

E-mails: jlosada@ub.edu; ca.casal@ucv.es; rmaneirodi@upsa.es

1. Introduction

The goalkeeper is a specific position of vital importance in a football team's game, since they are the player closest to the own goal and an error by this player can mean a drastic change in the course of the match, which entails a high degree of responsibility. But in today's football, the goalkeeper has taken a more active role, motivated primarily by regulatory changes. His functions have gone from being limited only to the defensive phase of the game, as the ultimate person responsible for preventing a goal, to also taking on an important role during the offensive phase, as its initiator or continuator. Consequently, goalkeepers have adapted to the new demands of the game, involving greater technical-tactical skill and behaving like the players in the last defensive line (Lapresa Ajamil et al., 2018) and the first offensive line (Pérez et al., 2016), as indicated by the results obtained in the study of Sainz de Baranda et al. (2019), who verified that the technical action most carried out by a goalkeeper together with that of blocking, was playing with their feet, that is, controlling and passing, i.e., a predominantly offensive action.

Despite the importance of this figure, the scientific interest it has aroused so far is negligible, with very little research work available. Most of the previous works have excluded the goalkeeper from any team and attacking game analysis (Otte et al., 2019). Instead, they have focused on their ability to stop penalty shots (Gelade, 2014; Lopes et al., 2012; Noël et al., 2015).

Accordingly, we aim to fill the gap with this study, whose main objective is to analyze the goalkeeper's participation in the offensive phase and, examine whether this participation influences the offensive performance of the team. The results of the work will provide greater insights into the influence of goalkeeper participation in the offensive phase of the game.

2. Method

2.1. Sample and design

The sample consisted of the matches corresponding to the 2018/2019 season of the Iberdrola League.

Out of the possible observational methodology options, a nomothetic, intersessional monitoring, multidimensional design was applied (Anguera, 1979). Nomothetic because a plurality of units were studied, intersessional over time and multidimensional because we analyzed the multiple dimensions that constituted the *ad hoc* observation instrument used. The systematic observation carried out was non-participant and active, using observational "all occurrence" sampling.

2.2. Observational tool

Distribution zone: Inside the box (IB), Outside the box (OB). *Distribution type:* Goal kick (GK), Free kick (FK), Open play after transition (OR), Open play to continued possession (OP). *Distribution*: Direct (DR), Indirect (ID), No distribution (ND). *Number of passes. Outcome*: Goal (GO), Attempt on target (AO), Attempt off target (AF), Set Piece (SP), Loss goalkeeper (LG), Loss player (LP), Returned to goalkeeper (RG). *Pitch zone of first pass by outfield*: Defensive (DF), Middle defensive (MD), Middle Offensive (MO), Offensive (OF), Central (CE). *Pitch zone by goalkeeper*: Defensive (DFG), Middle Defensive (MDG), Middle Offensive (MOG), Offensive (OFG), Central (CEG), Middle offensive (MOG), No distribution (ND). *Pitch zone of outcome:* Defensive (DFF), Middle Defensive (MDF), Middle Offensive (MOF), Offensive (OFF), Central (CE), Middle offensive (MO). *Defensive pressure*: High (HG) Low (LW).

2.3. Procedure

The images of the matches were obtained from the InStatScout platform (www.instatscout.com) which is a private platform dedicated to assessing the performance of teams in different leagues around the world. They were analyzed post-event by the systematic observation of two observers, who were trained following the protocol of Losada & Manolov (2014). Firstly, eight observation sessions dedicated to the training of the observers were carried out applying the criteria of consensual agreement (Anguera, 1990) between observers, so that the sessions were only recorded when agreement was produced. The quality control of the data was also carried out by means of an inter-observer agreement analysis using Cohen's Kappa coefficient. The Kappa values (Table 1) were excellent, taking Fleiss, Levin, & Paik (2003) as a reference.

2.4. Statistical analysis

The R Studio program was used as the analysis instrument. According to the objective of the work, two types of statistical analysis were carried out. First, a univariate descriptive analysis was carried out through the calculation of primary measures such as the analysis of proportions. Subsequently, a bivariate analysis was carried out to check the association of the analyzed behaviors with the result of the offensive play, using the Chi-square test for this. The level of significance was set at $p<0.005$.

3. Results

3.1. Descriptive analysis

Table 1. Absolute and relative frequencies

Category	Frequencies	%	Category	Frequencies	%
OB	169	14.9	AF	12	1.1
IB	965	84.9	AO	16	1.4
FK	73	6.4	FT	1	0.1
GK	251	22.1	GO	9	0.8
OP	510	44.9	LG	296	26.1
GR	301	26.5	LP	602	53.0
DR	460	40.5	RG	99	8.7
ID	523	46.0	SP	98	8.6
ND	152	13.4	CE	23	2.0
0	312	27.5	DF	69	6.1
1-3	530	46.7	MD	408	35.9
4-6	173	15.2	MO	1	0.1
>6	110	10.5	CEG	208	18.3
HG	220	19.3	DFG	54	4.8
LW	915	80.5	MDG	544	47.9

Category	Frequencies	%	Category	Frequencies	%
CEF	406	35.7	MOG	19	12.9
DFF	105	9.2	ND	147	12.9
MDF	150	13.2	MOF	305	26.8
OFF	169	14.9			

3.2. Bivariate analysis

Table 2. Results of Chi-square test

	GO	AO	AF	SP	LG	LP	RG	Pvalue
DR	2	4	2	42	150	249	11	0.000
ID	7	12	10	51	2	351	87	
ND	0	0	0	5	144	2	1	
0	0	0	0	12	296	3	1	0.000
1-3	2	4	3	58	0	382	81	
4-6	2	6	7	19	0	125	14	
>6	5	6	2	9	0	92	3	
CE	0	0	0	2	2	16	3	0.276
DF	1	1	0	6	18	37	6	
MD	2	7	6	38	97	217	40	
MO	0	0	0	0	0	1	0	
OF	0	0	0	0	0	0	0	
CEG	0	3	1	20	3	170	3	0.000
DFG	0	0	0	4	0	31	19	
MDG	8	13	11	52	2	379	76	
MOG	1	0	0	2	1	15	0	
OFG	0	0	0	0	0	0	0	
ND	0	0	0	5	142	0	0	
CEF	0	0	0	31	190	183	0	0.000
DFF	0	0	0	3	4	4	94	
MDF	1	0	0	17	62	65	4	
MOF	0	1	4	39	39	222	0	
OFF	8	15	8	8	1	128	1	
HG	1	2	2	16	102	86	11	0.000
LW	8	14	10	82	194	516	88	

4. Conclusions

Goalkeepers mainly distributed the ball from inside the box (84.9%), with no defensive pressure from the opposing team (80.5%), after receiving a pass from a teammate (44.9%) made from the middle defensive zone (35.9%). The goalkeeper normally made a pass to a player located in the defensive zone or middle defensive zone (46.0) that gave continuity to the offensive phase, making mostly between 1-3 passes (46.7%). The sequence ended in the central zone (35.7%) through an outfield loss (53%). Most goals were scored through indirect distribution (7, p=0.000) with possession of more than 6 passes (5, p=0.000), with the goalkeeper sending a pass to a player located in the middle zone (8, p=0.000) without defensive pressure from the opposing team (8, p=0.000).

References

Anguera, M. T. (1990). Metodología observacional. En J. Arnau, M. T. Anguera, & J. Gómez (Eds.), *Metodología de la investigación en Ciencias del Comportamiento* (pp. 125–236). Secretariado de Publicaciones de la Universidad de Murcia.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3ª). John Wiley y Sons.

Gelade, G. (2014). Evaluating the ability of goalkeepers in English Premier League football. *Journal of Quantitative Analysis in Sports*, *10*(2). https://doi.org/10.1515/jqas-2014-0004

Lapresa Ajamil, D., Chivite Navascués, J., Arana Idiakez, J., Anguera, M. T., & Barbero Cadirat, J. R. (2018). Análisis de la eficacia del portero de fútbol cadete (14 a 16 años) [Analysis of the Effectiveness of Under-16 Football Goalkeepers]. *Apunts Educación Física y Deportes*, *131*, 60–79. https://doi.org/10.5672/apunts.2014-0983.es.(2018/1).131.05

Lopes, J. E., Araújo, D., Duarte, R., Davids, K., & Fernandes, O. (2012). Instructional constraints on movement and performance of players in the penalty kick. *International Journal of Performance Analysis in Sport*, *12*(2), 331–345. https://doi.org/10.1080/24748668.2012.11868602

Losada, J. L., & Manolov, R. (2014). The process of basic training, applied training, maintaining the performance of an observer. *Quality & Quantity*. https://doi.org/10.1007/s11135-014-9989-7

Noël, B., Furley, P., van der Kamp, J., Dicks, M., & Memmert, D. (2015). The development of a method for identifying penalty kick strategies in association football. *Journal of Sports Sciences*, *33*(1), 1–10. https://doi.org/10.1080/02640414.2014.926383

Otte, F. W., Millar, S.-K., & Klatt, S. (2019). How does the modern football goalkeeper train? – An exploration of expert goalkeeper coaches' skill training approaches. *Journal of Sports Sciences*, 1–9. https://doi.org/10.1080/02640414.2019.1643202

Pérez, S., Domínguez, R., Rodríguez, A., & López García, S. (2016). Estudio de las acciones técnicas del portero de fútbol profesional a lo largo de una temporada: Implicaciones para el entrenamiento. *Revista Digital de Educación Física*, *42*, 22–37.

Sainz de Baranda, P., Adán, L., García-Angulo, A., Gómez-López, M., Nikolic, B., & Ortega-Toro, E. (2019). Differences in the Offensive and Defensive Actions of the Goalkeepers at Women's FIFA World Cup 2011. *Frontiers in Psychology*, *10*, 223. https://doi.org/10.3389/fpsyg.2019.00223

Systematic review in futsal: Impact on methodological quality

María Preciado[1], M. Teresa Anguera[2], Mauricio Olarte[3], and Daniel Lapresa[4]

[1]*Doctoral Program Communication and Change, University of Barcelona, Spain,*

[2]*Faculty of Psychology, University of Barcelona, Spain,*

[3]*National Administrative Department of Statistics, Bogotá, Colombia,*

[4]*Department of Educational Sciences, University of La Rioja, Logroño, Spain*

Abstract

This work, which is part of a broader investigation, aims to carry out a systematic review of futsal, focusing especially on the compliance with the methodological requirements of the primary documents, and within a mixed methods framework. The primary documents followed the observational methodology, corresponded to the 2009-2019 decade, were published in English, Spanish or Portuguese, and were obtained from various databases. In accordance with PRISMA specifications, 37 people were selected out of the initial 2410 participants, which met all the established requirements. Two aspects were considered in the systematic review carried out: substantive and methodological. The substantive aspect is the classical one, while the methodological one emphasized the revision of the procedural aspects contained in the GREOM guides, published in the EQUATOR Network. As a result of this review, specific primary document profiles were proposed, and proportional comparison analysis was also proposed to delve further into the diversity of primary documents in terms of their adjustment to the procedural structure of observational methodology.

Keywords: Procedural profiles; methodological quality; systematic review; futsal; direct observation; mixed methods.

Funding: This study has been supported by the Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated

project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mails: preciado030@gmail.com; tanguera@ub.edu; olartemauricio@gmail.com; daniel.lapresa@unirioja.es

1. Introduction

This systematic review aims to obtain a comprehensive synthesis of evidence (Higgins & Green, 2011) in relation to a space of knowledge in which various publications have been generated. In this work, a systematic review of futsal was carried out, although from a perspective that we consider novel (Preciado et al., 2019), in that it clearly emphasizes the interest in the procedural quality of primary documents (Pluye et al., 2011).

2. Method

2.1. Selection of primary documents

A computer search was carried out in the *Web of Science, Scopus, ProQuest, Medline, Google Scholar, Scielo* and *Dialnet* databases, and the systematic review guidelines in *Preferred Reporting Items for Systematic reviews and Meta-analyses Guidelines* (PRISMA) were applied (Liberati et al., 2009). The dates of publication of the primary documents corresponded to the period between 2005 and 2019.

The search was conducted in English, Spanish and Portuguese, and the keywords were: futsal, observational methodology, defense, match analysis, performance analysis, game analysis, tactical analysis, goal and game patterns, and their respective translations. The inclusion criteria of the primary documents were articles on futsal in which observational methodology was used, which contained some of the indicated keywords, and fitted at least two of the domains adopted from GREOM (Portell, et al., 2015).

2.2. Procedure

The search was performed by two main investigators simultaneously. These compared the articles, analyzed differences between them and debated the dubious situations that arose in their selection. These were analyzed with a third author to define the screening and reduce bias errors.

3. Results

3.1. Substantive systematic review

The following criteria were taken into account in the substantive systematic review: language, country of origin, journal in which the primary documents and impact factor were published, individual/shared authorship, samples, and object of study.

3.2. Systematic methodological review

For the systematic methodological review, each of the primary documents was analyzed from the three major domains established in the GREOM guide (Portell et al., 2015) which, respectively, consist of the type of observation (direct vs. indirect), the method (broken down into observational design, observation and recording instruments, parameters, and data quality control), and data analysis. Table 1 presents this analysis.

This procedural review revealed that there are different primary document profiles, depending on their methodological shortcomings compared to each of the indicated domains.

Table 1. Systematic review of primary documents from GREOM.

AUTHORS	Domain A	Domain B: Method				Domain C:	
		Observ.	Instrument		Parameters	Quality of data	Data analysis
	Kind of observation	design	Observation	Record			
Aires (2012)	There is not	No	Catalogue of behaviors (Scout)	No	Frequency	No	Descriptive analysis, EXCEL
Álvarez, García, et al. (2018)	Direct ob.	N/S/M	FC/SC	LINCE	Frequency	Agreement percent	Chi-square and Fisher test, SPSS
Álvarez, Medina, Murillo, et al. (2018)	Direct ob.	No	SC	LINCE	Frequency	Agreement percent	Descriptive analysis, t-Student, Kolmogorov and Smirnov, SPSS
Amaral y Garganta (2005)	Direct ob.	No	SC	Video, SDIS-GSEQ	Order	No	Lag sequential analysis and polar coordinate análisis, SDIS-GSEQ
Amatria et al. (2016)	Direct ob.	N/M/S	FC/SC	LINCE	Order	Kappa	Simple logistic regression, SPSS
Arruda, et al. (2011)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, MATLAB
Balyan y Vural (2018)	Direct ob.	No	Catalogue of behaviors	No	Frequency	No	ANOVA and Kruskal-Wallis, SPSS
Bortolini y Soares (2018)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, EXCEL
Botelho y Coppi (2010)	Direct ob.	No	Catalogue of behaviors	DVD	Frequency	No	Descriptive analysis, EXCEL
Bueno y Alves (2012)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, EXCEL
Campos (2013)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	ANOVA, Kruskal-Wallis, Shapiro Wilk test, Bonferroni test, STATISTICA and TACTO
Corrêa, Davids, et al. (2014)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	Kappa Correlation coefficient	Descriptive analysis, Mann-Whitney, t-test, TACTO and SPSS
Corrêa, Vilar, et al. (2014)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, EXCEL
De Paula y Barbosa (2019)	Direct ob.	No	Catalogue of behaviors	No	Frequency	No	Descriptive analysis, EXCEL
Franco, et al. (2014)	Direct ob.	No	Catalogue of behaviors	No	Frequency	Correlation coefficient	Descriptive analysis, Chi-square, ATRO and SPSS
Fukuda y de Santana (2012)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, EXCEL
García-Angulo y García-Angulo (2018)	Direct ob.	I/S/M	FC/SC	Video	Frequency	Kappa	Chi-square, Cramer coefficient and Phi coefficient, SPSS
Giani, et al. (2018)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, Cronbach Alpha, EXCEL
Giusti, et al. (2012)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis

Giani, et al. (2018)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis, Cronbach Alpha, EXCEL
Giusti, et al. (2012)	Direct ob.	No	Catalogue of behaviors	Video	Frequency	No	Descriptive analysis

The resulting profiles were as follows:

A: Direct observation, observational design, observation instrument, video as a recording instrument, frequency parameter, data quality control carried out, and descriptive data analysis.

B: Direct observation, observational design, observation instrument, video as a recording instrument, order parameter, data quality control carried out, and descriptive data analysis.

C: Direct observation, observational design, no observation instrument, video as a recording instrument, frequency parameter, no data quality control carried out, and descriptive data analysis.

D: Direct observation, no observational design proposed, no observation instrument, video as a recording instrument, frequency parameter, no data quality control carried out, and the data analysis was the detection of behavior patterns

E: Direct observation, no observational design proposed, no observation instrument, video as a recording instrument, order parameter, no data quality control carried out, and the data analysis was the detection of behavior patterns

F: Miscellaneous

4. Conclusions

The contributions of the main conclusions in the study highlight that it is feasible and useful to use the systematic methodological review, and, furthermore, that the proposed profiling may be useful for future characterization of primary documents and to increase the interest of the scientific community with a view to its optimization.

References

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine, 6*(7), e1000100.

Portell, M., Anguera, M. T., Chacón-Moscoso, S., & Sanduvete Chaves, S. (2015). Guidelines for reporting evaluations based on observational methodology. *Psicothema, 27*, 283–289. doi: 10.7334/psicothema2014.276

Pluye, P., Robert, E., Cargo, M., Bartlett, G., O'Cathain, A., Griffiths, F., & Rousseau, M. C. (2011). *Proposal: A mixed methods appraisal tool for systematic mixed studies reviews*. Montréal: McGill University, 2, 1–8.

Preciado, M., Anguera, M. T., Olarte, M., & Lapresa, D. (2019). Observational studies in male elite football: A systematic mixed study review. *Frontiers in Psychology, 10*, 2077. doi: 10.3389/fpsyg.2019.02077

Empirical research in observational methodology (3): Sport and physical activity (III)

José Luis Losada[1], M. Teresa Anguera[1]

[1]*University of Barcelona, Barcelona, Spain*

1. State of the art

Systematic observation, essentially characterized by focusing on the scientific study of spontaneous or habitual behavior in natural contexts, has not only been consolidated during the last decades, but the scope of application has been considerably expanded, revealing itself as flexible, useful, and of great rigor, characteristics that constitute its fundamental virtues. Its nature as a scientific method makes it suitable for psychologists in a wide spectrum of research and professional areas.

2. New perspectives and contributions

In this Symposium, four papers are presented, three of which refer to the field of physical activity and sport (mainly soccer, basketball, and fitness) and violin, and methodologically a special emphasis is made on: (1) *mixed methods*, which are applied to interviews, from an ecological and holistic perspective; (2) indirect observation, also from interviews, and carrying out *quantitizing* (analysis of polar coordinates) from an indirect observation instrument and the codes generated; (3) *mixed methods*, from quantitative data of a physiological nature and qualitative data obtained from a questionnaire, and (4) generalizability analysis, applied to the systematic observation of the interpretation in student handling of the violin.

3. Research and practical implications

More and more specific aspects are deepened in observational methodology, such as *quantitizing*, generalizability, coding in indirect observation, *T-Patterns* analysis, stability of sequential analysis, or polar coordinate analysis, among others, and as a consequence, a large number of works that use observational methodology have been published in journals with a high impact factor. Undoubtedly, the culture of systematic observation is progressively intensifying, being the only possible methodology in a large number of situations, whenever an interest exists in studying spontaneous or habitual behavior, in a non-artificial context, and ensuring that there is visual and/or auditory perceptivity. Furthermore, in this online 9th *European Congress of Methodology* we are interested in highlighting that we are working within the framework of *mixed methods*, which are currently in a phase of constant growth throughout the world, and we emphasize that observational methodology, according to its profile, can be considered as a *mixed method* in itself, taking into account the QUAL-QUAN-QUAL transition in its successive stages. This consideration opens up a relevant space for increased interest in quantitizing

within observational methodology, leading to a wide spectrum of practical implications in many substantive areas.

Keywords: *Mixed methods*; indirect observation; polar coordinate analysis; generalizability.

E-mails: jlosada@ub.edu; tanguera@ub.edu

Talented Portuguese football players – Genes or environment?

Hugo Sarmento[1], M. Teresa Anguera[2], Duarte Araújo[3]

[1] *University of Coimbra, Research Unit for Sport and Physical Activity (CIDAF), Faculty of Sport Sciences and Physical Education, Coimbra, Portugal,*

[2]*Faculty of Psychology, University of Barcelona, Barcelona, Spain,*

[3]*Spertlab, Facultade de Motricidade Humana, Universidade de Lisboa, Lisboa, Portugal*

Abstract

Purpose: The specificities of how expertise is achieved in Association Football, are being repeatedly investigated by many researchers through a variety of approaches and scientific disciplines (Sarmento et al., 2018). The purpose of this study was to compare training and practice, and psychosocial constraints of biographical histories of the Golden Generation of Portuguese football (under-20 world championships: 1989 (Riyadh) and 1991 (Lisbon)). *Method*: A mixed method design (QUAN/QUAL) was used in this study (Anguera et al., 2012), which adopted the holistic, ecological approach. The software QSR NVivo 10 was used in coding the transcripts of the interviews. Mann-Whitney U tests and the Friedman test were used to compare elite (players that represented the main national team at adult age) and sub-elite (players that never represented the main national team at adult age) groups. *Results*: The results reveal interesting patterns concerning: (1) specificity and volume of practice; (2) psychological factors; (3) technical and tactical skills; (4) anthropometric and physiological factors; (5) relative age effect; (6) performance-related genes, (7) injury-related genes, (8) body composition-related genes, and; (9) cardiac adaptations.

Keywords: Soccer, genetic, psychosocial influences, Textual units, Mixed methods.

Funding: HS and MTA gratefully acknowledge the support of a Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (Ministerio de Economía y Competitividad, Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema I+D+i), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mails: hugo.sarmento @uc.pt; tanguera@ub.edu; daraujo@fmh.ulisboa.pt

1. Introduction

Traditionally, genetic influences have been associated with specific psychological and physiological factors related to sporting performance (Sarmento & Araújo, 2020). Nevertheless, environment assumes a key role in the process of talent identification and development. To better understand this process, we should take into account three of the key theoretical assumptions that integrate the many facets of the ecological dynamics framework: (1) (expert) performance emerges from the performer-environment system; (2) to understand the performance of an individual, an analysis of the behaviors offered by his or her environment (i.e., affordances or opportunities for action) is necessary; and (3), performance emerges (as a result of self-organization) under interacting constraints for in-depth descriptions of the ecological dynamics approach to sporting expertise (for more details about the ecological dynamics approach to sporting expertise, see Araújo & Davids, 2011; Araújo, Dicks & Davids, 2019; Araújo, Hristovski et al., 2019; Davids et al., 2015; Davids et al., 2017).

2. Method

The present study adopted the holistic, ecological approach through a mixed-method design (QUAN/QUAL) (Anguera et al., 2012).

2.1. Participants

The present study included 31 U-20 Portuguese football world champions from the 1989 and 1991 world cup national teams. Of these players, 19 had been international players at senior level (hereafter, "experts"), while 15 footballers had not achieved international status as adult players (hereafter, "non-experts").

The two coaches of this generation of players were interviewed, as were another seven actors who played essential roles in the two championship wins (i.e., the World Football Championships of Riyadh and Lisbon).

2.2. Procedure

A retrospective interview was employed to trace players' entire careers. Each player was previously contacted by e-mail or by phone. Each participant took part in an in-depth, face-to-face interview (duration from 60 to 240 minutes) with the principal investigator. The structure was similar for all interviews which were all digitally recorded.

The data were collected via a Portuguese football-specific adaptation of the interview protocols of Côté, Ericsson, and Law (2005) and Fraser-Thomas, Côté, and Deakin (2008).

Coaches were interviewed using a separate interview structure and were asked to comment on (1) the general organization of football in Portugal during the Golden Generation; (2) the processes of talent identification/development; (3) training resources/facilities; (4) the training process; and (5) barriers and facilitators to players' development during different phases of their careers.

Medical staff members, club officers, teachers from sports faculties, members of sport-related governmental structures, and journalists were asked to comment on club/school values, macro-environmental and historical dimensions, and financial and human resources matters.

Additionally, the Portuguese Football Federation (FPF) website was also analyzed to collect data about players and games. National sports newspapers, personal documents provided by players, and official documents provided by the FPF officers were also analyzed.

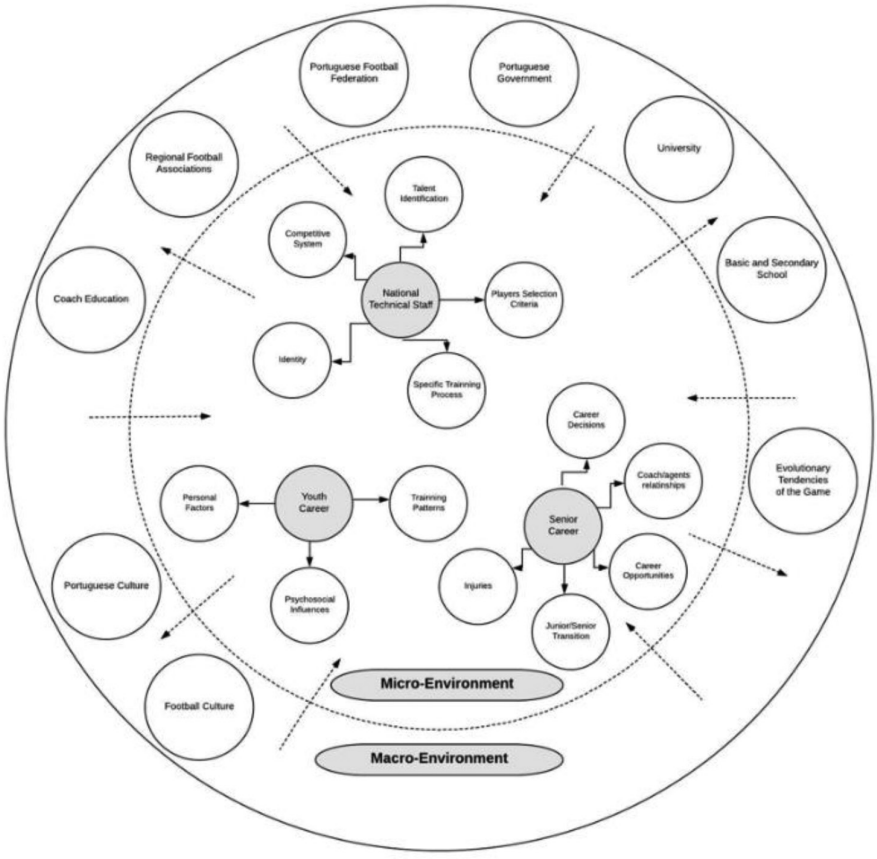
2.3. Data Analysis

Data were analyzed through quantitative and qualitative techniques. For quantitative variables, non-parametric procedures via IBM SPSS software were applied. Mann-Whitney U tests were used to compare groups at each stage, while Friedman tests with pairwise comparisons tested differences across the stages of development for each group. Concerning qualitative data, all interviews and notes from newspapers and official documents were transcribed and coded using a deductive-inductive approach. Transcripts were read repeatedly to promote familiarization with and immersion in the underlying data (Creswell, 2007). Deductive coding was based on a node tree that was built to reflect the working models and primarily involved high-order themes. Inductive coding expanded the node tree when new categories or ideas emerged. The software QSR NVivo 12 was used to code the transcripts of the interviews.

