Selected topics in Psychometrics

José L. Meliá Navarro
Department of Methodology
University of Valencia
2019

www.uv.es/psicometria

Index:

Foreword

This document contains teaching materials that have been developed as part of the set of didactic resources and learning activities employed in Psychometrics, a nine-credit compulsory subject taught in the second year of the Psychology degree at the University of Valencia.

Although we already use an extensive set of classroom materials –in the form of detailed PowerPoint-like presentations developed in the past five years of teaching in English in the 'Academic High Performance' group of the Psychology degree– in this document I provide more detailed explanations for a selection of topics that usually require more thorough treatment. Examining several of the more sensitive points in psychological measurement, these ancillary notes are intended to contribute to resolving some of the most frequently asked questions and encourage critical thinking by suggesting new issues and inquiries.

Psychometrics is a particular case within psychology because it is a technical subject. Based on mathematics and especially statistics, it involves a special need for formulas, Greek symbols, and specific vocabulary. In fact, it is possible to write a psychometrics paper or a psychometrics handbook using hardly any natural language at all and using only the language of mathematics. Psychology students do not have to be afraid of this possibility. Here, my purpose is to explain things in a didactic way, without hiding certain natural complexities, but avoiding a formal presentation. However, as a technical field based on mathematical symbols and statistical concepts, psychometrics has developed its own set of psychometric conventions about how to refer to psychometric concepts, how to symbolize them, and how to express particular psychometric properties or relationships. These implicit conventions have been developed in more than a century of research papers, specialized scientific journals, and many outstanding psychometric handbooks. Although I have tried to reduce the formulas to a minimum, they are certainly unavoidable when writing psychometrics.

Part I. Psychometrics and the Scientific Knowledge

The concept of psychometrics

Psychometrics, as the name suggests, is about psychological measurement. As in any other empirical science, Psychology tries to "measure all that can be measured and render measurable all that defies measurement", to follow the old adage attributed to Galileo Galilei.

The advantages of measurement –when proper and available– are so evident that scientists rarely take the time to explain them. For now, it would suffice to say that psychological measurement is a main pillar of psychological assessment and psychological intervention.

From a theoretical point of view, Psychometrics is a technical subject that provides mathematical models for psychological measurement, supports the measurement of all kinds of psychological constructs, and contributes procedures to estimate and interpret psychological scales.

From a practical point of view, psychometrics is mainly devoted to developing the methodologies for creating and checking the quality of tests and questionnaires, and by doing so it provides a more solid base for advancing psychology as an empirical science.

There is a complex route from a psychological construct to a psychological measurement –ultimately represented by a nude number– and back again from the number to the psychology through psychological interpretations. Psychometrics is there, behind any psychological interpretation of all psychological measures.

Before discussing how psychometrics helps psychology to be a science and professional psychological practice to be soundly based on scientific knowledge, it would be useful to set off on a short journey starting with reviewing the idea of science and how to differentiate scientific knowledge from other types of discourses. You may think that this path has been previously travelled, but let me show you why these epistemological bases are so important in your psychological training.

One step back: a brief epistemological background.

All sciences are made of propositions. Theories are merely complex structures of propositions arranged in ways that allow an internal deductive consistency and some empirical coherence. Propositions, and so theories, are statements or ordered sets of statements that must be tested or proven to be true or false.

Formal and empirical sciences

Based on the criteria for truth they use to accept or reject propositions, sciences can be classified into two main groups: formal sciences and empirical sciences.

Formal sciences –such as mathematics and involving main branches such as logic or geometry– are based on a criterion of internal consistency; i.e., given certain assumptions, deductions are true if they correctly follow from the assumptions, given the rules of the system. Theorems –not diamonds! – are forever. Even if this universe finally collapses –as many cosmologists predict today– theorems will remain true. In some way, mathematics does not speak about this world, though it certainly helps us to create the most useful models for this world. Therefore, formal sciences do not need to test their propositions based on the real world: a logic or mathematical deduction can be demonstrated to be true –or rejected as false– in the darkness of a solitary desk without any data coming from the real world.

Empirical sciences encompass astronomy, geology, meteorology, marine sciences, biology, paleontology, physics, and chemistry, but also archeology, history, sociology, physiology, histology, medicine, and psychology, and many others. All these sciences speak about different areas or levels of the real world.

Speaking of the real world, all empirical sciences must be based on facts. They all need data from the real world to decide whether a proposition is true or false. They all need the support of empirical pieces of evidence. And this is the crucial point: empirical scientific knowledge requires empirical evidence, i.e., checking whether real world data support –or reject– the theories and propositions.

Therefore, what characterizes an empirical science as a *science* is not its object –the subject matter or level of the real world it researches– but the method it uses, i.e., the use of any of the multiple science methodologies, all of which are always based on checking through empirical evidence whether hypothetical propositions can be sustained. All empirical sciences share this core of the scientific methodologies as the common characteristic that defines them as sciences. Thus, the resource of empirical evidence as the criterion of truth is not what differentiates between empirical sciences. The subject, -i.e., the part, level, or perspective of the real world they analyze- is what does differentiate them.

The scientific method adopts many forms to adapt to the particular requirements of the subject under study –because, clearly, it is not the same thing to research galaxies as it is to research cells, memory, societies, or human organizations. However, any scientific methodology is always ultimately based on empirical evidence. Do we have facts that can reject the theory –usually through its observable consequences? Do these facts support the observable consequences of this theory? As far as we know, does this theory, with all its propositions and observable consequences, agree with the available facts?

The kernel of the scientific method hinges on testing whether the proposed explanations and their consequences fit the facts, usually registered as data. Of course, there is no dogmatism in this. If a careful account of the facts that could reject a proposed theory does not refute it, we provisionally accept this theory, but without any abiding attachment. All accepted theories are just provisionally accepted theories; science is always looking for new ways to explain as much evidence as possible in the simplest form.

Other kinds of valuable knowledge that are not empirical science

Before introducing the various levels on which a proposition that claims to be empirical science can lie, I would mention that empirical science does not comprise all kinds of acceptable, useful, or valuable human knowledge. As previously suggested, some sources of valuable knowledge are not empirical scientific knowledge, cannot be empirical scientific knowledge, and should not try to be. Philosophical knowledge and its many branches are outstanding examples.

For instance, think of ethics. Ethics plays a major role in human life, even in the control of scientific activities. Even though according to some ethical theories, ethical and non-ethical acts can also be judged by their consequences, ethics itself is not scientific knowledge and probably cannot be. At first glance, this seems somewhat paradoxical because today ethical principles and the decision-making of ethics committees rule many choices about the acceptance, rejection, or modification of scientific research programs for all kinds of empirical sciences. The empirical science that can or cannot be performed hinges on ethics, but the ethical ideas underlying these decisions are not a matter of empirical science.

A second example of valuable knowledge that is not empirical science is epistemology, also known as gnoseology (or gnosiology), theory of knowledge or theory of science. Because epistemology discusses the bases and methodologies of scientific knowledge, it plays a founding and developing role in any science but, again, by itself, it is not empirical science.

Something similar happens with symbolic logic and logic in general as a principal and decisive branch of human knowledge. Logic can be incarnated into philosophical knowledge but it can also be included in the formal sciences. The principles and contributions of logic support any correct deduction and any kind of acceptable reasoning. Therefore, there is no science or empirical science without logic and, furthermore, it is hard to imagine any kind of science without logic. However, again logic itself is not a matter of empirical science.

In a time when empirical sciences and technology (based on and closely related to empirical sciences) have displaced philosophy as the main type of acknowledged

knowledge, the strong dependence of empirical sciences on ethics, epistemology, and logic seems to be the revenge of philosophy, to the delight of philosophers and enthusiastic admirers.

As mentioned above, perhaps the most notorious form of valuable knowledge that is not empirical science are the formal sciences, particularly mathematics. Mathematics is a nearly magical class of knowledge whose criterion of truth does not depend on empirical facts but many times provides the best models and the best intellectual instruments for understanding empirical facts –i.e., our scientific representation of the universe– and always the tools for calculations, the matter of precise deduction, and the concrete basis of many scientific laws.

The dependence of empirical sciences on mathematics is so strong that without mathematics a vast part of these sciences would be inconceivable. Without mathematics, physics, astronomy, chemistry, and vast parts of medicine, sociology, or psychology would not exist. What is worse, without mathematics, there would be no statistics, and without statistics there would be no way to test hypotheses under probabilistic conditions –which are the usual conditions.

In general, there is no conflict between these types of knowledge and empirical sciences. In some way, in the twenty-first century, after several centuries of discussion, each branch knows its position in the tree of human knowledge, and, as I intentionally mentioned, some of this non-empirical-science knowledge is not only useful but, in some way, a prerequisite and foundation for the empirical sciences.