Different techniques were utilized in this study to establish trustworthiness. Member checks are the most crucial technique for establishing credibility and they occurred twice in this study: (1) they first took place during the debriefing session that was held at the end of each interview; (2) the second phase, a full verbatim transcript of each interview was sent to the participants. The participants had another opportunity to clarify, add to, or eliminate any comments that they had made during the interview. Additionally, the trustworthiness of the data was ensured by a panel of three sports psychology experts who analyzed all meaning units, themes, and categories.

3. Results

Based on the analysis performed we propose the Ecological Model of Development of Portuguese Football's Golden Generation (EMDPFGG), considering both the macro- and micro-structure of contextualized histories and practices (Figure 1).



The analysis of training patterns and psychosocial influences reveals similar tendencies between experts and non-experts (Table 1).

Table 1. Descriptive, Friedman and pairwise statistics of experts and non-experts' training patterns and psychosocial influences

	Experts				Non-Experts			
	Stage 1 6-12	**Stage 2 13-15**	**Stage 3 16-18**	**Comparison across the stages**	**Stage 1 6-12**	**Stage 2 13-15**	**Stage 3 16-18**	**Comparison across the stages**
	Training Patterns							
Structured activities[1]	1.64 (.84)	1.29 (.61)	1.07 (0.27)	$\chi2(2) = 8.40$, $p = 0.02$	1.44 (.51)	1.06 (.25)	1.00 (0.00)	$\chi2(2) = 12.29$, $p = 0.00$
Hours of Structured activities[2]	697.9 (207.1)	767.7 (173.7)	681.6 (130.9)	$\chi2(2) = 12.29$, $p = 0.00$ Stage 1 ≠ Stage 3	675.6 (82.0)	753.1 (124.9)	835 (83.5)	$\chi2(2) = 10.07$, $p = 0.01$ Stage 1≠ Stage 2; Stage 2≠ Stage 3;
Unstructured activities[1]	1.79 (0.69)	1.5 (0.5)	1.4 (0.49)	$\chi2(2) = 10.52$, $p = 0.01$	1.69 (.70)	1.25 (.45)	1.06 (.25)	$\chi2(2) = 6.22$, $p = 0.04$
Hours of Unstructured activities[2]	1537.14 (449.8)	1475.5 (500.1)	398.6 (158.5)	$\chi2(2) = 29,79$, $p = 0.00$ Stage 1≠ Stage 3; Stage 2≠ Stage 3	1461..3 (293.6)	1340.0 (302.9)	458.8 (109.6)	$\chi2(2) = 26,09$, $p = 0.00$ Stage 1≠ Stage 3; Stage 2≠ Stage 3
	Psychosocial Influences							
Parent Support[3]	80.9 (14.5)	90.3 (8.1)	95.3 (6.2)	$\chi2(2) = 18,88$, $p = 0.00$ Stage 1 ≠ Stage 3	66.4 (32.9)	82.1 (30.2)	90.4 (26.7)	$\chi2(2) = 15,2$, $p = 0.00$ Stage 1 ≠ Stage 3
Parent Pressure[3]	1.9 (4.4)	2.8 (4.8)	2.8 (4.8)	$\chi2(2) = 4,00$, $p = 0.14$	1.1 (2.9)	3.1 (6.3)	2.5 (4.7)	$\chi2(2) = 6,00$, $p = 0.05$
Sibling Influence[4]	1.6 (1.8)	2.0 (1.9)	1.8 (1.9)	$\chi2(2) = 3,60$, $p = 0.12$	0.9 (1.2)	1.1 (1.4)	0.9 (1.2)	$\chi2(2) = 2,7$, $p = 0.10$
Coach support[3]	69.7* (11.6)	84.4 (11.5)	92.8 (6.8)	$\chi2(2) = 28,9$, $p = 0.00$ Stage 1 ≠ Stage 2,3	50.4* (28.0)	80.4 (25.3)	82.5 (27.9)	$\chi2(2) = 17,74$, $p = 0.00$ Stage 1 ≠ Stage 2,3
Sport peer influence[4]	3.1 (1.2)	4.6 (0.6)	4.8 (0.4)	$\chi2(2) = 20,93$, $p = 0.00$ Stage 1 ≠ Stage 2,3	2.4 (1.3)	3.8 (1.6)	4.3 (1.4)	$\chi2(2) = 17,57$, $p = 0.00$ Stage 1 ≠ Stage 2,3
School peer influence[4]	2.1 (0.9)	2.0 (0.9)	1.9 (0.9)	$\chi2(2) = 2,27$, $p = 0.32$	1.5 (0.9)	1.7 (1.1)	1.7 (1.3)	$\chi2(2) = 1,53$, $p = 0.47$

Note: 1 Number per year; 2 Number of hours per year; 3 percentage; 45-point Likert Scale; * There are no statistical differences between the two groups across the different variables, exception for the variable "Coach Support" in Stage 1 (6-12 years).

4. Conclusions

This study enables us to conclude that the success of this generation of footballers was not due exclusively to the talent of the players but also to the active leadership and holistic vision of their chief coach. Mr. Queiroz promoted a set of deep reforms (ranging from the training process to organizational issues) that defined football in Portugal. The results of this study suggest that coaches who intend to improve the performance of their athletes should adopt holistic perspectives (physical, technical, tactical, etc.). Focusing on these aspects helps footballers to overcome the traditional reductionism that characterizes most scientific studies and practices implemented in the field. In this sense, coaches, practitioners and scientists need to shift their attention to factors that go beyond the individuality of the athlete. These aspects include players' involvement in their assessments and practices, as well as enhanced opportunities/affordances that are available only in specific contexts at specific times.

References

Anguera, M. T., Camerino, O., & Castañer, M. (2012). Mixed methods procedures and designs for research on sport, physical education and dance. In O. Camerino, M. Castañer, & M. T. Anguera (Eds.), *Mixed Methods Research in the Movement Sciences - Case studies in sport, physical education and dance* (pp. 3–28). Oxon: Routledge.

Araújo, D., & Davids, K. (2011). Talent development: From possessing gifts, to functional environmental interactions. *Talent Development & Excellence Interactions, 3*(1), 23–25.

Araújo, D., Dicks, M., & Davids, M., (2019). Selection among affordances: a basis for changeling expertise in sport. In M. L. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology* (pp. 557–580). Cambridge, MA: The MIT Press.

Araújo, D., Hristovski, R., Seifert, L., Carvalho, J., & Davids, K. (2019). Ecological cognition: Expert decision-making behaviour in sport. *International Review of Sport and Exercise Psychology, 12*, 1–25.

Côté, J., Ericsson, K., & Law, M. (2005). Tracing the development of elite athletes using retrospective interview methods. *Journal of Applied Sport Psychology, 17*, 1–19.

Creswell, J. (2007). *Qualitativ Inquiry & Research design: Chosing among five approaches*. Thousand Oaks: Sage Publications.

Davids, K., Araújo, D., Seifert, L., & Orth, D. (2015). Expert performance in sport: An ecological dynamics perspective. In J. B. & D. Farrow (Eds.), *Routledge Handbook of Sport Expertise* (pp. 273–303). London: Routledge.

Davids, K., Güllich, A., Araújo, D. & Shuttleworth, R. (2017). Understanding environmental and task constraints on athlete development: Analysis of micro-structure of practice and macro-structure of development histories. In *Routledge Handbook of Talent Identification and Development in Sport* (Edited by J. Baker, S. Cobley, J. Schorer & N. Wattie, pp.192–206). London: Routledge.

Fraser-Thomas, J., Côté, J., & Deakin, J. (2008). Examining Adolescent Sport Dropout and Prolonged Engagement from a Developmental Perspective. Journal of Applied Sport Psychology, 20(3), 318-333. doi:10.1080/10413200802163549

Sarmento, H., & Araújo, D. (2020). Readiness for career affordances in high-level football: Two case studies in Portugal. *High Ability Studies*, 1–15. doi:10.1080/13598139.2020.1728191

Quantitizing in interviews with senior coaches in basketball: Vectorization of answers through a polar coordinate analysis

Hermilo Nunes[1], Xavier Iglesias[1], M. Teresa Anguera[2]

[1]*INEFC, University of Barcelona, Spain,*

[2]*Faculty of Psychology, University of Barcelona, Spain*

Abstract

This work is part of a broader investigation on ball screening in basketball, and it is intended to contrast the results obtained through direct observation with the expert opinion of the coaches involved in the analyzed team. In-depth interviews were conducted with 6 coaches, and therefore, this study focused on indirect observation. The interview guide, containing 17 questions, was prepared. Once the interviews with the coaches were carried out, after arranging the day and time, a custom indirect observation instrument was constructed, consisting of 2 dimensions, which gave rise to 4 and 15 subdimensions, respectively, from which category systems were built, and a code was assigned to each category. Using this instrument, the textual units that made up each of the interviews were coded, and then recoded *a posteriori*. The quality control of the intraobserver data was carried out, which was satisfactory. Polar coordinate analysis was applied to the records to ascertain the interrelation between two of the recoded focal behaviors (positive assessment and negative assessment) which we proposed, with the others, and whose relationships were vectorized.

Keywords: Basketball; pick and roll; ball screen; textual units; polar coordinate analysis; mixed methods.

Funding: This study has been supported by Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mails: hermilo@hotmail.com; xiglesias@gmail.com; tanguera@ub.edu

1. Introduction

Basketball is a highly complex collective sport, where teams, which share a common playing space, have antagonistic objectives conditioned by the presence or absence of ball possession. The game dynamics adjust to an internal logic, generating a set of defined and different behaviors for the players.

With the natural improvement of basketball, the appearance of the joint strategies created by the coaches and that currently predominate also begins. In this study, the objective is the analysis of in-depth interviews carried out with basketball coaches with extensive experience about the Unicaja Málaga team.

2. Method

2.1. Design

This communication presents a small part of a much broader work, in which a professional basketball team belonging to the ACB League (Spain), which is Unicaja *Málaga,* was analyzed in matches played with 17 different rivals. Since our objective here was to contrast the results obtained by direct observation with the expert opinion of the coaches involved in the analyzed team, in-depth interviews were conducted with high-level basketball coaches who were totally familiar with the sports behavior of Unicaja.

It was an indirect observation study, since the material was the transcripts of in-depth interviews, with an N/P/M (Nomothetic/Punctual/Multidimensional) design (Anguera et al., 2011). It was nomothetic because the interviews were done independently by the coaches, punctual because all the interview questions were answered in one session, and multidimensional because the answers to the questions asked required different dimensions in the indirect observation instrument.

2.2. Participants

The participants were 6 basketball coaches: Alejandro García Reneses and Jesús Mateo Díez as successive head coaches of the Málaga SAD Basketball Club - Unicaja Málaga in 2010-2011; Joaquim Costa Puig, Francisco Aurioles Moreno and Ángel Luis Sánchez-Cañete Calvo as the assistant coaches of the aforementioned team. And finally, Sergio Scariolo in his role as Spain's national male team basketball coach who could analyze the data collected in this study with greater neutrality.

We identify these coaches by name based on the fact that they are unique people, that their duties at this stage are publicly known, and that all of them have signed an informed consent authorizing us to express their opinions.

2.3. Indirect observation instrument

To carry out the interviews, a protocol containing 17 questions was prepared, which was previously sent to the interviewed coaches in case they wanted to reflect on it beforehand. As an illustration, two of the questions are detailed below:

Nº 1: *There are an average of 33 ball screens made per team and per game. Did you expect this result? Why? In addition, an average of 4 ball screen simulations were observed per team and per game. Did you expect this result? Why?*

Nº 12: *Knowing that 51% of the actions registered after the ball screen end in a shot, what do you think of these numbers? Would you continue to use ball screening as a weapon for your team's offensive actions? Why?*

A custom-made instrument (*ad hoc*) was constructed from the reading of the interview transcripts. Two dimensions were proposed: (1) Generic answer, and (2) Justification of the answer. From dimension (1) the sub-dimensions were set out: Significant content, non-significant content, emotional content, and limiting content. From dimension (2) the sub-dimensions were set out: Argumentation of the planning of the team, temporary argumentation of the game, special argumentation of the game, regulatory argumentation, argumentation of the results of the game, technical argumentation, Individual tactical argumentation, collective tactical argumentation, reasoning leading to different decision-making, physical argumentation, psychological argumentation, team argumentation, player/coach argumentation, comparisons, and miscellaneous.

An exhaustive and mutually exclusive system of categories was prepared from each of the subdimensions, and a code was assigned to each of them.

After registration, a recoding was performed in three macrocodes: R1 (positive responses), R2 (negative responses), and R3 (neutral responses).

2.4. Procedure

The transcription of the interviews was segmented into textual units, combining interlocutory and syntactic criteria (Anguera, in press; Krippendorf, 2013).

From each of the interviews, a matrix of codes was obtained, consisting of type II data (Bakeman, 1978), which were concurrent and event-based; that is to say, that in each row of the matrix, corresponding to a textual unit, there were co-occurrence codes corresponding to categories of different dimensions/subdimensions, since the criterion of mutual exclusivity was met in each of them.

2.5. Data quality control

A data quality control was carried out, calculating the intraobserver agreement from a second coding of 10% of the text of the transcripts, and a kappa coefficient (Cohen, 1960) of 0.97 was obtained, which was satisfactory. The calculation of the kappa coefficient was performed using the GSEQ5 program (Bakeman & Quera, 2011).

3. Results

3.1. Polar coordinate analysis

This analysis technique aims to build a map that shows the statistical association relationships that exist between the different codes of conduct, and specifically between the one that is considered central or nuclear, called *focal behavior*, and all the others, which are *conditioned behaviors*, which establish whether a relationship exists, and, if applicable, of what type and intensity are these relationships. The polar coordinate analysis, which considers the adjusted residuals obtained in the lag sequential analysis as data, works by complementing prospective (forward) and retrospective (backward) perspectives, and enables us to know how the relationship between focal behavior and the conditioned behaviors varies or evolves over time. Consequently, this analysis is based on the concepts of prospectivity and retrospectivity. Sackett (1980) applied Bakeman's (1978) concept of prospectivity flawlessly, but considered forward hindsight, from a negative lag, going from lag -5 to lag -4, from this to -3, and so on successively, which could be criticized.

Sackett (1980) had the great success of using the parameter Z_{sum} that Cochran (1954) had proposed, which materialized a large reduction in data, provided that they were independent.

Sackett applied it to the adjusted residual values obtained (which were independent values between them because each one responded to a different calculation since it was a different lag) considering the criterion behavior of the sequential analysis as focal and the conditioned behaviors on positive lags to obtain the Z_{sum} prospective values, and the adjusted residual values obtained considering the criterion behavior of the sequential analysis as focal and the conditioned behaviors in negative lags to obtain the Z_{sum} retrospective values. The number of positive and negative lags had to be the same (Sackett, 1980), and lags +1 to +5, and -1 to-5 were considered.

From the prospective and retrospective Z_{sum} values, Sackett (1980) proposed a vectorization of the relationships between focal and conditioned behaviors. Each vector had as length or radius $Length = \sqrt{(Z_{sum\ \text{prospective}})^2 + (Z_{sum\ \text{retrospective}})^2}$ and as an angle $\varphi = Arc\ sen \frac{Z_{sum\ \text{retrospective}}}{Length}$.

As many vectors as conditioned behaviors are obtained, and all graphically have the origin of the focal behavior. Because the Z_{sum} values (prospective and retrospective) have a positive or negative sign, the corresponding vectors are plotted taking into account that the prospective Z_{sum} is represented on the abscissa axis, and the Z_{sum} retrospective on the ordinate axis.

The calculation of the parameters corresponding to the quadrant, prospective Zsum, retrospective Zsum, radius length and angle corresponding to each of the conditioned behaviors was carried out using the HOISAN program (Hernández-Mendo, et al., 2012).

3.2. Results (partial) obtained

The analysis was performed for each of the 6 coaches, with various focal behaviors. As an illustration, we have selected the polar coordinate analysis corresponding to coach 2, considering the category "efficacy on the scoreboard" as the focal behavior, and as conditioned behaviors GUNI (Win Unicaja), DB2P (*Defender of B2 performs the push defense*), GRIV (Win rival), DB1P3 (Defender of B1 passes 3rd), B1B2DB7 (*Position of the screener and the ball handler at the time of pick and roll stands at B7*), BZONAB (Grouping from space observation zones B1 to B7), TBV (Vertical screen), E (Draw), TC (Defense all field), DB2PEN (Defender of B2 makes *open* defense), DB1N (Defender of B1 denies the screen), and DB1P4 (Defender of B1 passes 4th) were chosen, in addition to others that were not significant.

Table 1. Quadrant and parameters corresponding to the conditioned behaviors in the analysis of polar coordinates about the answers of Coach 2, with "effectiveness on the scoreboard" as the focal behavior.

Category	Quadrant	Prospective P.	Retrospective P.	Ratio	Length	Angle
GUNI	I	2.71	0.91	0.32	2.86 (*)	18.65
DB2P	I	2.13	2.16	0.71	3.08 (*)	45.4
GRIV	II	-2.65	0.16	0.06	2.65 (*)	176.45
DB1P3	II	-2.43	0.4	0.16	2.46 (*)	170.75
B1B2DB7	II	-0.56	2.23	0.97	2.3 (*)	104.09
BZONAB	II	-0.08	2.29	1	2.29 (*)	91.97
TBV	II	-1.61	1.37	0.65	2.11 (*)	139.66
E	III	-1.26	-2.13	-0.86	2.47(*)	239.37
TC	III	-0.15	-2.12	-1	2.12 (*)	265.83
DB2PEN	III	-3.27	-0,65	-0,19	3.33 (*)	191.22
DB1N	IV	1.87	-0.62	-0.32	1.97 (*)	341.57
DB1P4	IV	2.07	-0.66	-0.3	2.17 (*)	342.47

(*) Vectors with statistical significance (values > 1.96)

Table 1 presents the values corresponding to the quadrant, as well as the parameters prospective Z_{sum}, retrospective Z_{sum}, radius length and angle corresponding to each of the conditioned behaviors.

For the correct interpretation of the relationships between focal and conditioned behaviors, taking into account the quadrant in which each of the vectors is located, Table 2 shows the interpretive meaning of each of them.

Table 2. Interpretive meaning of each of the quadrants.

Quadrant	Sign of the prospective Zsum	Sign of the retrospective Zsum	Interpretive meaning
I	+	+	The focal and conditioned behaviors mutually activate each other
II	-	+	The focal behavior inhibits the conditioned one, and this one activates the focal behavior
III	-	-	The focal and the conditioned behavior mutually inhibit each other
IV	+	+	The focal behavior activates the conditioned one, and this one activates the focal behavior

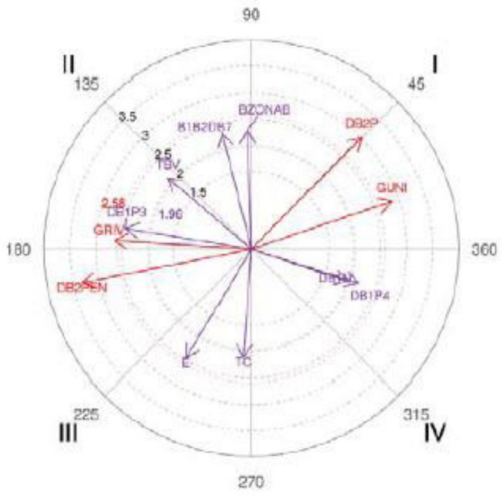


Figure 1. Vectors corresponding to the conditioned behaviors in the analysis of polar coordinates about the answers of Coach 2, with "effectiveness on the scoreboard" as the focal behavior.

4. Conclusions

The results of Table 1 and Figure 1 show how "effectiveness on the scoreboard", which is the focal behavior, presents mutual activation with the GUNI and DB2P categories, and mutual inhibition with E, TC and DB2PEN. There is an asymmetric relationship between the focal behavior, which inhibits GRIV, DB1P3, B1B2DB7, BZONAB and TBV while it is activated by these categories. And, finally, there is also an asymmetric relationship between the focal behavior, which activates DB1N and DB1P4, and is inhibited by them.

In short, the polar coordinate analysis technique is revealed to be extraordinarily suitable and adequate to obtain an interrelational map between codes of conduct, even in indirect observation studies, such as this one, in which the data comes from responses to in-depth interviews.

References

Anguera, M.T. (en prensa). Desarrollando la observación indirecta: Alcance, proceso, y habilidades metodológicas en el análisis de textos. En C. Santoyo (Coord.), *Patrones de habilidades metodológicas y conceptuales de análisis, evaluación e intervención en ciencias del comportamiento*. Ciudad de México: UNAM/PAPIIT, IN306715.

Anguera, M.T., Blanco-Villaseñor, A., Hernández-Mendo, A., y Losada, J.L. (2011). Diseños observacionales: ajuste y aplicación en psicología del deporte. *Cuadernos de Psicología del Deporte, 11*(2), 63–76.

Bakeman, R. (1978). Untangling streams of behavior: Sequential analysis of observation data. In G.P. Sackett (Ed.), *Observing Behavior, Vol. 2: Data collection and analysis methods* (pp. 63–78). Baltimore: University of Park Press.

Bakeman, R. & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge: Cambridge University Press.

Cochran, W. G. (1954). Some methods for strengthening the common $\chi 2$ tests. *Biometrics, 10*, 417–451.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Hernández-Mendo, A., López-López, J. A., Castellano, J., Morales-Sánchez, V., and Pastrana, J. L. (2012). Hoisan 1.2: programa informático para uso en metodología observacional. *Cuadernos de Psicología del Deporte 12*, 55–78.

Krippendorf, K. (2013). *Content analysis. An introduction to its methodology*, 3rd ed. Thousand Oaks, Ca.: Sage.

Sackett, G.P. (1980). Lag sequential analysis as a data reduction technique in social interaction research. In D.B. Sawin, R.C. Hawkins, L.O. Walker & J.H. Penticuff (Eds.), *Exceptional infant. Psychosocial risks in infant-environment transactions* (pp. 300–340). New York: Brunner/Mazel.

Observation system of body posture in violin interpretation: a study with elementary violin students

Daniel Lapresa[1], Angélica Bastida[1], Javier Arana[1]

[1] *Department of Educational Sciences, University of La Rioja, Logroño, Spain*

Abstract

Purpose: Within observational methodology, an ad hoc observation system ad hoc has been designed to observe, analyze and interpret the position that best supports violin performance.

Method: The observation instrument was constructed after an exhaustive theoretical revision. Concordance between the records was guaranteed from the violin stance observation instrument, the reliability of melodic error records in the performance, and the validity of the observation system designed in the theoretical framework of the theory of generalizability.

Results: Subsequently, to demonstrate the operability of the designed observation system, the postural performance was analyzed in the violin interpretation of a short melodic piece by non-professional musicians in their 2nd Elementary Grade violin studies.

Conclusions: The observation system has made it possible to determine specific technical aspects (adjustment and error) with respect to the ideal technical pattern, both intra-interpreter and inter-interpreter.

Funding: The authors gratefully acknowledge the support of a Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0]. In addition, authors thanks the support of the Generalitat de Catalunya Research Group, *GRUP DE RECERCA I INNOVACIÓ EN DISSENYS (GRID). Tecnología i aplicació multimedia i digital als dissenys observacionals* [Grant number 2017 SGR 1405].

E-mails:daniel.lapresa@unirioja.es; anglog91@gmail.com; javier-sabino.arana@unirioja.es

1. Introduction

The musical study of an instrument such as the violin requires dedication and melodic practice but, in addition, the acquisition of a very determined body technique that can affect performance. On the other hand, the unnatural nature of the movements performed when playing the violin can lead to physical and biomechanical problems that have a negative impact on the interpreter's health (Rosinés, 2010). The objective of this work is to design an observation system that observes, analyzes and intervenes in the body posture of the child violinist during melodic performance.

2. Method

Within the use of observational methodology (Anguera, 1979), an observational design was carried out (Anguera et al., 2011): of inter- and intra-sessional follow-up -three interpretations of each of the participants were analyzed, frame by frame-; Nomothetic -three performers from the 3rd year of Primary Education and the 2nd year of Elementary Education at the Professional Conservatory of Music- and multidimensional -there were several dimensions to be studied related to body posture in violin performance-. To end this section, it is relevant to point out that the observation carried out was direct -in relation to the position of the interpreter- and indirect -in relation to melodic errors-. Observation was non-participating. The performed piece, *Etude*, belongs to Volume I of the Suzuki method and was adjusted to the level of competence of the participants.

2.1. Observation instrument

The observation instrument (Table 1) was designed *ad hoc* from an exhaustive theoretical review of the postural technique in violin interpretation: Blacking (1973), Baillot, Rode and Kreutzer (1974), Rolland, Mutschler and Hellebrandt (1974), Suzuki (1978), Skrgatic, Krapac and Zergollern (1979), Galamian (1985), Tubiana, Chamagne and Brockman (1989) , Gregosiewicz, Okonski and Gil (1990), Turner-Stokes and Reid (1999), Rusinek (2004), Valvasori (2009), Klein-Vogelbach, Lahme and Spirgi-Gantert (2010) and Verrel (2012), among others.

Table 1. Schematic structure of the observation instrument.

Dimension	Category (codes)
Placement of the chin on the chin guard	Inside the chin guard (BM); Above the chin guard (BA); Below the chin guard (BD)
Violin position with respect to the head	85º or less (VA); between 86º and 94º (VR); between 95º and 110º (VOI); greater than 110º (VOII)
Positioning of the left elbow with respect to the violin	Elbow located in the middle of the violin (CM); elbow close to the body (CT); elbow near pins (CC)
Placement of the left arm bow on the violin	Less than 88º with respect to the violin strings (AA); 88º-92º with respect to the violin strings (AR); more than 92º with respect to the violin strings (AO)

Dimension	Category (codes)
Finger position on the bow	All fingers above marks (TM); all fingers outside the marks (TF); index finger inside the mark and the rest outside (ID); middle finger inside the mark and the rest outside (CD); ring finger inside the mark and the rest outside (AD); little finger inside the mark and the rest outside (MD)
Shoulder position in relation to the spine	Shoulders parallel to the ground and perpendicular to the spine at an angle of 85º-95º (HP); right shoulder below left (HD); left shoulder below right (HI).
Position of the feet	Left foot advanced up to one foot in relation to the right foot (IA); left foot forward more than one foot in relation to the right foot (IAA); right foot advanced up to one foot in relation to the left foot (DA); right foot forward more than one foot in relation to the left foot (DAA); feet aligned (DI): both feet are parallel.

2.2. Procedure

Filming was carried out from four different points: frontal, zenith, left lateral and right lateral points. A single piece was performed, repeated three non-consecutive times, by each of the participants. References were used to facilitate observation: lines on the floor that served as a guide for placing the feet of the lectern at a 90º angle between the front and the side; different colored stickers to mark the correct position on the arch of each finger; two strips perpendicular to the axis of the strings, in the center of the violin and at the end of the fingerboard.

Registration was carried out using Lince software (Gabin et al., 2012). It was recorded using the images from the front camera. When necessary, the filming of the upper chamber and the left lateral chamber was used, taking the performed melody as a reference. The agreement between records from the observation instrument -of the posture in violin performance- was carried out intra-observer, from the records of the first of the three executions of each of the participants. It was calculated through Cohen's Kappa coefficient, using the Lince software.

The agreement between records from the observation instrument -of the posture in violin performance- was carried out intra-observer, from the records of the first of the three executions of each of the participants. It was calculated through Cohen's Kappa coefficient, using the Lince software.

The recording of errors in each of the interpretations of each of the participants was made by two experts in musical language, through indirect observation, pointing out the specific note incorrectly performed in the score. In this way, the incorrect note was related to the body posture. For example, participant 1, in interpretation 1: incorrect note DO2C4; multi-event of the corresponding record: BM, VOI, AR, ICD, HP, DA, CT; frame 854.

An analysis of generalizability was also carried out (Cronbach, et al., 1972), based on the work of Blanco-Villaseñor (1993) and within the SAGT software from Hernández-Mendo, et al. (2016). There were three generalizability measurement plans: Participant, Interpretation / Categories; Interpretation, Categories / Participant and Categories, Participant, / Interpretation. To demonstrate the operation of the designed observation system, regular behavior structures were detected in the T-patterns register within the Theme program (Magnusson, 1996 and 2000).

3. Results

Regarding the agreement between the records from the violin posture observation instrument, the agreement was complete (1.0), except in two dimensions. In the dimension "Position of the shoulders in relation to the spinal column", Cohen's Kappa was 0.97, and in the dimension "Position of the fingers in the arch" it was 0.79. This category had the support of marks (*stickers*) that were placed on the arch of each interpreter, to increase the perceptibility of finger changes in order to anticipate the recording difficulties.

Regarding the reliability of the melodic error registers in the execution, it should be noted that there was total agreement in the judgments of the two experts in the nine interpretations; which, in addition, endorsed the competence of the selected observers-experts.

The generalizability analysis verified that the highest percentage of variance was found in the participant-category facet (63.80%), followed by the category facet (22.45%) and the interaction between the participant-interpretation-category facets (13.36%). The relative G Coefficient obtained (0.47) in the Interpretation, Categories / Participant measurement plan, led us to carry out an optimization plan that enabled us to affirm that 40 interpreters would be an appropriate reference value to complete the observational sampling (0.922). The relative G Coefficient obtained (0.951) in the Categories, Participant / Interpretation measurement plan, indicated that three interpretations were sufficient to guarantee the generalizability of the results of each participant, avoiding fatigue due to the repetition of the piece. The results corresponding to the Participants, Interpretations / Categories measurement plan referred to the validity of the observation instrument in the theoretical framework of the Theory of Generalizability. In this case, the relative generalizability coefficient (0.15) reflected the discrimination capacity of the categories facet.

The operation of the designed observation system was demonstrated with the data packages corresponding to each interpretation and in the regular behavior structures detected.

The data package corresponding to each interpretation of each of the participants involved the dumping of reality in a register that contemplated errors in posture, as well as the variation in the different dimensions of behavior contemplated in the observation instrument. In addition, the recording of melodic errors enabled the position adopted by the interpreter to be related to each melodic error.

The T-patterns revealed behavioral patterns (multi-events that make up each row of the record) in the posture adopted in each interpretation of each participant; in fact, only intra-participant T-patterns were detected; that is to say, each interpreter manifested a postural specificity in their interpretation. As an example, Table 2 presents the T-patterns detected in the postural execution of participant 1.

Table 2. T- patterns detected in participant 1.

T-pattern	Piece and occurrences (intra-piece)
((bm,vr,ar,icd,hp,da,ct bm,vr,ar,icd,hp,da,ct)(bm,vr,ar,icd,hp,da,ct bm,vr,ar,icd,hp,da,ct))	Piece 2 (1 occurrence) and 3 (1 occurrence)
(bm,vr,ar,icd,hp,da,ct (bm,vr,aa,icd,hp,da,ct bm,vr,ar,icd,hp,da,ct))	Piece 2 (5 occurrences) and 3 (1 occurrences)
((bm,vr,aa,icd,hp,da,ct bm,vr,ar,icd,hp,da,ct) bm,vr,aa,icd,hp,da,ct)	Piece 2 (5 occurrences) and 3 (1 occurrences)

T-pattern	Piece and occurrences (intra-piece)
(bm,vr,aa,icd,hp,da,ct bm,vr,ar,icd,hp,da,ct)	Piece 2 (5 occurrences) and 3 (2 occurrences)
(bm,vr,ar,icd,hp,da,ct bm,vr,aa,icd,hp,da,ct)	Piece 2 (5 occurrences) and 3 (2 occurrences)
(bm,vr,ar,icd,hp,da,ct bm,vr,ar,icd,hp,da,ct)	Piece 2 (4 occurrences) and 3 (3 occurrences)

4. Conclusions

From the theoretical and technical models of the violin, an observational tool was developed that observes, analyzes and intervenes in body posture during the violinist's performance. Evidence of reliability, generalizability and validity were provided. The operability of the designed observation system was demonstrated analyzing the postural performance in the violin interpretation of a short melodic piece by non-professional musicians.

References

Anguera, M.T. (1979) Observational Typology. *Quality & Quantity. European-American Journal of Methodology, 13*(6), 449–484.

Anguera, M.T., Blanco-Villaseñor, A., Hernández-Mendo, A., y Losada, J.L. (2011). Diseños observacionales: ajuste y aplicación en psicología del deporte. *Cuadernos de Psicología del Deporte, 11*(2), 63–76.

Anguera, M.T., Magnusson, M.S., y Jonsson, G.K. (2007). Instrumentos no estándar. *Avances en medición, 5*(1), 63–82.

Baillot, P., Rode, P., y Kreutzer R. (1974). *Méthode de violon*. Madrid: Antonio Romero.

Blacking, J. (1973). *How musical is man?* Seattle: University of Washington Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Gabín, B., Camerino, O., Anguera, M.T. y Castañer, M. (2012). Lince: multiplatform sport analysis software. *Procedia Computer Science Technology*,46, 4692–4694.

Galamian, I. (1985). *Principles of Violin Playing and Teaching*. Michigan: Englewood Cliffs, Prentice Hall.

Gregosiewicz, A., Okonski, M., y Gil L. (1990). Specific character of movement disorders in children studying string instruments. *Chirurgia Narzadów Ruchu I Ortopedia Polska,* 55, 191–194.

Hernández-Mendo, A., Blanco-Villaseñor, A., Pastrana, J.L., Morales-Sánchez, V., y Ramos-Pérez, F.J. (2016). SAGT: Aplicación informática para análisis de generalizabilidad. Revista Iberoamericana de Psicología del Ejercicio y el Deporte, 11(1), 77–89.

Klein-Vogelbach, S., Lahme, A., y Spirgi-Gantert I. (2010). *Interpretación musical y postura corporal.* Madrid: Ediciones Akal.

Magnusson, M.S. (1996). Hidden real-time patterns in intra- and inter-individual behavior. *European Journal of Psychological Assessment, 12*(2), 112–123.

Magnusson, M.S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers, 32*(1), 93–110.

Rolland, P., Mutschler, M., y Hellebrandt, F. A. (1974). *The teaching of action in string playing: Developmental and remedial techniques [for] violin and viola* (Vol. 1). Illinois: String Research Associates.

Rosinés, M. (2010). Músicos y lesiones. *Biomecánica: Órgano de la Sociedad Ibérica de Biomecánica y Biomateriales*, 18, 16–18.

Rusinek, G. (2004). Aprendizaje musical significativo. *Revista Electrónica Complutense en Educación Musical.* Universidad Complutense de Madrid. 2–5.

Skrgatic, M., Krapac, L., y Zergollern, J. (1979). Radiological analysis of the spine of professional musicians. *Lijecnicki Vjesnik*, 101, 6–379.

Suzuki, S. (1978). *Suzuki, violin school Volumen I.* Florida, Estados Unidos): Warner Bros Publications.

Tubiana, R., Chamagne, P. y Brockman, R. (1989). Fundamental positions for instrumental musicians. *Medical problems of perfoming artists.* 73–76.

Turner-Stokes, L. y Reid, K. (1999). Three-dimensional motion analysis of upper limb movement in the bowing arm of string-playing musicians. *Clinical Biomechanics*, *14*, 426–433.

Valvasori, F. (2009) *El violín tres cuartos y/o siete octavos como alternativa válida definitiva. Análisis de 5 casos particulares*. Conservatorio Superior de Música Felipe Boero.

Verrel, J., Pologe, S., Manselle, W., Lindenberger, U., y Woollacott, M. (2013). Coordination of degrees of freedom and stabilization of task variables in a complex motor skill: expertise-related differences in cello bowing. *Experimental brain research*, 224(3), 323–334.

Empirical research in observational methodology (4): Several fields

Gudberg Jonsson[1], M. Teresa Anguera[2]

[1]*University of Iceland, Reykjavik, Iceland,*
[2] *University of Barcelona, Barcelona, Spain*

1. State of the art

Systematic observation, essentially characterized by focusing on the scientific study of spontaneous or habitual behavior in natural contexts, has not only been consolidated in the last few decades, but the scope of application has been considerably expanded, revealing itself as flexible, useful, and of great rigor, characteristics that constitute its fundamental virtues. Its nature as a scientific method makes it suitable for psychologists in a wide spectrum of research and professional areas.

2. New perspectives and contributions

In this Symposium, two papers are presented, which are focused on interventions in the analysis of conversations and anonymized data. From a methodological side, they refer to: (1) a motivational interview used through the ELAN program, taking a multimodal perspective; and (2) *T-Pattern* analysis and polar coordinate analysis, which are combined in one study with anonymized data.

3. Research and practical implications

Observational methodology is increasingly focusing on specific aspects, such as *quantitizing*, generalizability, coding in indirect observation, *T-Patterns* analysis, stability of sequential analysis, or polar coordinate analysis, among others, and as a consequence, a large number of works that use observational methodology have been published in journals with a high impact factor. Undoubtedly, the culture of systematic observation is progressively intensifying, being the only possible methodology in a large number of situations, whenever an interest exists in studying spontaneous or habitual behavior, in a non-artificial context, and ensuring that there is visual and/or auditory perceptivity. Furthermore, in this online 9th *European Congress of Methodology,* we are interested in highlighting that we are working within the framework of *mixed methods*, which are currently in a phase of constant growth throughout the world, and we emphasize that observational methodology, according to its profile, can be considered as a *mixed method* in itself, taking into account the QUAL-QUAN-QUAL transition in its successive stages. This consideration opens up a relevant space for increased interest in quantitizing within observational methodology, leading to a wide spectrum of practical implications in many substantive areas.

Keywords: Polar coordinate analysis; ELAN coding; T-Patterns.

E-mails: gjonsson@hi.is; tanguera@ub.edu

Proposal of an observational instrument applied to a motivational interview using ELAN

Francisco Molinero-Ruiz[1]

[1]*Researcher, Mimesis, Spain*

Abstract

The motivational interview is a widely validated professional intervention method of facilitating change processes. Its intervention procedure is based on the communicative intervention of a professional in the flow of a conversation which conditions the effectiveness of said intervention. Despite being a collaborative model of change, the evaluation systems developed so far have been based on an assessment of the professional's statements, without contextualizing or specifying the specific flow of conversation of which these evaluations and prescriptions of intervention are derived. In this communication we propose the development of an observational instrument that contextualizes and specifies these conversational flows. This observational instrument incorporates the contributions made from the conversational analysis and microanalysis of communication in which the meanings are built on the continuum of the interactive sequences, and the development of language technologies.

Keywords: Motivational interview; conversational flow; gesture conversation; observational instrument; microanalysis of communication, language technologies.

Funding: This study has been supported by the Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mail: info@pacomolinero.net

1. Introduction

Conversational therapies and professional communication-facilitating interventions for change have proposed different models and interview techniques. These communication-based models aim to be evidence-based models of change.

Despite this, we confirm that there is a gap between the models, concepts and techniques of intervention, and their eventual performance in the conversational flows in which they theoretically take place.

In order to reconnect the models, techniques and concepts that guide the professional's actions with the conversational flows that are actually produced, it is necessary to develop observational instruments that systematize and operatively define the conversational sequence's concepts and techniques that have been developed in the different communication-based intervention models.

In this communication, we describe the components that need to be incorporated in observational instruments to make these models operational and we develop a grid with a system of categories and codes. We propose that this grid is based on a collaborative, interactive model of the construction of meanings in line with the communication microanalysis proposals made by the Bavelas group, and on the other hand, we suggest that it should reflect the multimodal nature of communication in which gestural, paralinguistic and verbal components complement and concatenate.

Likewise, we propose that the observational instrument must be flexible enough to adapt it to different intervention contexts, and to the application of mixed and collaborative methods that allow the use of these instruments in contexts of professional collaboration and reflection in relation to their professional practice.

We propose the use of the ELAN tool because it enables us to develop this grid qualitatively, with enough flexibility to make annotations from the most inductive and qualitative to systematized forms of observation that can naturally be integrated into other statistical programs of sequential analysis and temporal patterns.

2. Method

To draw up this observational instrument, as we have said, we followed the meta-method developed by the Victoria group of communication microanalysis and took a multimodal and sequential perspective of conversations to redefine the concepts and evaluation parameters of motivational interviews in conversational flow.

2.1. Victoria group proposals

The system of categories of the proposed observational instrument integrates the concept of interactive function developed by the micro-analysis of communication group developed by Janet Bavelas at the University of Victoria in Canada.

In this method, special importance is given to the concept of conversation as 'joint action' and the different acts of the interlocutors are defined in terms of their 'interactive functions', that is, the actions performed by each one are operationally inter-defined in reference to the action of the other. These actions can be sustained not only by verbal content but also by gestures or other aspects of multimodal communication.

In fact, what the group called the 'calibration' process or the establishment of shared meanings describes those meanings in terms of the interactive functions involved. Shared meanings

thus become an observable process in the conversation flow, in terms of the operational definitions of those interactive functions.

Some of the most important interactive functions are Questions and Formulations, both defined in terms of their interactive function in conversational flow (McGee, 2004) (Korman et al., 2013). A speech act is said to be a formulation when the interlocutor speaks or comments on something that the interlocutor has previously said. The formulations can be of different types and many motivational interview techniques and micro skills such as different kinds of formulations (reflections, summaries, affirmations, etc.) could be included. The same happens with questions that play a very important role in shaping the course of the interview.

2.2. The multimodal nature of communication

The incorporation into this grid of the categories and codes that describe the multimodal nature of communication applied to the motivational interview has consequences from both a methodological and a theoretical point of view.

From the methodological point of view, we propose an integration of the qualitative and quantitative approaches in accordance with the proposals of Dr. Anguera (Anguera, 2014). The phenomenon of 'serendipity', that is, the unexpected discovery that gives shape and meaning to data paradoxically is a characteristic of both qualitative inquiry and automated data exploration thanks to analysis algorithms (Hunyadi, 2020). Accordingly, together with the annotation of the interactive functions and the actions identified, following the intervention model of the motivational interview, the observation instrument needs to incorporate a coding system of observable behaviors (gestures and prosody) as well as the annotation of the verbal content that enables the description of the 'gesture conversation' as it unfolds in its temporal evolution (Hunyadi, 2016).

2.3 The evaluation systems of the motivational interview

As we have said, the motivational interview is a model of intervention that enhances the language of change, and therefore, bases its effectiveness on the discursive operations that the interviewer performs with that objective. In this sense, different evaluation tools have been developed for the application of the model by an interviewer.

Although the motivational interview model is a collaborative model that encourages the incorporation and recognition of the client-patient point of view, and therefore, their discourse of change, the evaluation systems of the professional's actions do not sufficiently take into account the interactive multimodal nature of communication (traditionally, only audio records have been used, which prevents incorporating the gestural dimension of the conversation).

In this sense, this proposal for an observational instrument adds the evaluated dimensions (MITI and EVEM) and enables the analysis of items with the conversational contexts in which the actions prescribed to the professional are carried out, as well as showing 'how' they are carried out, using the annotation of the interactive functions and their accumulated effects on the conversational sequences that constitute the interviews.

All this information was registered in the ELAN annotation system along with the annotations on the different dimensions of multimodal communication as proposed in the creation of the multimodal communication corpus (Hunyadi 2016). In this way, the observational instrument not only enriched the evaluation system used in motivational interviewing, but also facilitates the creation of a systematized corpus with applications in the field of motivational interviewing training, the supervision of professionals, and especially innovation and improvement

of the intervention model based on the professionals' reflections and the application of analysis algorithms with the discovery of patterns in conversational interactions.

3. Results

Figure 1 shows the structure of the category and code system arranged to record in ELAN with the different dimensions aligned over time. The flexibility of the tool enables the addition, removal or modification of other Annotation Tiers over time.

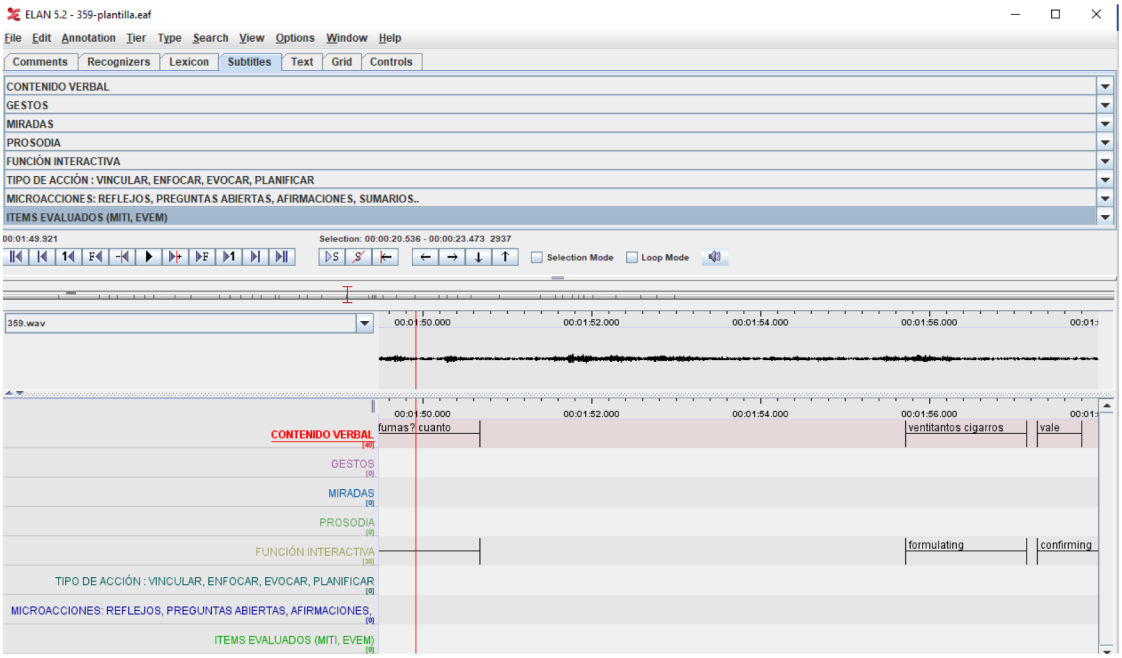


Figure 1. Display in ELAN of the categories and codes to annotate Motivational Interview Cases

This ELAN annotation system (in which the prosody data was imported directly from Praat) enabled us to carry out a qualitative analysis of specific moments that show a specific configuration of the different categories and codes, and their export to programs such as Theme, Lince and Hoisan to examine temporal and sequential patterns of different types of annotations.

4. Conclusions

The motivational interview has established itself as a communication-based intervention model to promote processes of change. Its application and development based on evidence fosters an enriching dialogue between professional practice and research.

Observational methodology is becoming increasingly popular thanks to the advances in technologies that develop integration strategies for qualitative and quantitative methods through the recording of sessions and their treatment with digital tools such as Elan, Praat, Hoisan, Lince and Theme.

In this communication we proposed the use of an observational instrument in ELAN designed to annotate a linguistic corpus that facilitated the treatment of recorded sessions whilst respecting their multimodal and interactive nature, aligning on the timeline the different codes for the different dimensions referring to the model of intervention of the motivational interview, and to the items which evaluate the correct application of the model.

This observational instrument can be used for and adapted to different purposes such as training in the use of the motivational interview model, the supervision of professional practice, and the systematic compilation of a corpus of communication that can be used in different research projects for the improvement of professional practice and basic research on communication processes.

References

Anguera, M.T Sanchez-Algarra, P. (June 27–29 2014) *Quantitizing and Qualitizing: The Benefits of Observational Methodology* Paper presented at the International Mixed Methods Conference Boston College

Bavelas, J., Gerwing,J., Healing,S., Tomori,C. (2017) 'Microanalysis of Face-to-face Dialogue. An Inductive Approach' Chapter in 'Researching Interactive Communication Behavior. A Sourcebook of Methods and Measures' VanLear C.A and Canary D.J Editors

Hunyadi, L. & Szekrényes, I. (Eds.) (2020) *The Temporal Structure of Multimodal Communication*. New York: Springer

Hunyadi, L., Váradi, T., & Szekrényes, I. (2016, December). *Language technology tools and resources for the analysis of multimodal communication.* Paper presented at the Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT-4DH), pages 117–124, Osaka, Japan

Korman, H., Bavelas, J. B., & De Jong, P. (2013). Microanalysis of formulations in solution-focused brief therapy, cognitive behavioral therapy and motivational interviewing. *Journal of Systemic Therapies, 32*, 31–45.

Searching for similarities between T-Patterns and polar coordinate analysis in direct observations

M. Teresa Anguera[1], Gudberg K. Jonsson[2], Pedro Sánchez-Algarra[3]

[1]*Faculty of Psychology, University of Barcelona, Spain,*

[2]*University of Iceland, Reykjavík, Iceland,*

[3]*Faculty of Biology, University of Barcelona, Spain*

Abstract

Purpose: The aim of this paper is the search for similarities in the results of T-Patterns and polar coordinate analysis when analyzing the same dataset. *Method*: Observational methodology is an ideal approach to study the hidden structures underlying real situations. First, we built *ad hoc* a previous observation instrument, that implied making a decision about a proposal of certain dimensions and a category system or catalogue of behaviors for each dimension, and we made a systematic record of episodes, using free software (GSEQ5, HOISAN, LINCE, LINCE PLUS, MOTS, etc.). *Results*: We worked with two databases of anonymized data, in order to comparatively analyze both data analysis techniques: the T-Pattern detection and polar coordinate analysis. These data analysis techniques have a common aim, which is to discover hidden relations between observed behaviors, although each one has a different algorithm and aims. We compared the degree of similarity between the results of code relations obtained from both techniques. We also proposed a guide to help researchers integrate results. *Conclusions*: Two data analysis techniques implied a possible convergence in results, irrespective of subject or field of study: detection of T-Patterns and polar coordinate analysis.

Keywords: T-Patterns analysis; polar coordinate analysis; parameters; observational instruments; direct observation.

Funding: This study has been supported by the Spanish government subproject *Integration ways between qualitative and quantitative data, multiple case development, and synthesis review as main axis for an innovative future in physical activity and sports research* [PGC2018-098742-B-C31] (2019-2021) (Ministerio de Ciencia, Innovación y Universidades / Agencia Estatal de Investigación / Fondo Europeo de Desarrollo Regional), which is part of the coordinated project *New approach of research in physical activity and sport from mixed methods perspective* (NARPAS_MM) [SPGC201800X098742CV0].

E-mail: tanguera@ub.edu

1. Introduction

The recent scientific literature shows how the analysis techniques for the detection of T-Patterns and polar coordinates occupy a relevant role in the direct observation of behavioral episodes in different areas (social interaction, sport, etc.).

The detection of T-Patterns implies discovering hidden relations that reveal aspects of social interaction that are not immediately observable. The recorded episodes of behavior are governed by structures of varying stability, and can be visualized by obtaining T-Patterns, which have proven to be an exceptional analytical tool. These temporal patterns can be detected with THEME 6.0. (Magnusson, 1996, 2000, 2020)

Polar coordinate analysis searches for a vectorial image of the complex network of interrelations between categories that make up the different dimensions of the observation instrument (Sackett, 1980; Anguera, Portell, Hernández-Mendo, Sánchez-Algarra, & Jonsson, in press). The values of length and angle of vectors, and their graphical representation is achieved by the free software HOISAN (Hernández-Mendo, López-López, Castellano, Morales-Sánchez, & Pastrana, 2012).

The aim of this paper is the search for similarities in the results of T-Patterns and polar coordinate analysis when analyzing the same dataset.

2. Method

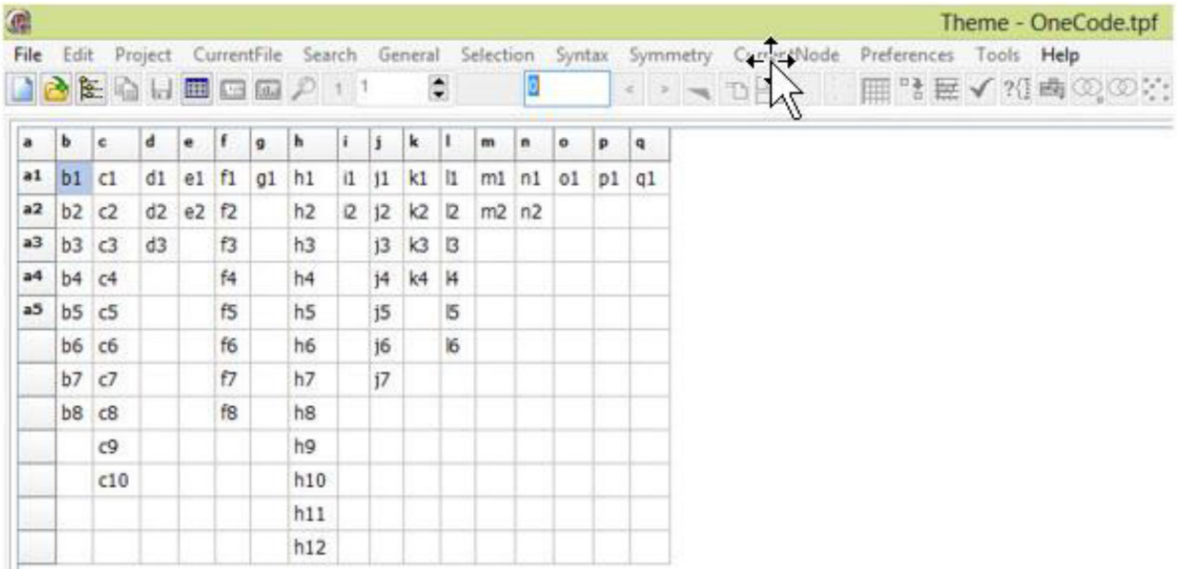
2.1. Databases

We worked with two fully anonymized databases that came from episodes of social interaction that had been registered through systematic observation.

2.2. Observation instrument

The observation instrument had 17 dimensions, and were used to construct a system of categories, which obviously met the requirements of exhaustivity and mutual exclusivity. Table 1 shows the observation instrument in the THEME program, which was used to record the data. It includes at least one code in each row, and one code at most for each dimension.

Table 1. Observation instrument (in THEME program).



a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
a1	b1	c1	d1	e1	f1	g1	h1	i1	j1	k1	l1	m1	n1	o1	p1	q1
a2	b2	c2	d2	e2	f2		h2	i2	j2	k2	l2	m2	n2			
a3	b3	c3	d3		f3		h3		j3	k3	l3					
a4	b4	c4			f4		h4		j4	k4	l4					
a5	b5	c5			f5		h5		j5		l5					
	b6	c6			f6		h6		j6		l6					
	b7	c7			f7		h7		j7							
	b8	c8			f8		h8									
		c9					h9									
		c10					h10									
							h11									
							h12									

3. Results

3.1. T-pattern detection

The detection analysis of T-Patterns was carried out using the THEME program, with a significance level of p<.005, and a minimum number of occurrences of 4 as parameters. Figure 1 shows T-Pattern 1, which is enlarged at the bottom of the figure.

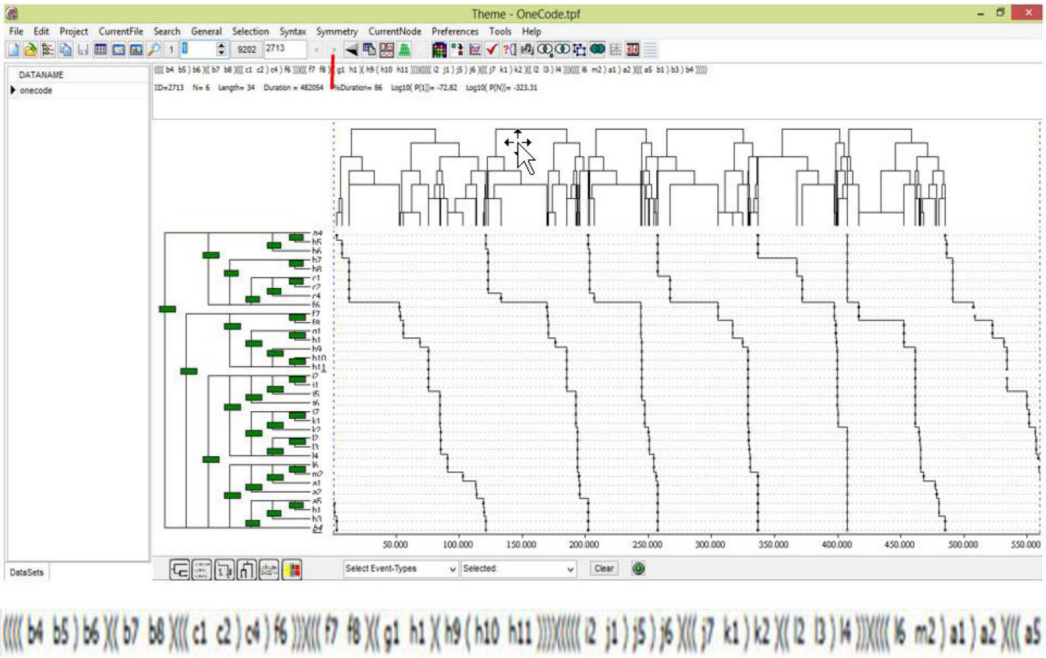


Figure 1. T-Pattern number 1 and tree structure enlarged at the bottom of the figure

The brackets at the beginning of the equation (there is no space to develop them all) show the direct relationship between b4 and b5, and their relationship with b6. In the extracted structure, the relationship between b7 and b8 is also evident, and accordingly between c1 and c2.

3.1. Polar coordinate analysis

For each of the relationships indicated in the first part of the extracted structure, a polar coordinate analysis was performed:

Firstly, using behavior b4 as a focal behavior, Table 2 shows its activating relationship with categories b5 and b6 (highlighted in bold).

Table 2. Parameters corresponding to behavior b4 as focal

Category	Quadrant	Prosp. Zsum	Retrosp. Zsum	Length	Angle
b_b1	II	−0.72	9.09	9.11(*)	94.56
b_b3	II	−0.74	9.1	9.13(*)	94.62
b_b5	**IV**	**9.1**	**−0.75**	**9.13(*)**	**355.38**
b_b6	**IV**	**9.09**	**−0.74**	**9.12(*)**	**355.37**
b_b7	IV	9.08	−0.74	9.11(*)	355.37
b_b8	IV	9.07	−0.74	9.1(*)	355.36

Secondly, using behavior b7 as a focal behavior, an activating relationship was found with category b8 (highlighted in bold).

Table 3. Parameters corresponding to behavior b7 as focal

Category	Quadrant	Prosp. Zsum	Retrosp. Zsum	Length	Angle
b_b1	II	−0.72	9.06	9.08(*)	94.57
b_b3	II	−0.74	9.07	9.1(*)	94.64
b_b4	II	−0.74	9.08	9.11(*)	94.63
b_b5	II	−0.74	9.09	9.12(*)	94.63
b_b6	II	−0.74	9.1	9.13(*)	94.62
b_b8	**IV**	**9.1**	**−0.74**	**9.13(*)**	**355.38**

Thirdly, using behavior c1 as a focal behavior, an activating relationship was found with category c2 (highlighted in bold).

Table 4. Parameters corresponding to behavior c1 as focal

Category	Quadrant	Prosp. Zsum	Retrosp. Zsum	Length	Angle
c_c2	**IV**	**9.1**	**−0.74**	**9.13(*)**	**355.38**
c_c4	IV	9.09	−0.74	9.12(*)	355.37
c_c5	IV	9.08	−0.74	9.11(*)	355.37
c_c6	IV	9.07	−0.74	9.1(*)	355.36
c_c7	IV	9.06	−0.74	9.09(*)	355.36
c_c8	III	−0.74	−0.74	1.04	235
c_c9	III	−0.74	−0.74	1.04	225
c_c10	III	−0.74	−0.74	1.04	225

4. Conclusions

The preliminary results found positive evidence of the relationships between codes shown in the first part of the equation corresponding to T-Pattern 1 in the respective polar coordinate analysis performed, with the respective vectors being significant (length>1.96).

References

Anguera, M. T., Portell, P., Hernández-Mendo, A., Sánchez-Algarra, P., and Jonsson, G. K. (in press). Diachronic analysis of qualitative data. In A.J. Onwuegbuzie and B. Johnson (Eds.), *Reviewer's Guide for Mixed Methods Research Analysis*. London: Routledge.

Hernández-Mendo, A., López-López, J. A., Castellano, J., Morales-Sánchez, V., and Pastrana, J. L. (2012). Hoisan 1.2: programa informático para uso en metodología observacional. *Cuadernos de Psicología del Deporte 12*, 55–78.

Magnusson, M. S. (1996). Hidden real-time patterns in intra- and inter-individual behavior: Description and detection. *European Journal of Psychological Assessment, 12*, 112-123. doi:10.1027/1015-5759.12.2.112

Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers, 32*, 93-110. doi:10.3758/BF03200792

Magnusson, M. S. (2020). T-Pattern Detection and Analysis (TPA) with THEMETM: A mixed methods approach. *Frontiers in Psychology, 10*, 2663. doi:10.3389/fpsyg.2019.02663

Sackett, G.P. (1980). Lag sequential analysis as a data reduction technique in social interaction research. In D.B. Sawin, R.C. Hawkins, L.O. Walker & J.H. Penticuff (Eds.), *Exceptional infant. Psychosocial risks in infant-environment transactions* (pp. 300-340). New York: Brunner/Mazel.

Current methodological trends in the analysis, assessment, and evaluation of intimate partner violence against women

Celia Serrano-Montilla[1], Manuel Martín-Fernández[2]

[1]*Department of Methodology for Behavioral Science, University of Granada, Spain,*

[2]*Department of Social and Methodological Psychology, Autonomous University of Madrid, Spain*

1. State of the art

Intimate partner violence against women (IPVAW) is a relevant public health and social problem. It is the most frequent form of violence experienced by women, with 23.8% being the average prevalence in Western societies (World Health Organization, 2013). Researching this issue is a crucial step to gain a better understanding of the problem and to develop intervention programs and strategies based on scientific evidence. However, the study of IPVAW presents some methodological challenges that we address through the symposium from both quantitative and qualitative approaches.

2. New perspectives and contributions

The studies we show in this symposium represent advances on research, as they provide insightful evidence on different issues related to IPVAW based on methodological approaches that are not commonly used to study this social and health problem. First, we highlighted the importance of using mixed methods to develop adequate measures to assess the attitudes toward intervention among law enforcers in IPVAW. Beyond the traditional approach which provides validity evidence in the later stages of scale development, the mixed methods model of scale development and validation analysis has its advantages. Thus, the validation process starts before item generation, being an ongoing process. Qualitative and quantitative data are also used to support intended uses and interpretations of test scores. Second, it is also important to point out the relevance of using a scoping review approach in order to gain knowledge about the ways social desirability is currently being addressed in research on psychological aggressions against a partner. In contrast with a more "traditional" systematic review, it is used when a research topic is of a controversial and complex nature, and can provide an overview of the type, extent and quantity of the available research. A scoping review approach was chosen to try to raise awareness about how researchers have paid little attention to social desirability assessments in research on psychological aggressions against a partner, in spite of the widespread recognition of the importance of the role of social desirability in underreporting IPV behaviors.

Third, meta-analysis could be regarded as a suitable method for determining whether scientific evidence about motivational strategies improves the effectiveness of IPVAW offender programs and can be generalized across intervention variations and settings. Finally, there is the use of structural equation modeling to identify key aspects in the study of public attitudes towards IPVAW and their potential mediation in the effect of socio-demographic variables in the willingness to intervene in cases of IPVAW.

3. Research and practical implications

This symposium also addressed some methodological challenges from both quantitative and qualitative approaches in the study of intimate partner violence. We showed how mixing both qualitative and quantitative approaches could be useful to develop new instruments to assess police attitudes and to enhance the understanding of Spanish police attitudes toward intervention in IPVAW. Specifically, research implication in psychometric studies regarding scale development or adaptation should take this approach in order to ensure the quality of new psychological measures and the interpretations of test scores. In the practical field, police officers' training could benefit from an assessment system which provides pre and post scores in order to conclude improvements after IPVAW training programs in the Spanish context.

We also presented how different approaches related to the synthesis and analysis of available empirical evidence (i.e., scoping review, systematic reviews and meta-analyses) could lead to future research providing information about response biases (social desirability) on psychological IPVAW assessment, as a first step in trying to address those problems in psychological aggression against a partner assessments, in order to eventually be more precise when detecting cases of intimate partner violence in clinical or forensic settings. Likewise, this approach leads us to point out crucial aspects in the assessment of clinical practice in IPVAW offender intervention programs. Therefore, it would be easier to develop clinical practice guidelines for IPV offender intervention programs and to assess methodological considerations such as the need to accurately report and follow up on the start points of interventions and dropouts, the establishment of follow-up periods which are long enough to assess the persistence of change, or the benefits of data triangulation to improve the validity of IPV offender programs.

Finally, through this symposium we illustrate how complex statistical modeling approaches such as structural equation modeling could be used to answer relevant research questions in the study of intimate partner violence and to identify key variables to increase the informal control of this type of violence.

Keywords: Assessment; Scoping review; Meta-analysis; Structural equation models; Intimate partner violence against women

E-mail: celiaserrano@ugr.es

Improving definition of police attitudes toward intervention in intimate partner violence against women in the Spanish context: A qualitative approach

Celia Serrano-Montilla[1], Luis-Manuel Lozano[1], José-Luis Padilla[1]

[1] *Department of Methodology for Behavioral Science, University of Granada, Spain*

Abstract

The aim of the present work was to apply a mixed method research approach (mainly, qualitative approach) to carry out the two first phases of the scale development to assess a social construct: police attitudes toward intervention in IPVAW. Specifically, the qualitative part aims (a) to build the semantic and syntactic definition, and (b) to generate the item pool. In addition, content validity evidence was obtained during both phases. The results of our in progress series of studies (systematic review, focus groups and expert appraisal) aimed to propose a preliminary definition of the target construct with two dimensions (reactive and proactive); four components (tolerance toward IPVAW, minimal police involvement, understanding of the complex nature of abuse, and IPVAW intervention as an important police task); and seventeen behaviors. Four types of determinants (individual, situational, organizational, and societal) enabled the construction of a nomological theoretical network. In general, content-based validity evidence was provided for both the scale specifications and the first draft of the item scale regarding domain definition, representation, relevance, and the appropriateness of test construction procedures. Implications in terms of the usefulness of the mixed method approach for the following steps are discussed.

Keywords: Qualitative data analysis; Qualitative research; Scale development; Police attitudes toward intervention in Intimate partner violence against women.

Funding: This research was made possible thanks to the financing provided by the FPU Program of the Spanish Ministry of Education, Culture, and Sport under Grant FPU16/03024

E-mail: celiaserrano@ugr.es

1. Introduction

Intimate partner violence against women (IPVAW) is recognized as the most common type of violence against women, affecting 30% of women worldwide. In this sense, research has pointed out attitudes toward IPVAW as a key variable to explain actual IPVAW prevalence across neighborhoods, communities, and countries (Sanz-Barbero et al., 2018), as well as public policy (Serrano-Montilla et al., 2020) and law enforcement responses to this social and health problem (Lila et al., 2013). That is why literature has focused on the study of police attitudes toward intervention in IPVAW, drawing on the distinction between reactive and proactive attitudes (Chu & Sun, 2014; DeJong et al., 2008). In general, reactive police attitudes refer to views of IPVAW incidents as private matters, the victims being responsible for the violence, police work on these issues as a secondary police task, etc., leading police officers to lower levels of willingness to be involved. Conversely, proactive attitudes toward intervention in IPVAW include the understanding of the importance of this police task, the complex nature of the violence, and the engagement with aggressive and empathic intervention, and being in favor of pro-arrest policies (Chu & Sun, 2014). However, there is neither a clear operational definition of police attitudes toward intervention in IPVAW nor suitable adapted measurement instruments to the Spanish cultural, legislative and linguistic context (Gracia, & Lila, 2015). The present work will show in progress studies to overcome methodological challenges related to the development of adequate self-report measures of police attitudes toward intervention in IPVAW. Therefore, we decided to develop an "item pool" intended to measure police attitudes, leading our work from the mixed methods model of scale development and validation analysis (MSDVA; Zhou, 2019), which have been adapted from Creswell and Plano Clark (2011)'s exploratory sequential instrument design. This model also introduces the validation process from the beginning (i.e., validity evidence is obtained across the process even before the item generation). The model includes five steps: (1) qualitatively investigating the scale construct by collecting content-validity evidence; (2) converting qualitative findings to scale items; (3) performing mixing validation studies to obtain content-validity evidence for the scale items; (4) administering test items and collecting item responses; and (5) conducting quantitative validation studies to analyze item psychometric characteristics. Items with poor psychometric properties should be revised and sent back to step three for another run of validation and pilot testing until they get through item analysis in step five (Zhou, 2019). The present extended summary is focused on the two first phases. The objective for the first phase was (a) to develop semantic and (b) syntactic definitions of police attitudes toward intervention (Study 1 and 2), and (c) to obtain content validity evidence for the semantic definition of such attitudes (Study 3). Regarding the second phase, the objective was to (d) develop the item pool and (e) to obtain validity evidence for the scale content (Study 4).

2. Method

2.1. First phase

In order to conceptualize police attitudes toward intervention and validate the proposed definition, we carried out three studies: a systematic review (Study 1), focus groups (Study 2), and expert appraisals (Study 3).

2.1.1. Study 1: Systematic review

Serrano-Montilla et al. (2021) followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (Moher et al., 2009) to develop the review protocol. First, a

comprehensive search through the main databases (Scopus, Web of Science Core Collection, ProQuest) was performed and results were imported to Mendeley for storage and management of references. Two authors screened the titles, abstracts, and full texts, and papers were selected by following the inclusion criteria: (a) to have provided original qualitative, quantitative, or mixed empirical findings, or reviews of empirical studies; (b) have been published between January 1990 and September 2019; (c) have been published in Spanish or English; (d) the need to include police officers in the sample; and (e) be focused on police attitudes toward intervention in IPVAW and their determinants. The intercoder reliability was 76%. After a discussion on the disagreements, 59 papers were selected but finally, only 57 empirical or review studies remained because two papers were removed after a more in-depth reading during the analysis. A coding scheme guided the collection of data through papers, focusing on obtaining information about police attitudes toward intervention in IPVAW and its determinants. Then content analysis was performed, dividing information into content units (i.e., specific attitudes and determinants) and then, frequency for police attitudes and determinants was calculated. More information can be obtained in Serrano-Montilla et al. (2021).

2.1.2. Study 2: Focus Groups

A non-probabilistic incidental and snowball sampling method was employed for recruitment. Police officers were selected if they had experience with official duties included in Spanish law, and protocols for intervention in IPVAW. To ensure that semantic definition included all nuances of police attitudes toward intervention in IPVAW, we formed focus groups based on: (a) belonging to a national or local police department (b) belonging to the IPVAW specialist unit, and (c) gender. Police officers from each focus group shared the police unit (i.e., IPVAW specialist or non-specialist role) and agency (i.e., National or Local department) but they were diverse in terms of gender (i.e., men and women shaped each group). Finally, 36 police officers (17 IPVAW specialists and 19 non-IPVAW specialists) from Local and National Spanish Law Enforcement departments, participated in six mixed gender groups: three all-IPVAW specialists, and three all non-IPVAW specialists. Following systematic review results (Serrano-Montilla et al., 2021), we developed a semi-structured protocol that the facilitator used to conduct focus groups (between 50-90 min) during police officers' working days at quiet places within the police departments. The facilitator applied for permission to audio record the sessions and informed written consents were filled out by police officers before the sessions started. At the end of the session, they were also asked to fill out a short demographic form.

Regarding data analysis, an assistant transcribed the audiotapes of the six focus groups verbatim. Then authors performed a content analysis to identify and organize themes and subthemes embedded in the police speeches through three phases (i.e., identifying content units), moving into the process of searching for a bigger content or themes, and reviewing the final coding scheme independently by each author). Once we established the coding scheme, the frequency of each code and theme appeared, identifying the prevalence and most common topics mentioned in our corpus of data. Analysis was supported by Atlas.ti *(version 7.5.18)*.

2.1.3. Study 3: Expert appraisal

A group of four academics and four police experts revised the operational definition (dimension, components, behaviors and indicators) we developed drawing from systematic review and focus groups. Through a matching task, academics and experts evaluated whether behaviors belonged to the intended components and dimensions. Academics and experts also evaluated

whether indicators were adequate to measure the intended behaviors. Content validity indices (CVIs) and Cohen's kappa coefficients (κ) were also calculated. A CVI higher than .79 indicated that the item was appropriate; between .70 and .79 that it might need revision, and values lower than .70 indicated that the item should be eliminated. Likewise, Kappa values above 0.74, between 0.60 and 0.74, and between 0.40 and 0.59 were considered excellent, good, and fair, respectively (Zamanzadeh et al., 2015).

2.2. Second phase

2.1.1. Study 4: Development of the first Spanish version of police attitudes toward intervention in an IPVAW item pool

The first pool of items was developed following DeVellis' (2012) recommendations. The criteria to select the best items for the pool were: (a) items that generate variability on responses according to participants' attitudes toward intervention, (b) those with the simplest grammatical structure, and (c) to be consistent with legal and police terminology. Likewise, measurement experts determined whether items were conformed to standard principles of quality item writing. A total of 15 experts (psychometricians and other academic psychologists) assessed the final pool of items. They had to assess items by using a rating task with a four-point Likert-type rating scale (from 1 = not at all representative to 4 = very representative).

3. Results

3.1. First phase

Systematic review (Study 1) provided a general structure or "etic" positions about police attitudes toward intervention in IPVAW. As Serrano-Montilla et al. (2021) showed, regarding the semantic definition, the reactive (vs. proactive) dimension of police officers' attitudes has received more attention and had a better conceptualization in the literature (404 reactive content units of police attitudes vs. 334 content units of police attitudes). Likewise, the components that belonged to the reactive and proactive dimensions from better to worse conceptualization were: tolerance toward intervention in IPVAW, understanding of the complex nature of abuse when intervening, minimal police involvement, view of the IPVAW intervention as an important police task, and supportive versus unsupportive attitudes toward the legal system and legislation against IPVAW. In addition, systematic review results showed that more attention has been paid to variables from an "ontogenic system", in particular, police demographic and background characteristics, professional background, attitudes toward women and gender-based violence, and police abilities and cognitions. Second, we found that microsystem-related variables were also widely studied (i.e., the situational characteristics of an IPVAW situation, the mental illness status of the perpetrator, the injury status of the victim, the type of violence and IPVAW, and other perpetrator and victim characteristics) whereas variables from the exosystem and macrosystem were underserved (more details of the systematic review results can be obtained by checking Serrano-Montilla et al., 2021)

Furthermore, focus groups (Study 2) shaped both psychometric definitions from an "emic" approach by introducing specific contents to adapt definitions to the Spanish linguistic, legislative and cultural context. Thus, content domains derived from this study (ranging from more to less frequent) were involved partners and nature of the IPVAW, self, philosophy of policing in IPVAW, resources, barriers in the case of IPVAW events, the law, and other involved professionals. After analyzing the results, the research team re-shaped the general structure of the

semantic definition (i.e., unsupportive and supportive attitudes toward the legal system and legislation against IPVAW were considered as a part of syntactic definition) and introduced behaviors and their respective indicators in order to define the four specific components for the Spanish context. For example, the minimal police involvement component included the behavior: "to believe there are barriers in IPVAW which make the intervention difficult" and within this behavior we developed several indicators that will let items be written easily (e.g., to think that involved partners make the intervention difficult). Besides, the perception of resources, and the feelings derived from the IPVAW intervention were included in the syntactic definition.

With regard to Study 3, content validity and Kappa indices indicated which behaviors and indicators had a good or poor performance from the experts' point of view. Hence, experts pointed out problems with the item pool specifications. Specifically, there was disagreement about the relationships of two behaviors with dimensions (i.e., reactive or proactive police attitudes), and eight behaviors with components (i.e., tolerance toward intervention in IPVAW, minimal police involvement, understanding of the complex nature of abuse when intervening, and IPVAW intervention as an important police task; Kappa index < .60, and CVI < .80). Based on the indices and comments of experts we re-wrote behaviors in order to avoid ambiguity, and we removed the term "intervention" for cognitive components (i.e., tolerance toward IPVAW, understanding of the complex nature of the abuse).

3.2. Second phase

In order to generate the item pool, authors introduced specific terms that police officers used during the focus groups. The first item pool was composed of 154 items. After the review, 63 items were assessed in Study 4. The results showed agreement between experts about 32 items (Kappa index > .60, and CVI > .80 and positive expert comment) were the best to evaluate police attitudes toward intervention in IPVAW.

4. Conclusions

Through four studies we have provided evidence about three steps of mixed methods approaches to test development: (1) qualitative inquiry (i.e., systematic review, focus group and expert appraisal) about the scale construct; (2) the translation of the qualitative findings toward scale specifications and items, and (3) mixing validation to review items' content-based validity. The following steps will be continuing the integration of qualitative and quantitative data during the administration of the item pools, collecting police officer responses to items, analyzing item properties and obtaining different validity evidence to support the proposed interpretation of instrument scores. Mixing both qualitative and quantitative methods is a useful framework, which could enhance the understanding of Spanish police attitudes toward intervention in IPVAW than either method by itself.

References

Chu, D. C., & Sun, I. Y. (2014). Reactive versus proactive attitudes toward domestic violence: A comparison of Taiwanese male and female police officers. *Crime & Delinquency, 60*(2), 216–237. http://doi.org/101177/0011128710372192

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research (2nd ed.)*. Sage.

DeVellis, R. F. (2012). *Scale development: Theory and application.* Sage.

Dejong, C., Burguess-Proctor, A., & Elis, L. (2008). Police officer perceptions of intimate partner violence: An analysis of observational data. *Violence and Victims, 23*, 683–697. http://doi.org/101891/0886-6708.23.6.683

Gracia, E., & Lila, M. (2015). *Attitudes towards violence against women in the EU* (Report prepared by ENEGE Network for the European Commission, DG Justice Unit D2). http://doi.org/102838/045438

Lila, M., Gracia, E., & García, F. (2013). Ambivalent sexism, empathy and law enforcement attitudes towards partner violence against women among male police officers. *Psychology, Crime and Law, 19*(10), 907–919. http://doi.org/101080/1068316X.2012.719619.

Moher, D., Liberati, A., Tetzlaff, J., & Douglas, A. G., & PRISMA Group. (2009). Reprint—Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Physical Therapy, 89*(9), 873–880. http://doi.org/101093/ptj/89.9.873

Sanz-Barbero, B., López, P., Barrio, G., & Vives-Cases, C. (2018). Intimate partner violence against young women: prevalence and associated factors in Europe. *Journal of Epidemiology and Community Health, 72,* 611–6.vhttps://doi.org/10.1136/jech-2017-209701

Serrano-Montilla, C., Valor-Segura, I., Padilla, J.L., & Lozano, L.M. (2020). Public helping reactions to intimate partner violence against women in European countries: the role of gender-related individual and macrosocial factors. *International journal of environmental Research and Public health 17*(17), 6314. https://doi.org/10.3390/ijerph17176314

Serrano-Montilla, C., Lozano, L.M., Alonso-Ferres, M., Valor-Segura, I., & Padilla, J.L. (2021). Understanding the components and determinants of police attitudes toward intervention in intimate partner violence against women: A systematic review. *Trauma, Violence, & Abuse*, 1-16. https://doi.org/10.1177/15248380211029398

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences,4*(2), 165–178. http://doi.org/10.15171/jcs.2015.01

Zhou, J. (2019). A mixed methods model of scale development and validation analysis. *Measurement Interdisciplinary Research and Perspectives, 17*(1), 38–47. https://doi.org/10.1080/15366367.2018.1479088

Social desirability in psychological aggression against a partner studies: A scoping review

M. Carmen Navarro-González[1], José-Luis Padilla[1], Carolina Díaz-Piedra[1]

[1]*Department of Methodology for Behavioral Science, University of Granada, Spain*

Abstract

Underreporting can undermine the validity of the assessment of socially unaccepted behaviors. Reducing the levels of underreporting in assessments of psychological aggressions against a partner is particularly difficult due to several factors, among others, social desirability (SD). Problems with SD definition and evaluation could make it difficult to include SD measures when assessing psychological aggression. We conducted a systematic search of the literature (scoping review) evaluating how SD has been assessed in studies on psychological aggression against a partner. A total of 391 studies that used at least one self-administered measure to assess psychological aggression against a partner in adult participants were included. Only 5.63% of studies covered assessed SD. All of them resorted to SD scales that understand this construct as a latent trait and did not take into account the assessment context. The majority of studies (63.64%) used correlations as an analytic strategy for SD data. Trying to detect SD in studies that assess psychological aggression is not frequent, despite the widespread recognition among both researchers and clinicians of the importance of underreporting. Several implications for improving the detection of SD when assessing psychological aggression against a partner are discussed.

Keywords: Social desirability; response biases; assessment; psychological aggression; intimate partner violence; scoping review.

E-mail: mcnavarro@ugr.es

1. Introduction

Underreporting is a well-known problem among researchers working on socially unaccepted behaviors, such as intimate partner violence (IPV). In fact, both researchers and clinicians working with IPV perpetrators recognize that they tend to minimize and deny their own violent behaviors against their partners (Follingstad & Rogers, 2013).

Underreporting rates of IPV could vary depending on several factors. Particularly, in the assessment of psychological aggression, one of these factors is the difficulty in validating this type of aggression with external records or proxies (Follingstad & Rogers, 2013). Another especially relevant factor that could explain underreporting levels in psychological aggression assessment is social desirability (SD) (e.g., Rosenbaum et al., 2006).

SD refers to the tendency of a person to give positive self-descriptions about him/herself when he/she is assessed (Paulhus, 2002). This response tendency would account for both unconscious and conscious attempts to safeguard one's self-image. It may be reasonable to think that a valid SD assessment could help to detect underreporting when assessing psychological aggression in IPV. However, controversies over the definition and measure of the SD construct itself (e.g., Leite & Cooper, 2010) could be a barrier to develop effective means to account for SD in psychological aggression assessment.

Regarding the operationalization of the SD construct, there are two big trends in the literature. Some researchers consider that SD is a "personality trait": each individual has a "stable" latent score in SD when responding to self-administered measures (e.g., Paulhus, 2002). On the other side, following new "contextual" approaches, new researchers propose that the effects of SD on a particular variable depend on the participant's characteristics, the nature of the "target" construct, and the contextual factors of the assessment (e.g., Krumpal, 2013; Leite & Cooper, 2010). For example, people are more likely to anticipate a higher level of risk if they admit "compromising" behaviors (Krumpal, 2013; Rosenbaum et al., 2006).

Therefore, as a first step in addressing the issue of underreporting in the assessment of psychological aggression against a partner, it becomes necessary to gain knowledge about the ways SD is currently being addressed. To achieve this goal, we explored the scientific literature regarding the assessment of psychological aggression against a partner through a scoping review. The objective of this scoping review is to gain knowledge about the way SD is being assessed, if assessed, in research on psychological aggressions against a partner. The proposed scoping review answered the following questions:

1) In what kind of studies is a SD assessment included?
2) What instruments or procedures are used to assess SD?
3) What are the psychometric properties of the SD scales used?
4) How do these instruments or procedures understand SD?
5) What analytic strategy did researchers perform with SD scores, if they performed any?

2. Method

2.1. Eligibility Criteria

We included empirical studies conducted in any context that had used, at least, one self-administered measure to assess the presence and/or the intensity of psychological aggression against the current or former romantic partner in adult participants of all genders, marital

statuses, and sexual orientations. The included studies were articles published in peer-reviewed scientific journals between January 1994 and June 2020, published in English, Spanish or Portuguese.

2.2. Information Sources

The search was executed in five databases: Scopus, PsycINFO, and ProQuest Psychology, Social Science, and Nursing & Allied Health databases. The most recent search was executed on 06/26/20.

2.3. Search

The search keywords were selected according to the American Psychological Association (APA) Thesaurus of Psychological Index Terms and the UNESCO Thesaurus. Then, the search strategy was adapted to the particularities of each database. The general search strategy was as follows:

(((“psychological violence” OR “psychological abuse” OR “psychological aggression” OR maltreat*) AND (partner OR “intimate partner” OR conjugal OR “gender-based violence” OR “gender violence”)) AND NOT (infan* OR child* OR adolesc*))

2.4. Selection of Sources of Evidence

After the search strategy was implemented, the citations were imported to Mendeley, where a first removal of duplicates (n = 925) was conducted. Then, the citations were imported to a free online software to help conduct systematic reviews: Rayyan QCRI. A second removal of duplicates was conducted with this software (n = 41).

To assess documents for eligibility, two reviewers independently first executed a title and abstract screening of the retrieved abstracts. The Cohen’s kappa between the two reviewers was .68, indicating substantial agreement (Landis & Koch, 1977). Then, a full-text screening of the included records was executed independently by two reviewers to decide whether articles were definitely included in the review. The Cohen’s kappa between the two reviewers was .88, indicating excellent agreement (Landis & Koch, 1977). In both screenings, the conflicts were solved by a third reviewer.

2.5. Data Charting Process

The data charting process was carried out using the software NVivo. A predefined list of categories for extraction was developed by the research team and implemented in the software. NVivo allowed a qualitative data extraction, in which the retrieved information of each study was used to categorize the study, following the predefined list of categories for extraction. This process was complemented with a quantitative data extraction conducted in SPSS Statistics 22.

2.6. Data Items

We extracted data on bibliometric indicators, methodological aspects of the studies, SD assessments, properties of SD scales used, the conceptualization of SD, and uses of SD assessments.

3. Results

3.1. Selection of Sources of Evidence

We identified 1966 records and, after duplicates were removed, we screened 1000 records. Based on the title and abstract screening, we excluded 343 records. Full texts were not found for 7 of the 657 remaining records. Therefore, we assessed 650 full-text documents for eligibility. Of these records, 271 were excluded. A total of 379 articles were finally included in the review. Some of those articles reported more than one study (n = 11). Thus, we reviewed 391 studies.

3.2. Frequency and Percentage of Studies that Assessed Social Desirability

Overall, only 22 of 391 (5.63%) studies included an SD assessment. In addition, there were 26 (6.65%) additional studies that did not assess SD but tried to control it in some way. The strategy used by almost all of those studies (25 of 26) consisted in either using a partner's report of the received psychological aggression perpetrated by his/her partner as an indicator of perpetration of the other member of the couple, or only choosing the higher of the partner's reports in psychological aggression scales to calculate the final score of "perpetration" of each member of the couple. The other study controlled SD by asking for the testimony of the caseworkers that followed the participants throughout the study.

3.3. Characteristics of Social Desirability Measurement Instruments

Of the 22 studies that assessed SD, 20 studies (90.91%) used a scale to assess the construct. The other two studies (9.09%) reported in the same article, used "two questions" which were not specified. The three main SD scales used were the Balanced Inventory of Desirable Responding (BIDR, used in 9 studies) (Paulhus, 1988), the Marlowe-Crowne Social Desirability Scale (M-C SDS, used in 3 studies) (Crowne & Marlowe, 1960), and the Social Desirability Scale of the Personal and Relationships Profile (PRP, used in 8 studies) (Straus et al., 1999), which is a version of the Reynolds' (1982) short form of the M-C SDS. All those scales understand SD as a personality trait.

3.4. Psychometric Properties of Social Desirability Scales Used

In general, the psychometric properties provided for the SD scales used were good: Cronbach's alphas ranged from .68 to .87 for the BIDR, from .73 to .88 for the M-C SDS, and from .68 to .72 for the Social Desirability Scale of the PRP. Of the 20 studies that assessed SD with a scale, 10 studies (50%) reported results from other studies that conducted a psychometric analysis and/or conducted a psychometric analysis themselves. Of these, two studies (10%) only reported results from other studies, without conducting any for the study sample. The remaining 10 studies (50%) gave no information about the psychometric properties of the SD scales used.

3.5. Use of Social Desirability Assessments (Analytic Strategies)

Of the 22 studies that assessed SD, five studies (22.73%) did not perform any analytic strategy with the SD data, or did not give any information. Of the 17 remaining studies (77.23%), some of them performed more than one analytic strategy. Ten studies (45.45%) explored the impact of SD in their variables of interest or controlled its effects using regression models. Six studies (27.27%) established group differences in SD. Two studies (9.09%) performed other analytic

strategies such as including SD as a covariate in an analysis of covariance. Twelve studies (63.64%) used correlations between SD and their variables of interest.

4. Conclusions

We reviewed 391 studies about psychological aggression against a partner and only 5.63% of them did assess SD. Furthermore, only in 12.28% of those studies was there a concern about SD and an attempt to, at least, account for it. This means that trying to detect or to control in some way SD in studies that assess psychological aggression against a partner is not a frequent practice or considered relevant. Furthermore, only half of the studies that assessed SD with a scale did report some information about the psychometric properties of the SD measures they had obtained. In addition, two of those studies only reported results from other studies but did not conduct any assessment of psychometric properties by sample studies.

These findings are consistent with some experts' opinions (e.g., Follingstad & Rogers, 2013): in spite of the widespread recognition of the importance of SD for assessing IPV behaviors, researchers have not paid enough attention to the issue when conducting studies aimed as assessing IPV behaviors. This almost paradoxical impression leaves us with more questions that could open a new path for future research. For example, it would be interesting to explore the main reasons why researchers do not pay that much attention to SD assessments in psychological aggression against a partner studies. In addition, according to the "ecological model" of item responding proposed by Zumbo (Chen & Zumbo, 2011), the contextual factors within and outside the test settings need to be addressed to obtain valid inferences from test scores. It becomes necessary, then, to conduct an evaluation of the psychometric properties each time a particular scale is used, because validity evidence can differ across samples and contexts.

On the other hand, all studies that assessed SD used procedures that understand this construct as a personality trait, and do not take into account the items' content or the assessment context. These scales were created decades ago, and might have become outdated, not including the more recent "contextual" approaches (e.g., Leite & Cooper, 2010). Given that people anticipate a higher level of risk when answering "compromising" items (Krumpal, 2012; Rosenbaum et al., 2006) and consequently, are more prone to underreporting, we might be confronting a validity issue here, That is, the most popular SD scales might be assessing only one component of the SD construct (the personal tendencies or participant's characteristics).

The relevance of this scoping review resides in trying to raise awareness about the issue of underreporting in assessments on psychological aggressions against a partner, and the lack of importance given to the assessment context when detecting SD. This review represents a first step in trying to address those problems in assessments dealing with psychological aggressions against a partner, in order to eventually be more precise when detecting cases of IPV.

References

Chen, M. Y., & Zumbo, B. D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA Data. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 53–68). Springer International.

Crowne, D.P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24(4)*, 349–354.

Follingstad, D.R., & Rogers, M.J. (2013). Validity Concerns in the Measurement of Women's and Men's Report of Intimate Partner Violence. *Sex Roles, 69,* 149–167.

Krumpal, I. (2013). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality and Quantity: International Journal of Methodology, 47(4),* 2025–2047.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorial data. *Biometrics, 33,* 159–74.

Leite, W.L., & Cooper, L.A. (2010). Detecting Social Desirability Bias Using Factor Mixture Models. *Multivariate Behavioral Research, 45*, 271–293.

Paulhus, D.L. (1988). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding.* University of British Columbia.

Paulhus, D.L. (2002). Socially desirable responding: The evolution of a construct. In H.I. Braun, D.N. Jackson, & D.E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp. 49-69). Lawrence Erlbaum Associates, Inc.

Reynolds, W. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 38(1),* 119–125.

Rosenbaum, A., Rabenhorst, M.M., Reddy, M.K., Fleming, M.T., & Howells, N.L. (2006). A Comparison of Methods for Collecting Self-Report Data on Sensitive Topics. *Violence and Victims, 21(4),* 461–471.

Straus, M.A., Hamby, S.L., Boney-McCoy, S., & Sugarman, D. (1999). *The Personal and Relationships Profile (PRP).* Family Research Laboratory, University of New Hampshire.

Advances in the effectiveness of intervention programs for intimate partner violence offenders: A systematic review and meta-analysis of RCTs

Faraj A. Santirso[1], Marisol Lila[1], Enrique Gracia[1]

[1]*Department of Social Psychology, University of Valencia, Spain*

Abstract

Systematic reviews and meta-analyses are essential to support the development of clinical practice guidelines and inform clinical decision-making. In this paper, we conducted a systematic review with meta-analysis to analyze whether the inclusion of motivational strategies improves the effectiveness of intimate partner violence (IPV) offender programs based only on Randomized Controlled Trials (RCTs), following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses. The methodological quality of the trials was assessed according to the Cochrane Risk of Bias tool. The main summary measures were the standardized mean difference and odds ratio, and the degree of heterogeneity (I^2) was calculated. Data entry and statistical analysis were carried out using Review Manager Software. Results showed that participants in the motivational IPV offender intervention programs showed a non-statistically significant reduction in physical and psychological IPV and official recidivism. Also, they received a significantly higher dose of intervention and showed significantly less dropout. Our findings showed the importance of conducting systematic reviews and meta-analysis studies to support the development of clinical practice guidelines of IPV offender programs based on scientific evidence. These methods are suitable for determining whether scientific evidence can be generalized across treatment variations, subsamples or settings.

Keywords: Meta-analysis; experimental designs; publication bias; intimate partner violence offender programs.

Funding: This study has been supported by the Spanish Ministry of Health, Consumption and Social Services, National Drugs Plan (PND2018/021). Faraj A. Santirso was supported by the FPU Program of the Spanish Ministry of Science, Innovation and Universities (grant number FPU15/00864).

E-mail: faraj.santirso@uv.es

1. Introduction

Systematic reviews and meta-analyses are essential to support the development of clinical practice guidelines and inform clinical decision-making (Moher et al., 2015; Gopalakrishnan & Ganeshkumar, 2013). In the framework of intimate partner violence (IPV) prevention, offender intervention programs are one of the most widespread prevention strategies (Voith et al., 2018). It has been suggested that the effectiveness of these programs could improve with the inclusion of motivational strategies (e.g., motivational interviewing, retention techniques, stages-of-change-based interventions) (Babcock et al., 2016). Randomized Controlled Trials (RCTs) are considered the gold standard to compare different interventions, enhancing control over bias such as regression to the mean or spontaneous remission (Lilienfeld et al., 2014, 2018). Taking all it together, we conducted a systematic review with meta-analysis to analyze whether the inclusion of motivational strategies improves the effectiveness of IPV offender programs based only on Randomized Controlled Trials (RCTs).

2. Method

2.1. Systematic review

We carried out a systematic review and meta-analysis following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations (Moher et al., 2009). Before the search, a study protocol was submitted through PROSPERO, an international prospective register for review protocols (registration CRD42018110107). We used the following databases: Cochrane Collaboration, MEDLINE, EMBASE, PsycINFO, and CINAHL. Documents were imported in a reference manager software (Endnote Version X9). The search strategy included combining terms for intervention programs for IPV offenders, motivational strategies, and randomized controlled trials. The decision to include only RCTs was taken to prevent possible confounding factors, as well as to increase the precision and replicability of the results. For identifying RCTs, the Cochrane Highly Sensitive Search Strategy (Lefebvre et al., 2011) was used. The Cochrane Risk of Bias tool (Higgins et al., 2011) was used to evaluate the methodological quality of the trials.

2.2. Meta-analysis

Data entry and statistical analysis were carried out using Review Manager Software, following the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al., 2019). Both the software and the handbook were available for academic use on the Cochrane website.

We used the random-effect model, in which the summary effect of all studies included in the meta-analysis is an estimate of the mean of a distribution of true effects. We used this model since we assumed that the observed differences between study results were due to both the effects of the intervention and the effect of chance. Additionally, following the Cochrane Handbook recommendations, possible differences between the fixed-effect model and the random-effect model were checked (Higgins et al., 2019).

Regarding the statistical method used for meta-analysis, we used the DerSimonian and Laird method (1986), since it is adequate for random-effect models. In this method, the weighting of each study is set as the inverse of the variance of the effect estimate, so that studies with larger samples (with smaller standard errors) have more weight than studies with smaller samples (and larger standard errors).

Continuous and dichotomous outcomes were analyzed. As continuous outcomes, we considered physical and psychological intimate partner violence, and intervention dose. For these outcomes, the main summary measure was the standardized mean difference (SMD). As dichotomous outcomes, dropout and official recidivism were considered, and the odds ratio (OR) was computed as a summary measure. The degree of heterogeneity (I2) was computed to examine whether studies included in the meta-analysis were consistent. The following cut-offs were considered to interpret heterogeneity: I2 of 25%: low; I2 of 50%: moderate; and I2 of 75%: high (Higgins et al., 2003).

3. Results

3.1. Methodological quality of the trials

The risk of bias for included studies is shown in Table 1.

Table 1. Risk of bias for included studies

Study	Q1	Q2	Q3	Q4	Q5	Q6
Alexander et al. (2010)	H	U	H	L	U	H
Bahia (2016)	L	U	H	H	U	U
Chermack et al. (2017)	U	U	U	U	L	U
Crane & Eckhardt et al. (2013)	L	U	U	L	L	L
Kraanen et al. (2013)	L	L	U	U	L	U
Lila et al. (2018)	L	U	U	U	L	L
Mbilinyi et al. (2011)	L	U	U	U	U	L
Murphy et al. (2018)	L	U	U	U	L	L
Murphy et al. (2011)	L	U	H	L	L	U
Schumacher et al. (2011)	L	U	U	U	U	L
Stuart et al. (2013)	L	L	U	U	L	L
Wooding and O'Leary (2010)	U	U	H	H	L	H
Percentage of low risk criteria	75%	17%	0%	25%	67%	50%

Note. H = high risk; L = low risk; U = unclear risk; Q1 = random sequence generation (selection bias); Q2 = allocation concealment (selection bias); Q3 = blinding of participants and personnel (performance bias); Q4 = blinding of outcome assessment (detection bias); Q5 = incomplete outcome data (attrition bias); Q6 = selective reporting (reporting bias).

3.2. Meta-analysis results

Regarding physical and psychological IPV, participants in the motivational IPV offender intervention programs showed a non-statistically significant reduction (SMD = 0.08, 95% CI [-0.09,0.25]; and SMD = 0.09, 95% CI [-0.21, 0.38], respectively). Heterogeneity was low for physical IPV (I2 = 0%), whereas it was moderate for psychological IPV (I2 = 53%).

Additionally, participants in the motivational IPV offender intervention programs received a significantly higher dose of intervention and showed significantly lower dropout rates than

those of the interventions without motivational strategies (SMD = 0.27, 95% CI [0.08, 0.45]; and OR = 1.73, 95% CI [1.04, 2.89], respectively). Heterogeneity was low for both outcomes (I2 = 0%).

Finally, participants in the motivational IPV offender intervention programs showed a non-statistically significant reduction in official recidivism (OR = 1.46, 95% CI [0.76, 2.80]). Heterogeneity was low (I2 = 33%).

4. Conclusions

This study showed the importance of conducting systematic reviews and meta-analysis studies to support the development of clinical practice guidelines for IPV offender programs based on scientific evidence. These methods are suitable for determining whether scientific evidence can be generalized across treatment variations, subsamples or settings.

Some methodological considerations should be underlined. The follow-up start points of interventions should be accurately reported. In this regard, the use of post-intervention as a reference point may improve the comparability of results. Additionally, follow-up should be long enough to assess the persistence of change. When examining dropouts, a clear definition is needed (i.e., the number of participants who leave the intervention before it ends instead of a pre-defined participation percentage). Finally, data triangulation (i.e., using data from perpetrators, police records and partners or ex-partners) could be useful to improve the validity of IPV offender program outcomes.

References

Abcock, J., Armenti, N., Cannon, C., Lauve-Moon, K., Buttell, F., Ferreira, R., Cantos, A., Hamel, J., Kelly, D., Jordan, C., Lehmann, P., Leisring, P. A., Murphy, C., O'Leary, K. D., Bannon, S., Salis, K. L., & Solano, I. (2016). Domestic violence perpetrator programs: A proposal for evidencebased standards in the united states. *Partner Abuse, 7*(4), 355–460. https://doi.org/10.1891/1946-6560.7.4.355

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Gopalakrishnan, S., & Ganeshkumar, P. (2013). Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *Journal of Family Medicine and Primary Care, 2*(1), 9–14. https://doi.org/10.4103/2249-4863.109934

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Lefebvre, C., Manheimer E., & Glanville, J. (2011). Chapter 6: Searching for studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions (Version 5.1.0, updated March 2011)* (pp. 103–148). Cochrane. https://training.cochrane.org/handbook

Lilienfeld, S. O., McKay, D., & Hollon, S. D. (2018). Why randomized controlled trials of psychological treatments are still essential. *The Lancet Psychiatry, 5*(7), 536–538. https://doi.org/10.1016/S2215-0366(18)30045-2

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspectives on Psychological Science, 9*(4), 355–387. https://doi.org/10.1177/1745691614535216

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Stewart, L. A, & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1–9. https://doi.org/10.1186/2046-4053-4-1

Voith, L. A., Logan-Greene, P., Strodthoff, T., & Bender, A. E. (2018). A paradigm shift in batterer intervention programming: A need to address unresolved trauma. *Trauma, Violence, & Abuse, 19*(1)1–15. https://doi.org/10.1177/1524838018791268

Predicting the willingness to intervene in cases of intimate partner violence: A mediation analysis

Manuel Martín-Fernández[1], Miriam Marco[2], Arabella Castro[2], Enrique Gracia[2] & Marisol Lila[2]

[1]*Department of Social Psychology and Methodology, Autonomous University of Madrid, Spain,*

[2]*Department of Social Psychology, University of Valencia, Spain*

Abstract

Intimate partner violence against women (IPVAW) is a serious public health problem of global proportions. The willingness to intervene in cases of IPVAW reflects the level of tolerance of this type of violence. Previous research has showed a strong relationship between the willingness to intervene and several sociodemographic variables (e.g., gender, age, study level, and nationality). However, these relationships may be mediated by the attitudes towards IPVAW (i.e., attitudes of acceptability, victim-blaming attitudes, ambivalent sexism). The aim of this study was to assess the effect of public attitudes towards IPVAW in the willingness to intervene in cases of this type of violence. A sample of 1,000 participants was collected and their levels of public attitudes towards IPVAW and willingness to intervene in cases of IPVAW were assessed. Results showed that the attitudes of acceptability, victim-blaming, and ambivalent sexism could be modeled as a second order factor. A complete mediation SEM model was estimated, in which the relation between the sociodemographic variables and willingness to intervene was completely mediated by the second-order attitudinal factor. This study emphasizes the importance of using SEM models to predict the willingness to intervene in cases of IPVAW, and the key role of the attitudes towards this type of violence.

Keywords: Mediation, SEM, Attitudes, Willingness to intervene, Intimate partner violence

Funding: This study has been supported by the Spanish Ministry of Science (PSI2017-84764-P).

E-mail: manuel.martin@uv.es

1. Introduction

Intimate partner violence against women (IPVAW) is a serious public health problem and a major social issue (World Health Organization, 2021). It is considered the form of violence most commonly suffered by women, with severe consequences for their physical and psychological well-being (Ellsberg et al., 2008; Vilariño et al., 2018).

Public attitudes toward IPVAW are a risk factor at a macrosocial level, as they can encourage or deter its occurrence in societies (Heise, 2011; Powell & Webster, 2018; Waltermaurer, 2012). Previous research has acknowledged the relevance of public attitudes toward IPVAW, as they are closely related to reporting and incidence rates, help-seeking behaviors of the victims, and professional and law enforcement responses (Rizo & Macy, 2011). These attitudes have a well-documented relationship with several sociodemographic variables, such as gender, age, educational level, and immigrant status (Heise, 2011). In particular, higher levels of acceptability of IPVAW, victim-blaming attitudes, and hostile sexism has been found among men in comparison to women, in people with lower educational levels, and among immigrants in comparison to people with Spanish nationality (Martín-Fernández et al., 2018, 2021). Age has also been regarded as a potential risk factor of IPVAW, as older people tend to present higher levels of tolerance of this type of violence (Flood & Pease, 2009).

There is also a close link between these sociodemographic correlates of IPVAW and the willingness to intervene in cases of IPVAW. In particular, male gender, an older age, having a low educational level, and immigrant status have been related to lower levels of willingness to intervene in cases of this type of violence (Gracia et al., 2018).

There is, however, a gap in the literature as no studies have assessed the effect of these sociodemographic variables (e.g., gender, age, educational level, and immigrant status) in the willingness to intervene in cases of IPVAW controlling the effect of public attitudes towards IPVAW.

The aim of this study is, hence, to assess the potential mediation of public attitudes towards IPVAW in the relationship between the willingness to intervene in cases of IPVAW sociodemographic correlates of this type of violence.

2. Method

2.1. Participants

A two-stage clustered stratified sampling design was followed, sampling individuals from the different census block groups in each neighborhood of Valencia. A total pool of 4656 responses were collected. A subset of 1000 participants were randomly drawn from the pool of responses, filtering by the sociodemographic characteristics of the population of the city of Valencia (Table 1).

Table 1. Sociodemographic characteristics of the sample

	N	%
Gender		
Male	499	49.9
Female	501	50.1
Age		
18-24	149	14.9

	N	%
25-34	132	13.2
35-44	158	15.8
45-54	181	18.1
55-64	170	17.0
65+	210	21.0
M (SD)	47.19 (17.56)	
Educational Level		
Primary (<10 years of formal education)	102	10.2
Secondary (10 years of formal education)	64	6.4
Upper secondary (12 years of formal education)	56	5.6
Technical (13-14 years of formal education)	214	21.4
Graduate (13-15 years of formal education)	300	30.0
Postgraduate (>14 years of formal education)	140	14.0
Immigrant Status		
Immigrant	103	10.3
Spanish Nationality	897	89.7

2.2. Variables

Attitudes of the acceptability of IPVAW (A) (Martín-Fernández et al., 2021). This measure was an 8-item scale assessing the acceptability of IPVAW (e.g. "It is acceptable for a man to shout at his partner if she is constantly nagging/arguing"). The response format was a 3-point Likert-type scale (1 = "Not acceptable at all", 3 = "Acceptable"). The internal consistency of the scale was fair (McDonald's $\omega_{total} = 0.68$).

Victim-blaming attitudes in cases of IPVAW (VB) (Martín-Fernández et al., 2018). This instrument was a 5-item scale measuring attitudes of victim-blaming (e.g., "Men are violent towards their partners because they make them jealous"). Participants had to rate the items on a 5-point Likert-type scale (1 = "Strongly disagree", 5 = "Strongly agree"). The internal consistency of the scale was good ($\omega_{total} = 0.79$).

Hostile sexism (Expósito, Moya & Glick, 1998; Rollero et al., 2014). The short form of the hostile sexism subscale of the ambivalent sexism inventory was used for this study. The scale included 6 items assessing hostile sexism (e.g., "Women are too easily offended"), with a 6-point Likert-type scale response format (1 = "Completely disagree", 6 = "Completely agree"). The internal consistency of the scale was very good ($\omega_{total} = 0.86$).

Willingness to intervene in cases of IPVAW (WI) (Gracia et al, 2018). This scale comprises 9 items assessing willingness to intervene. It measures three specific factors —calling the police, not my business, direct intervention— and one general factor of willingness to intervene (e.g., "If I heard a man shouting violently at his partner in the communal area of my building, I would intervene to stop the situation"). The response format of the items was a 6-point Likert-type scale (1 = "Very unlikely", 6 = "Extremely likely"). The internal consistency of the specific

factors and the general factor was adequate ($\omega_{\text{calling police}} = 0.79$, $\omega_{\text{not my business}} = 0.66$, $\omega_{\text{direct intervention}} = 0.66$, $\omega_{\text{willingness to intervene}} = 0.84$).

Socio-demographic variables. Sociodemographic information about participants' gender (0 = female, 1 = male), age (mean centered for the analyses), educational level (1 = Primary, 2 = Secondary, 3 = Upper Secondary, 4 = Technical, 5 = Graduate, 6 = Post-graduate), and immigrant status (0 = Spanish nationality, 1 = immigrant) were collected.

2.3. Data Analysis

A second order factorial model (Att) was first estimated including attitudes of acceptability (A), victim blaming (B), and hostile sexism. A bi-factor model was estimated for the willingness to intervene in cases of IPVAW (WI). Weighted least squares with mean and variances adjusted (WLSMV) was used as the estimation method. Model fit was evaluated with a combination of fit indices (CFI, TLI, RMSEA). A complete mediation SEM model was estimated afterwards, in which the relation between the sociodemographic variables (i.e., gender, age, educational level, and immigrant status) and the scores of the willingness to intervene scale was mediated by the second-order attitudinal factor. To obtain the confidence intervals, a bootstrap of 10,000 iterations was conducted. The completely standardized indirect effect was computed as a size-effect measure (MacKinon, 2008).

3. Results

3.1. Measurement Model

The measurement model showed a good fit (CFI = 0.959, TLI = 0.954, RMSEA [95%CI] = 0.037[0.34, 0.040]), indicating that a second-order attitudinal factor could be posited to model the relationship between the attitudes of acceptability, victim-blaming in cases of IPVAW, and hostile sexism. The standardized loadings of all items were above 0.60 for the acceptability of IPVAW, victim-blaming, and hostile sexism, pointing out a strong relationship between the items and their constructs. The standardized loadings of the first-order factors were also above 0.60 which, in turn, indicated that these three constructs contributed substantially to the second-order attitudinal factor (Figure 1).

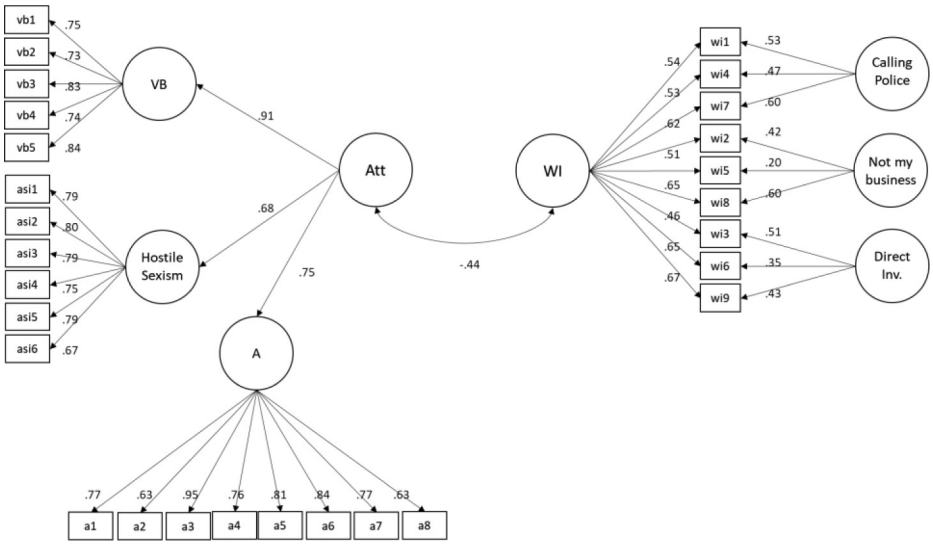


Figure 1. Measurement Model

Regarding the willingness to intervene in cases of IPVAW, the standardized loadings of the specific factors were above 0.30 for all items, except for item 5 which was 0.20. The loadings of the general factor were moderate, above 0.40 for all items. The correlation between the second-order attitudinal factor with the general factor of willingness to intervene was -0.44, pointing out that participants with higher levels of acceptability of IPVAW, victim-blaming attitudes, and hostile sexism tend to present lower levels of willingness to intervene in cases of IPVAW.

3.2. Mediation Model

The mediation model also showed a good fit (CFI = 0.956, TLI = 0.951, RMSEA[95%CI] = 0.034[0.31, 0.37]). We found a negative path between the second-order attitudinal factor and the willingness to intervene in IPVAW cases, indicating that as the scores in the attitudinal factor increased, the levels of willingness to intervene decreased. In particular, the factor scores in the willingness to intervene decreased -0.49 (Figure 2) for each standard deviation increase in the attitudinal factor.

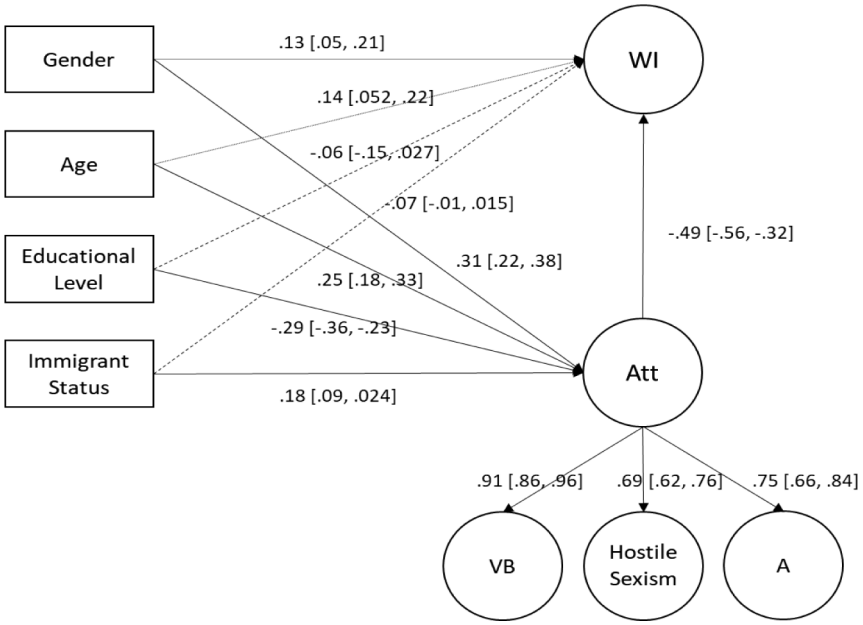


Figure 2. Mediation Model
Note: the 95% confidence intervals of the parameters are presented in square brackets

Complete mediation was found in the relationship between the socio-demographic variables and the willingness to intervene when the attitudes towards IPVAW were taken into account in the analysis (Table 2). In particular, the total effect of gender, age, educational level and immigrant status were not significant, whereas the indirect effect was significant in all cases.

Table 2. Standardized Total and Indirect Effects

	Total effect (95% CI) c = c' + ab	**Indirect Effect (95% CI) ab**	**Completely Standardized Indirect Effect (ab_{cs})**
Gender	−0.01 (−0.08, 0.06)	−0.14 (−0.20, −0.10)	−0.08 (−0.12, −0.05)
Age	0.03 (−.005, 0.10)	−0.11 (−0.17, −0.07)	−0.23 (−0.35, −0.14)

	Total effect (95% CI) c = c' + ab	**Indirect Effect (95% CI) ab**	**Completely Standardized Indirect Effect (ab_{cs})**
Educational Level	0.07 (−0.01, 0.16)	0.13 (0.08, 0.18)	0.03 (0.02, 0.04)
Immigrant Status	−0.01 (−0.09, 0.06)	−0.08 (−0.12, −0.04)	−0.03 (−0.04, −0.01)

However, the effect size of this mediation, measured as the completely standardized indirect effect (MacKinon, 2008), was almost negligible for educational level and immigrant status. A small effect size was found for gender, indicating that the effect of the second-order attitudinal factor in the willingness to intervene in cases of IPVAW was, on average, 0.08 standard deviations lower for men than for women. A moderate size effect was also found for age, pointing out that for each year above the average age, the willingness to intervene decreased 0.23 standard deviations indirectly via the second-order attitudinal factor.

4. Conclusions

This study emphasizes the importance of using SEM models to predict the willingness to intervene in cases of IPVAW, and the key role of the attitudes towards this type of violence, as they mediate the relation found in previous studies between the willingness to intervene and its socio-demographic correlates.

The results of this study, however, presented some limitations. First, the cross-sectional design of the study impeded an assessment of how the mediation found in our results may change over time. Second, our design was correlational and no manipulation was made. Third, the total effect of the sociodemographic variables was not significant (Baron & Kenny, 1986). Although some authors argue that the indirect effect is sufficient to posit mediation (MacKinnon, 2008; Hayes, Preacher, & Myers, 2011), this issue is still open to debate.

Attitudes towards IPVAW can be shaped through intervention programs and public policies in order to increase individuals' willingness to intervene in cases of IPVAW, and hence increase the informal control of this type of violence in order to eradicate it.

References

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51(6)*, 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Ellsberg, M., Jansen, H. A., Heise, L., Watts, C. H., & Garcia-Moreno, C. (2008). Intimate partner violence and women's physical and mental health in the WHO multi-country study on women's health and domestic violence: an observational study. The *Lancet, 371(9619),* 1165–1172. https://doi.org/10.1016/S0140-6736(08)60522-X

Expósito, F., Moya, M. C., & Glick, P. (1998). Sexismo ambivalente: Medición y correlatos. [Ambivalent sexism: Measurement and correlates]. *Revista de Psicología Social, 13(2),* 159–169. https://doi.org/10.1174/021347498760350641

Flood, M., & Pease, B. (2009). Factors influencing attitudes to violence against women. Trauma, Violence, & Abuse, 10(2), 125–142. https://doi.org/10.1177/1524838009334131

Gracia, E., Martín-Fernández, M., Marco, M., Santirso, F. A., Vargas, V., & Lila, M. (2018). The Willingness to Intervene in Cases of Intimate Partner Violence Against Women (WI-IPVAW) scale: Development and validation of the long and short versions. *Frontiers in Psychology, 9,* 1146. https://doi.org/10.3389/fpsyg.2018.01146

Hayes, A. F., Preacher, K. J., & Myers, T. A. (2011). Mediation and the estimation of indirect effects in political communication research. *Sourcebook for political communication research: Methods, measures, and analytical techniques, 23(1),* 434–65.

Heise, L. (2011). What works to prevent partner violence? An evidence overview. STRIVE.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Lawrence Erlbaum Associates, Inc

Martín-Fernández, M., Gracia, E., & Lila, M. (2018). Assessing victim-blaming attitudes in cases of intimate partner violence against women: Development and validation of the VB-IPVAW scale. *Psychosocial Intervention, 27(3)*, 133–143. https://doi.org/10.5093/pi2018a18

Martín-Fernández, M., Gracia, E., & Lila, M. (2021). A short measure of acceptability of intimate partner violence against women: development and validation of the A-IPVAW-8 Scale. *Assessment,* ahead of print. https://doi.org/10731911211000110

Powell, A., & Webster, K. (2018). Cultures of gendered violence: An integrative review of measures of attitudinal support for violence against women. *Australian & New Zealand Journal of Criminology, 51(1),* 40–57. https://doi.org/10.1177/0004865816675669

Rizo, C. F., & Macy, R. J. (2011). Help seeking and barriers of Hispanic partner violence survivors: A systematic review of the literature. *Aggression and Violent Behavior, 16(3),* 250–264. https://doi.org/10.1016/j.avb.2011.03.004

Rollero, C., Glick, P., & Tartaglia, S. (2014). Psychometric properties of short versions of the Ambivalent Sexism Inventory and Ambivalence Toward Men Inventory. *TPM: Testing, Psychometrics, Methodology in Applied Psychology, 21,* 149–159. https://doi.org/10.4473/TPM21.2.3

Vilariño, M., Amado, B. G., Vázquez, M. J., & Arce, R. (2018). Psychological harm in women victims of intimate partner violence: Epidemiology and quantification of injury in mental health markers. *Psychosocial Intervention*, *27*(3), 145–152. https://doi.org/10.5093/pi2018a23

World Health Organization. (2021). *Violence against women prevalence estimates, 2018. Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence*. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/341337/9789240022256-eng.pdf

Psychometrics and orectic variables

Eduardo García-Cueto[1], Marcelino Cuesta[1]

[1]*Department of Psychology, University of Oviedo, Spain*

1. State of the art

Nobody denies the importance of cognitive factors in many areas of psychology. However, what cannot be explained by them? This question is quite relevant in the fields of academic performance, addiction and mental health, among others. In recent years, researchers have become more and more interested in the study of orectic variables. They represent a wide range of attitudinal, behavioral, emotional, motivational, and psychosocial skills and dispositions, and they have received different names, such as soft skills, non-cognitive variables, and 21st-century skills. Even though the assessment of such abilities and attributes is complex, researchers are placing special emphasis on them as it is believed that they greatly influence many areas of everyday life. One such area is academic excellence. One example of this contemporary outlook on academic performance is found in the importance given to orectic variables by large international organizations such as the Teaching of 21st-Century Skills Consortium, The University of Chicago Consortium on Chicago School Research, and the Collaborative for Academic Social and Emotional Learning, to name but a few. However, the relevance given to orectic variables is not deserving only because of the attention they receive but it is also supported by data. For example, it has been shown that learning programs focused on orectic variables improve not only students' social, emotional, and academic areas, but their wellbeing as well.

2. New perspectives and contributions

A very common practice in non-cognitive research is the use of self-report questionnaires. Given the issues brought by such tools (for example, various response biases), the demand for valid and reliable questionnaires is ever more present. Nowadays, tools for measuring variables such as gender roles, experiential avoidance, the concern for appearance on social network and grit, among others, need to be developed and validated in Spain so as to better research and understand the complexities of non-cognitive variables and their impact in the general and clinical populations. Thus, the aim of this chapter is to show the development and validation of different tools aimed at different orectic variables.

3. Research and practical implications

These new tools can be used in practical and research contexts. The measurement of these variables through the new Spanish instruments will enable them to be analyzed in those contexts in which they have an impact, such as academic, clinical, organizational and social contexts.

Keywords: Psychometric properties, assessment, orectic variables.

E-mails: cueto@uniovi.es; mcuesta@uniovi.es

Development and initial validation of Oviedo Grit Scale

Álvaro Postigo[1], Álvaro Menéndez-Aller[2], Jaime García-Fernández[1], Marcelino Cuesta[1]

[1]*Department of Psychology, University of Oviedo, Spain,*

[2]*Health Research Institute of Asturias, Spain*

Abstract

Grit is one of the non-cognitive variables that has received considerable attention in recent years given its relationship to and influence in various aspects of life. There are very few reliable, valid instruments to evaluate it in Spanish-speaking countries. Accordingly, the aim of this study is the development and initial validation of a new scale to evaluate grit in Spanish-speaking contexts. We used a sample of 222 Spanish participants from the general population. The members of the sample were aged between 18 and 83 years old, with a mean age of 38.60 and a standard deviation of 14.90 years. We carried out an Exploratory Factor Analysis to confirm the unidimensional structure of the instrument. We calculated the instrument's reliability. The factorial analyses confirmed the unidimensionality of the instrument. The new grit scale demonstrated excellent reliability from the classical perspective. The new scale for evaluating grit (Oviedo Grit Scale) is essentially unidimensional, and its scores exhibit excellent indicators of reliability and validity.

Keywords: Grit, assessment, scale, dimensionality, initial validation, reliability

Funding: This study has been supported by a predoctoral grant from the Principality of Asturias (BP17-78)

E-mail: alvaro.postigo.gtz@gmail.com

1. Introduction

Grit has been a subject of much attention in the literature (Fernández-Martín et al., 2020; Postigo et al., 2021a, 2021b) since the well-known study by Duckworth et al. (2007), in which the authors defined it as follows: "Grit entails working strenuously toward challenges, maintaining effort and interest over years despite failure, adversity, and plateaus in progress. The gritty individual approaches achievement as a marathon; his or her advantage is stamina" (Duckworth et al., 2007, p. 1087). For this reason, grit is considered a positive trait based on an individual's perseverance combined with their passion for reaching a long-term goal.

Despite the research on grit in recent years, there is no consensus about its evaluation or measurement. The first instrument proposed for measuring grit was the Grit Scale (Duckworth et al., 2007), in which Duckworth and Quinn (2009) developed a short version (Grit-S). From that point on, most researchers interested in the construct used this scale, which has been validated in many countries and cultures**.** The Grit-S scale has two dimensions (with four items each): perseverance of effort and consistency of interests. Despite the boom in grit research, there are various ongoing debates about measuring this construct. In terms of dimensionality, the Grit-S scale was initially validated with two first-order factors (perseverance of effort and consistency of interests) and one second-order factor (grit; Duckworth & Quinn, 2009). However, this higher-order view of the structure of grit does not appear to be correct. In this regard, some recent studies have proposed a unidimensional structure with a single first-order factor (Areepattamannil & Khine, 2018; Gonzalez et al., 2020), or a two-factor structure with independent factors (Abuhassán & Bates, 2015; Datu et al., 2016). The underlying reason for these different results may be due to an overlap of the two dimensions, making it difficult to distinguish which items fit in one or the other. Something to note about the dimensionality of the test is that one of the two dimensions, consistency of interests, has all of its items in an inverse form, which may have helped the Factorial Analysis fit with two differentiated factors in the initial study in which the instrument was created (Duckworth & Quinn, 2009). This is because human beings tend to respond differently depending on the meaning of the question owing to the cognitive processing of direct and inverse items not necessarily being the same, particularly when reading ability is low (Marsh, 1986). For this reason, inverse items in Likert-type scales, and even including inverse and direct items in the same questionnaire, can have a negative impact on psychometric properties, and it is advisable to formulate all items in a direct manner (a more positive answer is associated with a higher level of the construct being evaluated; Suarez-Alvarez et al., 2018; Vigil-Colet et al., 2020).

In this line, the aim of this study is the development and initial validation of a new scale to evaluate grit in Spanish-speaking contexts.

2. Method

2.1. Participants and Procedure

The sample was initially made up of 222 participants from the general Spanish population. The sampling type was incidental. The members of the sample were aged between 18 and 83 years old, with a mean age of 38.60 and a standard deviation of 14.90 years.

We made individual contact with potential participants who met the inclusion criteria (being aged 18 or over). They were asked to respond to the questionnaire online, and to provide email addresses for other potential participants. The same process was repeated with these new potential participants. The anonymity of each participant was carefully respected, confidentiality was maintained, and we ensured strict compliance with current data protection laws.

2.2. Instrument

Oviedo Grit Scale (EGO; Escala Grit de Oviedo). In developing the Oviedo Grit Scale (EGO), we followed the criteria laid down by the European Federation of Psychological Associations (EFPA) for test evaluation and the Standards for Educational and Psychological Evaluation (AERA, APA, NCME, 2014), along with the recommendations from current psychometric literature (Muñiz & Fonseca-Pedrero, 2019). We constructed a sufficiently broad set of items (50 items) to cover each aspect of the two dimensions that *a priori* made up grit: perseverance of effort and consistency of interests. All of the items were written in a direct form (Suárez-Álvarez et al., 2018; Vigil-Colet et al., 2020). The response item was a Likert-type with 5 alternatives (1 completely disagree, 5 completely agree).

2.3. Data Analysis

The first phase of the study involved performing quantitative and qualitative analyses to assess how representative the content was (Sireci & Faulkner-Bond, 2014). In the next step, we asked 57 experts in psychometry or psychological evaluation from various Spanish universities to assign each of the 48 items to one of two dimensions that theoretically make up grit: perseverance of effort and consistency of interests. The level of inter-rater agreement about which dimension items belonged to was examined. In addition, we performed a chi-square test for each of the items to determine whether there were statistically significant differences between belonging to one or the other dimension.

Once we had obtained the 20 items for the questionnaire (10 per dimension), we made a preliminary application of it to a sample of 222 people taken from the general Spanish population (M = 34.23, SD = 15.85 years) for a preliminary evaluation of the quality of the item set. We performed an Exploratory Factor Analysis (EFA) to examine the dimensionality of the instrument. We used KMO and the Bartlett statistic to assess the suitability of the data for factorial analysis. The EFA was performed on the Pearson correlation matrix, using Exploratory Robust Maximum Likelihood (RML) as the method of estimation. We determined the dimensionality of the instrument by optimal implementation of parallel analysis with 1,000 random correlation matrixes. In addition, we used Unidimensional Congruence (UniCo), Explained Common Variance (ECV), and Mean of Item REsidual Absolute Loadings (MIREAL) to examine how well the data fit a single dimension. The following values support treating the data as essentially unidimensional: UniCo > .95; ECV > .85; MIREAL < .30 (Calderón-Garrido et al., 2019). We used the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA) as indices of fit, establishing a good fit when CFI > .95 and RMSEA < .06. Following this, we used a mixed statistical-substantive strategy to choose the final 10 items for the questionnaire. The strategy consisted of choosing the items that differed most between each other from those that had a factorial loading over .50. In addition, we kept in mind that there should be at least 3 items from each domain and that there should be items related to perseverance in long-term objectives, as well as consistency and passion for interests. Once the 10 final items were chosen, we performed an EFA to assess the dimensionality of the instrument, using all of the indicators and indices described above. We also examined the reliability of the instrument.

3. Results

In the first step, these items were reviewed by 24 graduate psychologists who scored each item between 1 and 10 in vocabulary and wording. The scores the judges assigned were evaluated using Aikens V index, which for vocabulary produced a value of .93 [.87-.96 CI= 95%], and for

wording .92 [.86-.95 CI= 95%], indicating excellent agreement. Nonetheless, two items were eliminated after scoring less than 8 in either vocabulary or wording. Furthermore, we removed 28 items for one of the following reasons: a) the initial assignment of the item was to a different dimension from the experts' assignment; b) there were no significant differences between belonging to one dimension or the other, according to the experts ($p > .05$); and c) the item had inter-rater agreement about which dimension it should be in which was below 80%. This allowed us to construct a preliminary instrument of 20 items (10 per dimension) to be analyzed in the quantitative pilot study.

In the first EFA, both the KMO (.96) and Bartlett's statistic (<.001) demonstrated that the data was suitable for factorial analysis. With the results we obtained, it seemed wise to reject a bidimensional structure for grit and maintain the hypothesis that a single factor was sufficient to demonstrate the psychological processes that could explain grit (Calderón-Garrido et al., 2019). A single factor explained 52% of the total variance, the optimal implementation of parallel analysis suggested a single dimension, and we found the following indicators for a unidimensional structure, UniCo = .956, ECV = .901, MIREAL = .174, CFI = .988, and RMSEA = 0.057. Following this, and using the mixed statistical-substantive strategy described previously, we selected the 10 final items for the questionnaire. We performed an EFA with the 10 final items, looking at the dimensionality of the instrument, indicating the data was suitable for factorial analysis (KMO: .96, Bartlett's statistic: <.001). Again, the results pointed toward rejecting a bidimensional structure for grit, and we maintained the hypothesis that a single factor was sufficient to explain the psychological processes underlying grit (UniCo: .972; ECV: .905; MIREAL: .155; CFI: .999 and RMSEA: .001). From an exploratory perspective, this enabled us to determine the instrument as essentially unidimensional.

Finally, the new scale (10 final items) demonstrated excellent reliability ($\alpha = .94$; $\omega = .94$).

4. Conclusions

The aim of this study was the development and initial validation of the Oviedo Grit Scale. This new instrument, in addition to being in Spanish, is an attempt to overcome some of the psychometric issues found in prior grit scales related to dimensionality as well as reliability and validity (Arco-Tirado et al., 2018; Clark & Malecki, 2019; Gonzalez et al., 2020). From an exploratory perspective, the new 10-item EGO demonstrates an essentially unidimensional internal structure (Calderón-Garrido et al., 2019), confirming previous studies that had shown grit to be unidimensional (Areepattamannil & Khine, 2018; Gonzalez et al., 2020).

References

Abuhassàn, T. C., & Bates, A. (2015). Grit. Journal of Individual Differences, 36, 205–214. https://doi.org/10.1027/1614-0001/a000175

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Psychological Association.

Areepattamannil, S., & Khine, M. S. (2018). Evaluating the psychometric properties of the original Grit Scale using Rasch Analysis in an Arab adolescent sample. Journal of Psychoeducational Assessment, 36(8), 856–862. https://doi.org/10.1177/0734282917719976

Calderón-Garrido, C., Navarro-González, D., Lorenzo-Seva, U., & Ferrando, P. J. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor-analytic approach for

assessing the dimensionality of tests and items. Psicothema, 31(4), 450–457. https://doi.org/10.7334/psicothema2019.153

Datu, J. A. D., Valdez, J. P. M., & King, R. B. (2016). Perseverance counts but consistency does not! Validating the Short Grit Scale in a collectivist setting. Current Psychology, 35, 121–130. https://doi.org/10.1007/s12144-015-9374-2

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. Personality Processes and Individual Differences, 92(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT-S). Journal of Personality Assessment, 91(2), 166–174. https://doi.org/10.1080/00223890802634290

Fernández-Martín, F. D., Arco-Tirado, J. L., & Hervás-Torres, M. (2020). Grit as a predictor and outcome of educational, professional, and personal success: A systematic review. Psicología Educativa, 26(2), 163–173. https://doi.org/https://doi.org/10.5093/psed2020a11

Gonzalez, O., Canning, J. R., Smyth, H., & Mackinnon, D. P. (2020). A psychometric evaluation of the Short Grit Scale: A closer look at its factor structure and scale functioning. European Journal of Psychological Assessment, 36(4) 646-657. https://doi.org/10.1027/1015-5759/a000535

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. Developmental Psychology, 22(1), 37–49. https://doi.org/10.1037/0012-1649.22.1.37

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. Psicothema, 31(1), 7–16. https://doi.org/10.7334/psicothema2018.291

Postigo, Á., Cuesta, M., Fernández-Alonso, R., García-Cueto, E., & Muñiz, J. (2021a). Academic grit modulates school performance evolution over time: A latent transition analysis. Revista de Psicodidáctica, 26(2), 87–95. https://doi.org/10.1016/j.psicoe.2021.03.001

Postigo, Á., Cuesta, M., Fernández-Alonso, R., García-Cueto, E., & Muñiz, J. (2021b). Temporal stability of grit and school performance in adolescents: A longitudinal perspective. Educational Psychology, 27(1), 77–84. https://doi.org/10.5093/psed2021a4

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema, 26(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Suárez-Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. Psicothema, 30(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. Psicothema, 32(1), 108–114. https://doi.org/10.7334/psicothema2019.286

Development and psychometric properties of the Concern for Appearance on the SN scale

Covadonga González-Nuevo[1], Marcelino Cuesta[1], Álvaro Postigo[1], Álvaro Menéndez-Aller[2]

[1]*Department of Psychology, University of Oviedo, Spain,*

[2]*Health Research Institute of Asturias, Spain*

Abstract

Social networks (SNs) are part of the new digital context in which people currently operate, offering numerous opportunities, but also some risks. The objective of this paper is to analyze the relationships between the use of SNs and the risk of eating disorders (ED). A total of 576 women over 18 years old participated with an average age of 28.88 years (SD = 11.14). The use of the two most used SNs (Facebook and Instagram), the concern for appearance on SNs, and eating attitudes were evaluated (the latter through the Eating Attitudes Test-26). The psychometric properties of the questionnaire developed to assess concern for appearance on SNs were excellent. Strong relationships were found between concern for appearance on SNs and the risk of ED. Instagram use predominates among younger women, while Facebook prevails among older women. The concern for appearance on SNs is an indicator of ED risk. The implications of the results for the prevention of ED were discussed.

Keywords: Eating disorder risk, Social networking, Appearance comparison, Body image, Facebook use, Instagram

Funding: This study has been supported by a predoctoral grant from the Principality of Asturias (BP19-032)

E-mail: covadongagonz@gmail.com

1. Introduction

The internet and social networks (SNs) are widely used. There are an estimated 4,388 billion internet users in the world, representing 57% of the total population, and 3,484 billion SN users (Global Digital Report, 2019). The most widely used SN is Facebook (FB), generally used by all ages (Pew Research Center, 2018). On the other hand, Instagram (IG) is the most highly visual social medium used (Global Digital Report, 2019), and is particularly popular with 18 to 25 year-olds (Pew Research Center, 2018). The widespread use and implementation of SNs have triggered interest from behavioral researchers. Eating disorders (ED) are one of the most widely studied disorders in relation to SNs due to appearance comparison factors (Burnell et al., 2019; Fardouly & Vartanian, 2014). Appearance comparison has been studied in connection with traditional media and the thin ideal body it shows, which has caused body dissatisfaction (Grabe et al., 2008). However, SNs have been proven to produce more body dissatisfaction and higher risks of ED than traditional media (Cohen & Blaszczynski, 2015). The use of SNs in relation to the risk of EDs has been measured in various ways. Most studies, according to Holland and Tiggemann (2016), have measured overall use as frequency of use, which is an insufficiently detailed measure because frequency of use alone is not a reliable predictor of ED, whereas appearance-related behaviors are.

In this regard, appearance-related use on SNs are those activities that are specifically related to appearance (e.g., looking at photographs) (Mingoia et al., 2017). The difference between appearance-related use and concern about appearance on SNs is in the consequences of the use. Appearance-related use of SNs is inherent in the use of SNs, as much of the functionality is based on exposure to photos and numerical indicators of social acceptance (e.g., likes). However, the problem arises when said use begins to cause excessive worry. In short, concern about appearance on SNs is defined as preoccupation about physical appearance (e.g., how I look in a photograph) or social appearance (e.g., having more likes than other photos do) on SNs which has negative consequences on a person's life. As far as we are aware, ours is the first questionnaire to evaluate this.

The present study attempts to overcome this problem, designing a new specific measuring instrument for the evaluation of concern about appearance on SNs. The instrument will be validated initially only with women, as they have a higher prevalence of ED (Striegel-Moore et al., 2009).Within this context, the objective of the present study is to develop a new measuring instrument to assess concern about appearance on SNs.

2. Method

2.1. Participants

The sample was comprised of 576 women, with ages ranging from 18 to 62 years old (M = 28.88 years; SD = 11.14). Participants who did not have or did not use FB and IG were removed from the study.

2.2. Instruments

Eating Attitudes Test-26 (EAT-26). The EAT-26 assesses ED risk (Garner et al., 1982). We used the Spanish version by Gandarillas et al. (2003) in the present study. The EAT-26 consists of 26 Likert items with a range from 1 ("never") to 6 ("always"). The items are divided into three subscales: Dieting, Bulimia, and Food Preoccupation and Oral Control, with a total score also being obtained. The internal consistency (Cronbach's coefficient α) is .88 for the total scale,.88

for the Dieting subscale, .77 for Bulimia and Food Preoccupation, and .79 for Oral Control (Gandarillas et al., 2003).

Concern about appearance in Social Networks (CONAPP). The CONAPP questionnaire was developed specifically for this study, and its objective is the evaluation of concern about appearance on SNs. Until now, there has been no measuring instrument for evaluating this construct, so it was necessary to develop a new one. We followed the recommendations in current psychometric literature in constructing it (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Downing, 2006; Downing & Haladyna, 2006; Drasgow, 2016; Irwing et al., 2018; Lane et al., 2015; Linn, 2011; Muñiz & Fonseca-Pedrero, 2019; Van der Linden, 2017). To that end, the first step was to define the variable to be measured. We performed a literature review and developed four theoretical areas for concern about appearance on SNs. These content areas were intended to thoroughly sample the construct being evaluated to ensure content validity. At no time were they considered as possible factors or dimensions because the aim was to produce an essentially unidimensional scale, which would allow the production of a global score for the individuals being evaluated (Calderón et al., 2019). The next step was the construction of a sufficiently broad bank of 52 items to address these theoretical areas. These items were assessed via two strategies: expert assessment and psychometric analysis.

They used a judgmental approach, evaluating aspects related to the appropriate formulation of items, and their knowledge and expertise, attempting to avoid any kind of ambiguity that items might produce in the participants. If there were doubts raised about any items, they were eliminated. This led to 11 items being removed. With the 41 remaining items, we calculated various psychometric indicators, discarding 10 items with low discriminative power (discrimination indices below .40) and with lower factorial loadings. The final scale comprised 31 items.

The questionnaire was presented to the participants in Spanish. The items were Likert type with 6 response options, 1 being "Never" and 6 "Always." No reversed items were included to avoid the biases that these may produce (Suárez et al., 2018).

Overall use of Social Networks. The assessment of overall use of SNs was carried out using the same question for each of the SNs studied: How much time do you spend looking at FB on any given day? How much time do you spend looking at Instagram on any given day? The question had six categories. For the analyses, scores from 1 to 6 were assigned to each of the categories of the scale.

2.3. Procedure

The participants completed an online survey anonymously and voluntarily, giving their informed consent before starting. Participants were initially contacted through various SN pages and sites.

2.4. Data Analysis

The psychometric properties of the tests used were analyzed, both from the point of classical theory and item response theory (Muñiz, 2018). Exploratory factor analysis (EFA) was used to study the internal structure with the unweighted least squares method (Lloret-Segura et al., 2014). EFA was chosen because in this first validation study for the instrument, the authors believed it was a risk to pose strict hypotheses about the dimensionality of the instrument, a requirement needed by a CFA. The Polychoric Correlation Matrix was used between items because the items were Likert type and the distribution did not approximate normality. The

procedure for determining the number of factors was the optimal implementation of parallel analysis (PA) (Timmerman & Lorenzo-Seva, 2011). Reliability was estimated using the Cronbach (1951) coefficient α, and McDonald (1999) coefficient ω. Within the item response theory (IRT) framework, the reliability was estimated using the information function.

The EFA was carried out using the FACTOR program (10.10.01 version) (Lorenzo-Seva & Ferrando, 2006). The IRTPRO (4.2 version) (Muraki & Bock, 2003) was used for the IRT analyses. The other analyses were performed using the statistical package SPSS (22.0 version).

3. Results

All the discrimination indices (DI) were above .40 since the original items with lower values were eliminated in order to enhance the unidimensionality of the questionnaire, and thus allow a global score to be produced. IRT-a parameter values ranged from .88 to 2.36.

The adequacy of the data for factor analysis of the CONAPP scale items was tested with the KMO test (KMO = .94) and the Bartlett test ($p \leq .001$). The PA (Calderón et al., 2019; Timmerman & Lorenzo-Seva, 2011) suggested the presence of a single factor. This result was supported by the GFI indicators (.96), which were greater than .95, so it is considered that there is a good fit (Ferrando & Anguiano-Carrasco, 2010); the explained variance (42.06%), and the RMSR (.079), indicated an acceptable fit (Ferrando & Anguiano-Carrasco, 2010). The factor loadings varied between .47 and .76. All these results indicate that the test can be understood as essentially unidimensional (Calderón et al., 2019).

Correlations were calculated between the scores of the participants in the CONAPP test and the total score and sub-scores in the EAT-26. The highest correlation was between EAT-TOTAL and CONAPP ($r = .379$, $p \leq .001$). The subscale with the highest correlation was Bulimia and Food Preoccupation and CONAPP ($r = .334$, $p \leq .001$), followed by Dieting and CONAPP ($r = .328$, $p \leq .001$), and the lowest correlation was found between Oral Control and CONAPP ($r = .205$, $p \leq .001$).

The reliability of the CONAPP had an alpha coefficient of $\alpha = .952$ (Cronbach, 1951), and an omega coefficient of $\omega = .953$ (McDonald, 1999), which can be considered optimal values according to the European model of test quality assessment (Muñiz, 2018). On the other hand, the information function indicates that the CONAPP questionnaire is more accurate for the relatively high levels of the trait evaluated (above $\Theta = -1$).

4. Conclusions

A new questionnaire has been developed to assess concern about appearance on social networks (SNs). The new instrument consists of 31 Likert-type items and shows an essentially unidimensional structure and excellent reliability. Clear relationships were found between the level of concern about appearance on SNs shown by the participants and eating disorder (ED) risk.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (DSM 5) (6th ed). Washington, DC, EEUU: American Psychiatric Press.

Burnell, K., George, M. J., Vollet, J. W., Ehrenreich, S. E., & Underwood, M. K. (2019). Passive social networking site use and well-being: the mediating roles of social comparison and the fear of missing out. Cyberpsychology: *Journal of Psychosocial Research on Cyberspace*, *13*(3), 1–14. http://dx.doi.org/10.5817/CP2019-3-5

Calderón, C., Navarro, D., Lorenzo-Piera, U., & Ferrando, P. J. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor analytic approach for assessing the dimensionality of tests and items. *Psicothema*, *31*(3), 450–457. https://doi.org/10.7334/psicothema2019.153

Cohen, R., & Blaszczynski, A. (2015). Comparative effects of Facebook and conventional media on body image dissatisfaction. *Journal of Eating Disorders*, *3*(23), 1–11. https://doi.org/10.1186/s40337-015-0061-3

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Drasgow, F. (Ed.) (2016). *Technology and testing*. Routledge

Fardouly, J., & Vartanian, L. R. (2014). Negative comparisons about one's appearance mediate the relationship between Facebook usage and body image concerns. *Body Image*, *12*, 82–88. https://doi.org/10.1016/j.bodyim.2014.10.004

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología [Factor analysis as a research technique in psychology] [Monograph]. *Papeles del Psicólogo*, *31*(1), 18–33. http://www.cop.es/papeles

Gandarillas, A., Zorrilla, B., Sepúlveda, A. R., & Muñoz, P. E. (2003). *Trastornos del comportamiento alimentario. Prevalencia de casos clínicos en mujeres adolescentes de la Comunidad de Madrid* (Documentos Técnicos de Salud Pública No. 85) [Eating behavior disorders. Prevalence of clinical cases in adolescent women of the Community of Madrid (Public Health Technical Documents No. 85)]. http://www.madrid.org/sanidad

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The Eating Attitudes Test: Psychometric features and clinical correlates. *Psychological Medicine*, *12*(4), 871–878. https://doi.org/10.1017/S0033291700049163

Global Digital Report. (2019*). Essential insights into how people around the world use the internet, mobile devices, social media and e-commerce.* https://hootsuite.com/pages/digital-in-2019

Grabe, S., Ward, L. M., & Hyde, J. S. (2008). The role of the media in body image concerns among women: A meta-analysis of experimental and correlational studies. *Psychological Bulletin*, *134*(3), 460–476. https://doi.org/10.1037/0033-2909.134.3.460

Holland, G., & Tiggemann, M. (2016). A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. *Body Image*, *17*, 100–110. https://doi.org/10.1016/j.bodyim.2016.02.008

Irwing, P., Booth, T. & Hughes, D. J. (Eds.) (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development.* John Wiley & Sons Ltd.

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). The exploratory factor analysis of items: guided analysis based on empirical data and software. *Anales de Psicología*, *30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

McDonald, R.P. (1999). *Test theory: A unified treatment*. Erlbaum

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales (4.20 version) [Computer software]. Skokie, IL: Scientific Software International, Inc.

Muñiz, J. (2018). *Introducción a la psicometría [Introduction to psychometry]*. Pirámide.

Mingoia, J., Hutchinson, A. D., Wilson, C., & Gleaves, D. H. (2017). The relationship between social networking site use and the internalization of a thin ideal in females: A Meta-Analytic Review. *Frontiers in Psychology*, *8*, 1–10. Article 1351 https://doi.org/10.3389/fpsyg.2017.01351

Pew Research Center. (2018). Social media use in 2018. https://www.pewinternet.org/2018/03/01/social-media-use-in-2018

Striegel-Moore, R. H., Dohm, F. A., Kraemer, H. C., Taylor, C. B., Daniels, S., Crawford, P. B., & Schreiber, G. B. (2003). Eating disorders in white and black women. American *Journal of Psychiatry*, *160*(7), 1326–1331. https://doi.org/10.1176/appi.ajp.160.7.1326

Suárez, J., Pedrosa, I., Lozano, L., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149-158. https://doi.org/10.7334/psicothema2018.33

Tiggemann, M., & Barbato, I. (2018a). "You look great!": The effect of viewing appearance-related Instagram comments on women's body image. *Body Image*, *27*, 61–66. https://doi.org/10.1016/j.bodyim.2018.08.009

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209. https://doi.org/10.1037/a0023353

Spanish validation of the Acceptance and Action Questionnaire-II

Álvaro Menéndez-Aller[1], Jaime García-Fernández[2], Covadonga González-Nuevo[2], Eduardo García-Cueto[2]

[1]*Health Research Institute of Asturias, Spain,*

[2]*Department of Psychology, University of Oviedo, Spain*

Abstract

Experiential Avoidance is a trans-diagnostical variable that contributes to the development of psychological problems. AAQ-II measures this construct but the sample used for its validation with the general population was not representative. In this study, a validation of AAQ-II on the general Spanish population was carried out. Through an online form, the Experiential Avoidance, personality, emotional intelligence, anxiety and depression of 964 participants (52.8% women) from all over the Spanish territory were evaluated. Moreover, the influence of sex and age in Experiential Avoidance was analyzed. Finally, through structural equations, it was contemplated whether Experiential Avoidance influenced the effect of personality over psychopathological variables. AAQ-II is a valid instrument for use in the general Spanish population. Experiential Avoidance decreases with age and differs significantly between men and women. The proposed path-analysis model shows a good fit. Experiential Avoidance is an important psychological variable and it is possible to study how it affects more variables than the clinical ones.

Keywords: Experiential Avoidance, AAQ-II, Anxiety, Depression, sex, age

E-mail: alvaromaller@outlook.es

1. Introduction

The Acceptance and Commitment Therapy (ACT) has established itself as an effective and efficient therapy in the last 30 years (Samaan et al., 2020). ACT proposes a trans-diagnostical model for the development of psychological issues: psychological inflexibility, which is defined as a tendency by which an individual may rigidly avoid coming into contact with unpleasant internal experiences (for example, uncomfortable or hurtful thoughts, memories …) (Hayes et al., 2019). ACT states that such treatment of negative internal experiences is counterproductive, because the same experiences may be reexperienced, sometimes even with greater intensity (Hayes et al., 2019). Of all of the components of Psychological Inflexibility, Experiential Avoidance (EA) is said to be the most detrimental. EA has been defined as the reluctancy to deal with negative internal experiences, which push the individual to perform certain actions so as to avoid or escape such experiences (Monestès et al., 2016). As stated above, such actions may lead to more intense and disruptive internal events (anxiety, sadness, intrusive thoughts…). Therefore, EA is considered an important variable in the development of psychological problems.

The relationship between EA and non-clinical variables has also been observed. For example, Cobos-Sánchez et al. (2017) observed that a reduction of psychological inflexibility (and, therefore, a lower EA) through training influences Emotional Intelligence; more specifically, higher Emotional Clarity and Emotional Reparation, and lower Emotional Attention. Previous research has also found that EA correlates positively with Neuroticism and negatively with the remaining Big Five traits (Steenhaut et al., 2018). Moreover, it mediates the relationship between personality and several measures of psychological well-being (Steenhaut et al., 2018).

Many questionnaires have been created in order to measure EA. The Acceptance and Action Questionnaire-II (AAQ-II; Bond et al., 2011) is one of them. The scale has been previously validated in a clinical Spanish population (Ruiz et al., 2013). Moreover, in the same paper, the authors attempted to validate the instrument in the general Spanish population. However, the sample used for this validation was comprised exclusively of university students and teachers from a city in the south of Spain. Thus, sample bias could have taken place.

The main aim of this research is the validation of the AAQ-II in the general Spanish population. This objective breaks down into multiple secondary objectives, such as the assessment of internal consistency, dimensionality, item discrimination indexes and evidence of validity of the instrument; more specifically, the relationship between EA and depression, anxiety, personality and emotional intelligence was measured. Finally, a structural equation model was generated to explain the interaction between personality, emotional intelligence, EA and clinical variables such as depression and anxiety.

2. Method

2.1. Participants

The sample was comprised of 964 participants (52.8) from all over Spain, with ages ranging from 18 to 84 years old (M = 43.43 years; SD = 15.27).

2.2. Instruments

Acceptance and Action Questionnaire (AAQ-II; Bond et al., 2011). The Spanish adaptation from Ruiz et al. (2013) was used. This is a self-report with 7 Likert-type items with 7 response options, where 1 means "*never true*" and 7 means "*always true*". The reliability of the adapted version was $\alpha = .88$.

Educational-Clinical Anxiety and Depression Questionnaire (CECAD; Lozano et al., 2010). This is a self-report with 50 Likert-type items split into two dimensions: (a) Depression, with α = .95 in the manual; and (b) Anxiety, with α = .91. All of the items were Likert-type, with five response options, where 1 means "*Completely disagree*" and 5 means "*Completely agree*".

Trait Mood-Meta Scale (TMMS; Salovey et al., 1995). The Spanish adaptation by Fernández-Berrocal et al. (2004) was used, which is a reduced version of the original, called the TMMS-24. This is a self-report with 24 items split into three dimensions: (a) Emotional Attention, with α = .90 in the adapted version; (b) Emotional Clarity, with α = .90; (c) Emotional Repair, with α = .86. Each dimension has eight Likert-type items with five response options, where 1 means "*Completely disagree*" and 5 means "*Completely agree*".

Overall Personality Assessment Scale (OPERAS; Vigil-Colet et al., 2013). This is a self-report with 40 items split into the Big Five traits: (a) Extraversion, with α = .86 in the original study; (b) Neuroticism, with α = .86; (c) Conscientiousness, with α = .77; (d) Agreeableness, with α = .71; and (e) Openness to Experience, with α = .81. Each trait has 7 Likert-type items with five response options, where 1 means "*Completely disagree*" and 5 means "*Completely agree*".

2.3. Procedure

The scales were mixed randomly in an online form, with the only condition that two consecutive items did not measure the same construct. Next, the online form was distributed through snowball sampling.

2.4. Data Analysis

The data was analyzed with IBM SPSS software (Version 24), Factor (Version 10.10.02) and MPlus (Version 8).

To confirm the fit of AAQ-II to a unidimensional structure, we carried out a confirmatory factor analysis using robust unweighted least squares and a matrix of polychoric correlations. We used two different indices to confirm a good fit of the data (Kline, 2011): CFI, which had to be above .90, and RMSEA, which had to be below 0.08 (Hoyle, 2012). To calculate the item discrimination indices for the AAQ-II items, the corrected correlation coefficient between each item and the overall score in the test was calculated. Items were considered to demonstrate adequate discrimination if this index was over .3.

To estimate the internal consistency of the used scales, Cronbach's alpha was used. We used the Pearson correlation coefficient to study the relationship of AAQ-II with the CECAD, TMMS-24, and OPERAS.

To assess the influence of personality over clinical variables, such as anxiety and depression, moderated by Emotional Intelligence and EA, a path-analysis model was proposed. In order to study the fitness of the model, a cross-validation method was used. Therefore, the sample was split in two random samples. In the first, modification indices were employed so as to find the best fit possible; the analysis were, then, repeated in the second sample using the same parameters. Two indices were used in both to confirm the fitness of the model (Kline, 2011): CFI, whose value had to be greater than .90, and RMSR, which had to be below 0.10 to indicate a correct fit (Hoyle, 2021).

3. Results

A confirmatory factor analysis was performed first. We found a good fit to a unidimensional structure. The CFI was over .90 (CFI = .996) and the RMSEA was below 0.08 (RMSEA = 0.064). Moreover, all of the items had indices of discrimination well above the .30 criterion, ranging from .687 to .796.

In terms of reliability of the instruments used in the study, the AAQ-II (α = .93) and the scales of Anxiety (α = .90), Depression (α = .95), Emotional Attention (α = .90), Emotional Clarity (α = .91), Emotional Repair (α = .85), Extraversion (α = .85), and Neuroticism (α = .88) demonstrated excellent internal consistency (over .80) (Hernández et al., 2016). Good internal consistency (over .70) was exhibited by the scales for Conscientiousness (α = .77), Agreeableness (α = .74), and Openness to Experience (α = .79).

Table 1 shows the correlation coefficients between the AAQ-II and the rest of the instruments employed in the current work. All of the correlations followed the expected patterns in the expected direction.

Table 1. Pearson correlations between the AAQ-II and the CECAD, TMMS-24, and OPERAS scales.

Scales	AAQ-II
Anxiety	.59
Depression	.77
Emotional Attention	.52
Emotional Clarity	−.38
Emotional Repair	−.35
Extraversion	−.23
Neuroticism	.74
Conscientiousness	−.30
Agreeableness	−.24
Openness to Experience	−.11

The results of the cross-validated path analysis are presented in Table 2, along with the results of the path analysis carried out using the whole sample. In the three instances, the indices showed an adequate fit, with similar values between analyses. A representation of the model is shown in Figure 1, along with the statistically significant regression coefficients extracted from the path analysis carried out with the whole sample.

Table 2. Cross-validated path analysis fit indicators

Samples	CFI	RMSR
Sample 1 (n = 486)	.913	0.094
Sample 2 (n = 478)	.914	0.100
Total sample (n = 964)	.911	0.097

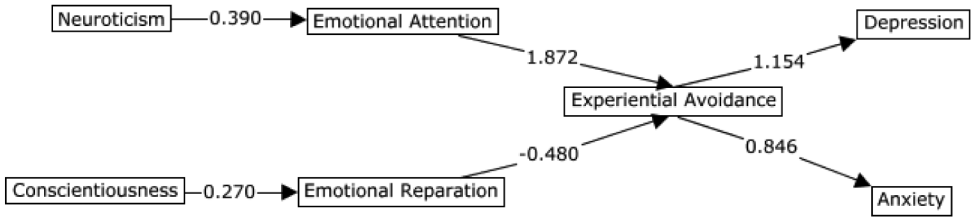


Figure 1. Visual scheme of the proposed model

4. Conclusions

The trans-diagnostical model proposed by ACT known as Psychological inflexibility has been linked to various mental health issues (Hayes et al., 2019). EA is the most detrimental of all of its components, as it deals with the avoidance and reexperiencing of negative internal events (emotions, thoughts or physiological states that may be uncomfortable or hurtful to the individual) (Monestès et al., 2016). Because of this, several questionnaires have been developed in order to measure and research the extent of the influence of EA. The AAQ-II is one of them. The questionnaire had previously been validated in the general Spanish population, yet the sample used was taken exclusively from a university context in only one Spanish city (Ruiz et al., 2013). Thus, the main objective of this research was to validate the AAQ-II in the general Spanish population with a more representative sample, taken from all over Spain.

As proven in the current paper, the AAQ-II is essentially unidimensional (CFI = .996; RMSEA = 0.064). It has shown excellent reliability (α = .93) and the discrimination indices of items are all correct (greater than .3). The correlations coefficients shown in Table 1 serve as the evidence of validity with other variables. In conclusion, the AAQ-II is a valid and reliable instrument for its use in the general Spanish population. This questionnaire would not only help to detect possible at-risk cases in the non-clinical population, but could also facilitate further research on the impact of EA on non-clinical variables.

The fit indicators shown in Table 2 indicate that the proposed model may serve as a reliable scheme through which the relationship between personality and clinical variables may be understood. As presented in Figure 1, personality traits such as Conscientiousness and Neuroticism propitiate a higher level of Emotional Reparation and Emotional Attention, respectively. These Emotional Intelligence factors have a particular impact on EA; particularly, a high score in Emotional Attention (described as the frequency in which the individual pays to attention to his emotions) will result in a high score in EA while a high score in Emotional Reparation (the ability to deal positively with one's emotions) will decrease EA. EA, in turn, acts as a risk factor in clinical variables such as anxiety and depression. In short, this model could be considered as a first stepping stone to understanding how EA acts in conjunction with non-clinical variables and how those dynamics may influence clinical outcomes.

References

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., Waltz, T., & Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire–II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy, 42*(4), 676–688. https://doi.org/10.1016/j.beth.2011.03.007

Cobos-Sánchez, L., Flujas-Contreras- J. M., y Gómez-Becerra, I. (2017). Intervención en flexibilidad psicológica como competencia emocional en adolescentes: Una serie de casos. *Revista de Psicología Clínica con Niños y Adolescentes*, *4*(2), 135–141.

Hayes, S. C. (2019). Acceptance and Commitment Therapy: Towards a unified model of behavior change. *World Psychiatry, 18*(2), 226–227. https://doi.org/10.1002/wps.20626

Hoyle, R. H., (2012). *Structural equation modeling*. The Guilford Press.

Kline, R.B. (2011). *Principles and Practice of Structural Equation Modeling.* The Gilford Press.

Lozano, L., García-Cueto, E. & Lozano, L. M. (2010). *Cuestionario Educativo-Clínico: Ansiedad y Depresión (CECAD)* [Educational-Clinical Questionnaire: Anxiety and Depression]. TEA Ediciones.

Monestès, J. L., Karekla, M., Nele, J., Michaelides, M. P., Hooper, N., Kleen, M., Ruiz, F. J., Miselli, G., Presti, G., Luciano, C., Villatte, M., Bond, F. W., Kishita, N., & Hayes, S. C. (2016). Experiential Avoidance as a common psychological process in European cultures. *European Journal of Psychological Assessment*, *34*, 247–257. https://doi.org/10.1027/1015-5759/a000327

Ruiz, F. J., Langer Herrera, A. I., Luciano, C., Cangas, A. J., & Beltrán, I. (2013). Measuring experiential avoidance and psychological inflexibility: The Spanish version of the Acceptance and Action Questionnaire-II. *Psicothema, 25*(1), 123–129. https://doi.org/10.7334/psicothema2011.239

Salguero, J. M., Fernández-Berrocal, P., Balluerka, N., & Aritzeta, A. (2010). Measuring Perceived Emotional Intelligence in the adolescent population: Psychometric properties of the Trait Meta-Mood Scale. *Social Behavior and Personality: An International Journal, 38*(9), 1197–1209. https://doi.org/10.2224/sbp.2010.38.9.1197

Salovey, P., Mayer, J. D., Goldman, S., Turvey, C, & Palfai, T. (1995). Emotional attention, clarity, and repair: Exploring emotional intelligence using the Trait Meta-Mood Scale. In J. W. Pennebaker (Ed.), *Emotion, disclosure, and health* (pp. 125–154). American Psychological Association.

Samaan, M., Diefenbacher, A., Schade, C., Dambacher, C., Pontow, I.-M., Pakenham, K., y Fydrich, T. (2020). A clinical effectiveness trial comparing ACT and CBT for inpatients with depressive and mixed mental disorders. *Psychotherapy Research,* 1–14. https://doi.org/10.1080/10503307.2020.1802080

Steenhaut, P., Rossi, G., Demeyer, I., & De Raedt, R. (2018). How is personality related to well-being in older and younger adults? The role of psychological flexibility. *International Psychogeriatrics*, 1–11. https://doi.org/10.1017/s1041610218001904

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., & Lorenzo-Seva, L. (2013). Development and validation of the Overall Personality Assessment Scale (OPERAS). *Psicothema, 25*(1), 100–106. https://doi.org/10.7334/psicothema2011.411

Development of a Gender Roles Scale

Jaime García-Fernández[1], Eduardo García-Cueto[1], Álvaro Postigo[1], Covadonga González-Nuevo[1]

[1]*Department of Psychology, University of Oviedo, Spain*

Abstract

While sex is defined as the biological difference between men and women, gender has been associated to behavioral, social, and cultural differences related to men and women. Differences in personality traits based on sex are still being found in recent research. The aim of the study was the creation of a self-report battery to assess gender roles (the ERGO) and to observe whether the differences by sex in personality traits are controlled by gender role adherence. The sample was made up of 612 Spaniards. They completed the ERGO scale online, as well as other personality measures. The psychometric parameters of the ERGO were evaluated: factor structure, test reliability and evidence of validity. Differences were analyzed according to sex, with gender as the control variable. The ERGO proved to be a reliable instrument and showed good evidence of both internal and external validity. Differences between men and women were found in some personality traits, although these differences changed when gender was controlled. Future research should consider gender role adherence when analyzing differences by sex.

Keywords: Psychometric properties; assessment; gender roles.

E-mail: cueto@uniovi.es

1. Introduction

Since John Money established the distinction between sex and gender (Money et al., 1955), researchers started to gradually displace the term sex, and replacing it with gender, mainly in the fields of humanities and social sciences (Haig, 2004). This distinction remains in force, defining sex (male or female) as the physical and biological traits which distinguish men and women, and gender (masculinity or femininity) as the differences in behavioral, social, and cultural aspects (VandenBos, 2015).

Gender attributes have varied in accordance with the sociohistorical context (García-Cueto et al., 2015; López-Sáez and García-Dauder, 2020). In the early years, masculinity and femininity were considered the endings of a continuum, where low masculinity implied high femininity levels (López-Sáez y García-Dauder, 2020). Throughout the last century, this model was heavily criticized, leading to a consideration where masculinity and femininity constitute different dimensions (Constantinople, 1973; Fernández, 2011; López-Sáez y García-Dauder, 2020). Nevertheless, in the last twenty years the bidimensional conceptualization of gender has also been questioned, given the poor amount of explained variance (Fernández, 2011). Choi y Fuqua (2003) conducted a revision about the factorial structure of the Bem Sex Role Inventory. They concluded that a multidimensional structure showed a better fit than a bidimensional one, assigning one factor to femininity, and two or more to masculinity. These results have been supported by further investigations (Choi et al., 2006), leading to a multidimensional conceptualization of gender.

Differences in gender regarding sociodemographic variables have also been observed. Women with a higher educational level score lower in femininity (Bringas-Molleda et al., 2016; Paino et al., 2017), and youths are lower than adults in masculinity and femininity scales (Fernández et al., 2014; Fernández-Rodríguez et al., 2018), suggesting that new generations have adopted more equal gender roles, rejecting the traditional behaviors classically assigned to women and men (Andrade, 2016). However, gender investigations still show differences regarding sex (Kachel et al., 2016). Within the Big Five Model, women score higher in Neuroticism and Amiability, and high scores in Extraversion and Openness are related to men (Pedrosa et al., 2010; Smith et al., 2019). Differences in more specific personality traits have also been reported. Postigo et al. (in press) analyzed the Entrepreneur Personality Evaluation Battery (BEPE; Muñiz et al., 2014) scores in 1170 participants. They found significant differences in Stress Tolerance and Risk Taking, with males scoring higher than females. It has also been reported that women have more pessimistic expectations than men in the first year of university (Araújo et al., 2019), and tend to score higher on self-handicapping (Ferradás et al., 2018). These differences regarding sex may be mediated by gender. Thereby, differences in verbal fluency, spatial orientation, neuroticism, and abstract reasoning disappear when controlling by gender (Pedrosa et al., 2010).

This research has two objectives: (1) The first objective is the creation and validation of the Oviedo Gender Roles scale (ERGO). (2) The second objective consists of analyzing the influence that gender and sex may have on general personality traits (Agreeableness and Openness) and some aspects of enterprising personality (Internal Locus of Control, Achievement motivation, Stress Tolerance, Risk Taking).

2. Method

2.1. Participants

A sample of 612 legal-age Spaniards was recruited, with 55.4% of women. The age range went from 18 to 83 years (M=34.2; SD=15.9).

2.2. Instruments

Oviedo Gender Roles Scale (ERGO). Self-measure of 44 five-point Likert items, which constitute two dimensions: masculinity and femininity. Several steps were followed in its development: (1) 30 people from the general population wrote down attitudes and behaviors typically regarded as masculine or feminine. 41 typically feminine and 36 typically masculine traits were obtained. (2) 128 psychologists assessed each trait classifying them as masculine, feminine or neutral. The items which not considered typical of one gender by at least 75% of the psychologists were dropped. Neutral traits were also excluded. 25 feminine and 23 masculine traits remained as being typical of each gender. (3) Three psychometricians revised the wording of each item, following the recommendations on the construction of gender role scales (Baber and Tucker, 2006). Three items from the femininity and one from the masculinity scale were dropped. Thus, the final scales were composed of 22 items each.

NEO five Factor Inventory traits (NEO-FFI). Self-report measure of the Five Factor Model of Personality developed by Costa y McCrae (1985) with a Spanish adaptation by Cordero et al. (2008). Two subscales were used: (1) ***Amiability:*** Being altruistic, pleasant and caring ($\alpha = .89$). (2) ***Openness:*** Interest in original, artistic, or novel things ($\alpha = .89$).

Entrepreneur Personality Assessment Battery (BEPE). This is a self-report measure developed in Spanish by Muñiz et al. (2014). Four subscales were used in this research: (1) ***Internal Locus of Control:*** The causal attribution that the consequences of a behavior depend on oneself ($\alpha = .94$). (2) ***Achievement Motivation:*** The desire to achieve standards of excellence, i.e., achieving and improving objectives ($\alpha = .95$ (3) ***Stress Tolerance***: The resistance to perceive environmental stimuli as stressful thanks to the adequate use of coping strategies ($\alpha = .91$). (4) ***Risk Taking:*** The tendency and willingness of people to take on certain levels of insecurity that will allow them to achieve a goal that brings greater profits than the possible negative consequences ($\alpha = .96$).

Attention Scale. Nine items assessed the attention degree, asking the participant to select a specific alternative. This ensure the participant's responses to not be hazardous.

2.2. Procedure

The scales were applied through an online questionnaire. The items were randomized with the only requirement of not having two items on the same scale consecutively. A snowball sampling was applied, disseminated by close contacts and social media.

2.2. Data analysis

An Exploratory Factor Analysis (EFA) was conducted to assess the ERGO dimensionality. Polychoric correlations were used for the assemblance of the inter-item correlation matrix. The data suitability for the EFA was assessed with the KMO index and the Bartlett statistic. The advised number of dimensions was obtained using Parallel Analysis (Lorenzo-Seva y Ferrando, 2020). Factorial weights were estimated with a robust (no-weighted) least squares method and rotated using a Promin rotation. Items with similar weights in more than one factor were dropped iteratively. As data fit indexes, CFI (advised value over 0.9) and RMSR (advised value under 0.1, ideally under 0.08) were obtained (Hoyle, 2012; Hu and Bentler, 1999).

The Discrimination Index (corrected) was estimated for all the ERGO items, excluding items with indexes lower than .3 (Hernández et al., 2016). Cronbach's alpha was used as an estimation of the scale's reliability. An interdimensional correlation matrix was estimated. The AVE index was obtained in order to assess the convergent-divergent validity of the ERGO. This index is

proof of convergent validity when its value is higher than 0.5. If the squared correlations between scales are lower than each scale AVE index, divergent validity can be accepted (Hair et al., 2009).

The differences in personality traits regarding sex were tested with an ANOVA, applying Bonferroni's correction and estimating the size effect (Lenhard y Lenhard, 2016). Later, an ANCOVA with the significant differences was performed, controlling by gender.

Data was analyzed with IBM SPSS (24 version) and Factor (10.10.02 version; Lorenzo-Seva and Ferrando, 2020).

3. Results

The ERGO scale showed data suitability for conducting an EFA (Bartlett's statistic $p < 0.01$; KMO = .79). The Parallel Analysis recommended the extraction of three dimensions. The items' factorial weights were coherent with the multidimensional theory of gender (Choi and Fuqua, 2003). Following the given criteria, 28 items were removed from the scale, as they had similar factorial weights in two or more dimensions. The 16 items in the ERGO final version showed a good fit to a tridimensional structure (CFI = 0.961; RMSR = 0.041; % explained variance = 54.54%). The first dimension is related to socioemotional issues generally assigned to women. The second dimension refers to sexual comparisons. The third dimension is related to aggressiveness.

The items showed discrimination indexes higher than .3, with ranges of .63−.71 (Socioemotional), .72−.77 (Comparison), .57−.70 (Aggressiveness). The reliability estimation of the scales was above $\alpha = .70$ (Socioemotional $\alpha = .75$; Comparison $\alpha = .81$; Aggressiveness $\alpha = .77$). The correlation between the Comparison and Aggressiveness scale was positive ($r = .47$), with the correlations of Socioemotional with Comparison ($r = -.25$) and with Aggressiveness ($r = -.29$) being negative. Convergent validity could not be concluded (Socioemotional AVE = 0.36; Comparison AVE = 0.43; Aggressiveness AVE = 0.48). However, the squared correlations between scales were lower than each AVE index, bringing evidence of discriminant validity.

The results of the differences between personality traits by sex are showed in Table 1. Only the ERGO scales, the Agreeableness scale of the NEO-FFI and the Achievement motivation and Stress Tolerance scale of the BEPE showed significant differences ($p < .001$). A second analysis was run with the personality traits with significant differences, this time controlling by gender. As can be seen in Table 1, when controlling by the ERGO scales, some of the differences by sex disappeared.

Table 1. ANOVA of personality traits by sex, and ANCOVA controlling by gender.

Gender & Personality traits	ANOVA			ANCOVA
	p (<.016)	partial η^2	group	p (<.016)
ERGO – Socioemotional	< .001	.108	Female	−
ERGO – Comparison	< .001	.324	Male	−
ERGO – Aggressiveness	< .001	.147	Male	−
NEO−FFI – Agreeableness	< .001	.035	Female	.130
BEPE – Achievement Motivation	< .001	.020	Female	< .001
BEPE – Stress Tolerance	< .001	.067	Male	.460

Note. group = sex with the higher main score. In brackets, the significance level to compare with after applying the Bonferroni correction. The non−significant differences have been excluded from the table.

4. Conclusions

The present research offers a psychometric instrument to assess gender roles on the Spanish population and, at the same time, it has assessed whether these roles mediate sexual differences in personality traits.

The EFA has revealed the ERGO as a scale with great psychometric properties, having a multidimensional structure coherent with the previous literature (Choi & Fuqua, 2003; Fernández et al., 2014). The three scales that shape the questionnaire (Socioemotional, Comparison and Aggressiveness) show great reliability and evidence of discriminant validity. The absence of convergent validity could be explained by the scale's nature. This means that, although gender is not a unidimensional construct, masculine and feminine roles tend to have an inverse relationship, as can be drawn from the correlations between factors. Bearing this in mind, the Socioemotional factor could be related to femininity, and the Comparison and Aggressiveness factors with masculinity. This is also supported by the significant differences in the scores and their size effects, with women scoring higher in Socioemotional and men lower in Comparison and Aggressiveness.

The differences by sex observed in Agreeableness and Stress Tolerance disappeared when controlled by the ERGO scores. This finding implies that classical differences usually associated with sex may be caused by gender. Thus, the basis for these differences is not biological but psychosocial, as other authors had concluded (Pedrosa et al., 2010). Nevertheless, this statement it is not applicable to all traits. For example, the sexual differences did not disappear after controlling by gender, meaning that this difference could have a biological base or, more likely, is mediated by other variables than gender (e.g., socioeconomic status).

Finally, some limitations must be mentioned. The main one is the sample imbalance, with the vast majority of the participants being college students. In addition, some important variables were not included, such as the gender identity or sexual orientation of the participant. Future research should consider studying this topic in an older sample, as well as collecting more variables that could mediate in the individual gender perception.

In summary, the ERGO has emerged as a compelling scale for assessing gender roles in young Spaniards. The research also showed how gender can mediate in some personality differences between males and females. Given these results, future research should consider gender role adherence when analyzing differences by sex.

References

Andrade, C. (2016). Adaptation and factorial validation of the attitudes toward gender roles scale. *Paideía, 26*(63), 7–14. https://doi.org/10.1590/1982-43272663201602

Araújo, A. M., Assis-Gomes, C. M., Almeida, L. S., & Núñez, J. C. (2018). A latent profile analysis of first-year university students' academic expectations. *Annals of Psychology*, *35*(1), 58–67. https://doi.org/10.6018/analesps.35.1.299351

Baber, K. M., & Tucker, C. J. (2006). The Social Roles Questionnaire: A new approach to measuring. *Sex Roles, 54*, 459–467. https://doi.org/10.1007/s11199-006-9018-y

Baquerín, I. S. (2017). *Masculinidades, sexualidad y género* [Acta]. Congreso internacional de la Red española de Filosofía, Madrid, España

Bringas-Molleda, C., Méndez, E., Rodríguez-Franco, L., y Rodríguez Díaz, F.J. (2016). Análisis diferencial del SRQ-R por sexo y nivel educativo en jóvenes españoles. En A. Andrés

Pueyo, F. Fariña Rivera, M. D. Seijo Martínez & M. Novo Pérez (Eds.), *Avances en psicología jurídica y forense* (pp. 11–21). Sociedad Española de Psicología Jurídica y Forense

Choi, N., & Fuqua, D. R. (2003). The structure of the Bem Sex Role Inventory: A summary report of 23 validation studies. *Educational and Psychological Measurement*, *63*(5), 872–887. https://doi.org/10.1177/0013164403258235

Choi, N., Fuqua, D.R., & Newman, J.L. (2006). Hierarchical confirmatory factor analysis of the Bem Sex Role Inventory. *Educational and Psychological Measurement, 67*, 818–832.

Constantinople, A. (1973). Masculinity-femininity: an exception to a famous dictum? *Psychological Bulletin, 80*(5), 389–407.

Cordero, A., Pamos, A., & Seisdedos, N. (2008). *NEO PI-R, Inventario de Personalidad NEO Revisado*. TEA Ediciones.

Costa, P. T., & McCrae, R. R. (1985). The NEO personality inventory. *Journal of Career Assesment 3*(2), 123–139.

Fernández, J. (2011). Un siglo de investigaciones sobre masculinidad y feminidad: Una revisión crítica. *Psicothema*, *23*(2), 167–172.

Fernández, J., Quiroga, A., Escorial, S., & Privado, J. (2014). Explicit and implicit assessment of gender roles. *Psicothema, 26(2),* 244–251. https://doi.org/10.7334/psicothema2013.219

Fernández-Rodríguez, M. A., Dema, S., & Fontanil, Y. (2018). La influencia de los roles de género en el consumo de alcohol: estudio cualitativo en adolescentes y jóvenes en Asturias. *Adicciones, 31(4),* 260–273. https://doi.org/10.20882/adicciones.1003

Ferradás, M.-del-M., Freire, C., Rodríguez-Martínez, S., & Piñeiro-Aguín, I. (2018). Profiles of self-handicapping and self-esteem, and its relationship with achievement goals. *Annals of Psychology*, *34*(3), 545-554. https://doi.org/10.6018/analesps.34.3.319781

García-Cueto, E., Rodríguez-Díaz, F. J., Bringas-Molleda, C., López-Cepero, J., Paíno-Quesada, S., & Rodríguez-Franco, L. (2015). Development of the gender role attitudes scale (GRAS) amongst young spanish people. *International Journal of Clinical and Health Psychology*, *15*(1), 61–68. https://doi.org/10.1016/j.ijchp.2014.10.004

Haig, D. (2004). The inexorable rise of gender and the decline of sex: Social change in academic titles, 1945–2001. *Archives of Sexual Behavior*, *33*(2), 87–86.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7th ed.). Prentice Hall.

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., & Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicologo*, *37*(3), 192–197.

Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. Guilford press.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kachel, S., Steffens, M. C., & Niedlich, C. (2016). Traditional masculinity and femininity: Validation of a new scale assessing gender roles. *Frontiers in Psychology, 7*.

Lenhard, W., & Lenhard, A. (2016). *Calculation of effect sizes*. Psychometrica. https://www.psychometrica.de/effect_size.html.

López-Sáez, M. A., & García-Dauder, D. (2020). Los test de masculinidad/feminidad como tecnologías psicológicas de control de género. *Athenea Digital*, *20*(2), e2521. https://doi.org/10.5565/rev/athenea.2521

Lorenzo-Seva, U., & Ferrando, P. J. (2020). *Factor* (Versión 10.10.03) [Computer software]. Universitat Rovira i Virgili. https://psico.fcep.urv.es/utilitats/factor/Download.html

Money, J., Hampson, J. G., & Hampson, J. L. (1955). An examination of some basic sexual concepts: the evidence of human hermaphroditism. *Bulletin of the Johns Hopkins Hospital*, *97*(4), 301–319.

Muñiz, J, Suárez-Álvarez, J., Pedrosa, I., Fonseca-Pedrero, E., & García-Cueto, E. (2014). Enterprising personality profile in youth: Components and assessment. *Psicothema*, *26*(4), 545–553. https://doi.org/10.7334/psicothema2014.182

Paino, S., Gutierrez, D., & Aguilera, N. (2017). Gender role attitudes amongst young spanish people. En C. Bringas & M. Novo (Eds.), *Coleccion Psicologia y Ley* (pp. 41–57). *Sociedad Española* de *Psicología Jurídica* y *Forense.*

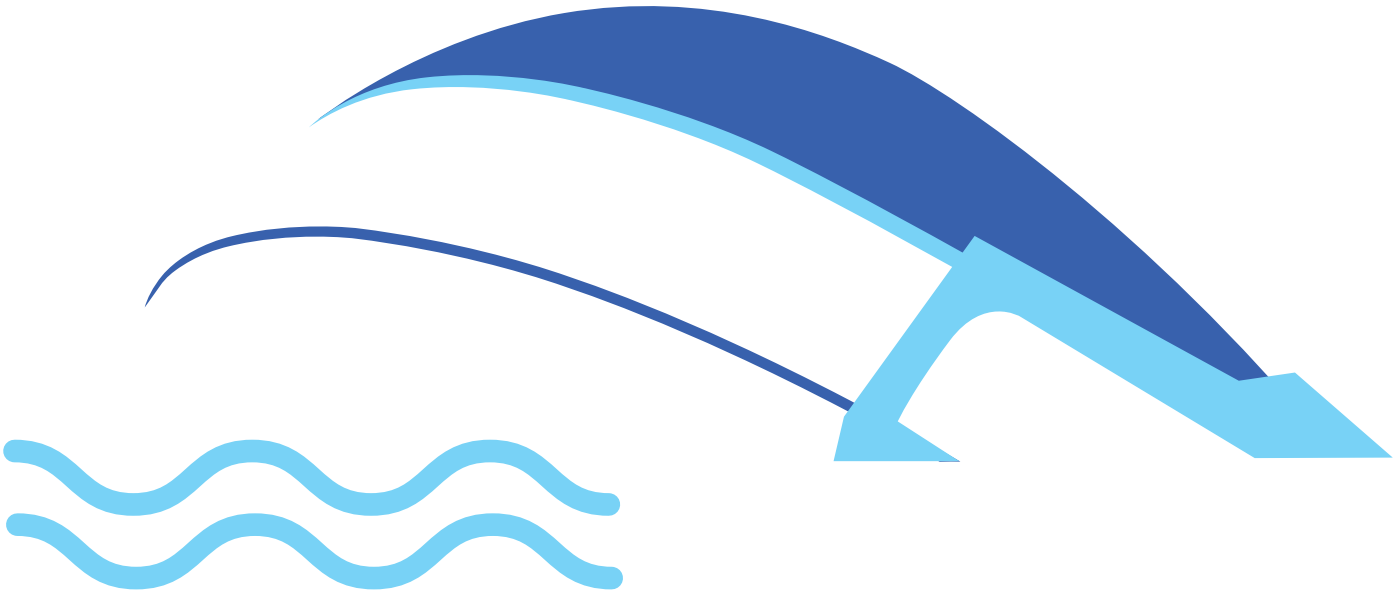
Pedrosa, I., Suárez-Álvarez, J., Pérez-Sánchez, B., & García-Cueto, E. (2010). Efecto del sexo y de la masculinidad-feminidad sobre algunos aspectos cognitivos y de personalidad. *Revista de psicología general y aplicada*, *63*(1-2), 23–32.

Postigo, Á., García-Cueto, E., Muñiz, J., González-Nuevo, C., & Cuesta, M. (in press). Measurement invariance of entrepreneurial personality in relation to sex, age, and self-employment. *Current Psychology.*

Smith, M., Sherry, S., Vidovic, V., Saklofske, D., Stoeber, J., & Benoit, A. (2019). Perfectionism and the five-factor model of personality: A meta-analytic review. *Personality and Social Psychology Review*, *23*(4), 367–390. https://doi.org/10.1177/1088868318814973.

VandenBos, G. R. (Ed.). (2015). *APA dictionary of psychology* (2nd ed.). American Psychological Association. https://doi.org/10.1037/14646-000

9th European Congress of Methodology



RESEARCHDESIGNBIGDATASTATISTICSRIGOUR
MEASUREMENTTRANSPARENCYREPLICATION

Encouraging Advance in Methodology
European Association of Methodology

EUROPEAN ASSOCIATION OF METHODOLOGY

VNIVERSITAT DE VALÈNCIA Faculty of Psychology
Department of Methodology of Behavioral Sciences

GENERALITAT VALENCIANA
Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital

aemcco

sepex
sociedad española de psicología experimental

VNIVERSITAT DE VALÈNCIA Relacions Internacionals 2.0

FUNDACIÓ PER AMOR A L'ART

AJUNTAMENT DE VALÈNCIA

VLC VISIT VALÈNCIA
www.visitvalencia.com

francisco j. bermell ELECTROMEDICINA
EQUIPOS PARA EL DEPORTE Y LA MEDICINA DEL DEPORTE