The 5 levels at which a proposition that claims to be empirical science may lie

In the realm of empirical sciences, every effort must be made to distinguish the different levels at which a proposition –or a set of propositions articulated as a theory– may lie from an epistemological point of view. This may be especially important in disciplines such as medicine and psychology where the consequences of assuming untestable, untested, or rejected propositions to be true may have serious effects on public health (Meliá, 1990).

Class I. Propositions that cannot be tested using empirical evidence

Not all propositions or theories are eligible for empirical science, even if their proponents wish to incorporate them into the knowledge of empirical science. Science does not judge intentions, purposes, or goals.

Some propositions just cannot be tested with empirical evidence and so are beyond the realm of empirical science. There are some supposed “theories” that cannot be tested at all because they do not clearly refer to an empirical object, i.e., they do not speak about

the empirical world in a well-defined way. Others exhibit a lack of clear definitions or do not connect their statements with observable consequences. Theories that are not precise enough, use ambiguous language, allow many interpretations, or cannot be rejected by empirical facts cannot be a matter of scientific discussion. By definition, untestable propositions cannot be possible objects for empirical science unless they define terms, establish operative definitions, and are able to be rejected by the empirical evidence. Those theories whose ambiguity allows them to explain everything, leading both to a fact and its contrary, and all facts that can be nicely explained or interpreted without the possibility of being shown to be wrong are therefore untestable and so outside the range of empirical science.

Class II. Propositions suitable for empirical science but not yet tested

Second, there are theories that might be tested and may or may not be true but have not yet been put to the test. These theories might be reasonably well-defined and able to be contrasted with the facts through empirical research (and so able to be rejected by the facts), but a lack of research keeps them from being admitted as current science.

These theories must be checked empirically by the usual procedures of the empirical sciences before we can admit them as scientific knowledge. Until then, they cannot be accepted as part of the scientific corpus and should not be used as a basis for interventions or professional treatments no matter how interesting, attractive, or verisimilar they seem.

Class III. Propositions that are currently under testing procedures

A third group consists of all those theories that are being empirically tested now. Commonly, these theories present some aspects that seem to be supported by the facts while others are not. The evidence might be complex, incomplete, and inconclusive. It is often necessary to test a theory several times from different perspectives, perhaps using several methodologies and by several research teams, to start to figure out how it fits the facts if it fits the facts. These theories being tested would need a good deal of research involving many empirical tests before they could be accepted as suitable contributions. Again, it may be inappropriate or unacceptable to use them as a basis for methods for intervention or treatment in the professional field, and only under exceptional circumstances and under a special patient agreement could experimental treatments whose testing has been advanced through previous trials be introduced into professional practice.

Class IV. Propositions currently accepted

The fourth group consists of all those theories that have been empirically tested many times and survived these tests. These theories constitute the available scientific knowledge. Strictly speaking, we will never completely know if they are true but, based on the available evidence, they may be upheld. They are our best available explanations up to now.

All tests of scientific hypotheses explicitly involve the possibility of rejecting the proposition being tested although, in fact, this is usually done the other way around, –i.e., by establishing a threshold with a known probability for the rejection of a null hypothesis.

As any student of statistics in any scientific discipline knows, the technical details of how scientific hypotheses are translated into statistical hypotheses, and how these statistical hypotheses are submitted to statistical tests is rather complex –and beyond the scope of this chapter. These really sophisticated (though not infallible) tests are the bridge between facts and theories, provisionally accepting those propositions that best fit the available evidence.

The ways of testing some empirical hypotheses are really specific, and some require specific types of research with specific mechanisms to prevent many of the errors or artifacts that challenge any research. There are many types of research designs that are useful for different kinds of hypotheses in different research settings.

Some hypotheses have to be tested using the best available research designs to guarantee the validity of the results as much as possible, especially the internal validity which is related to our primary confidence about the effects of one or more factors (independent variables) on one or more responses (dependent variables).

Outstanding examples of propositions requiring a special type of research design are those of the type "treatment X is useful for the disorder/illness/syndrome Y under circumstances C in population type P".

In the case of research related to human health, especially studies involved in testing health-related treatments such as pharmacological or psychological treatments, an especially demanding kind of research design called a randomized clinical trial is required. For instance, only after a complex research process involving many steps and including at least one randomized clinical trial, all with enough acceptable positive results, can a new pharmacological principle be approved for a specific indication by the health authorities. In the case of pharmacological substances, this process usually takes about ten years and its estimated financial cost is around €1,170,000,000 –a really impressive figure partially due to the fact that only 1.1% of research processes are successful.

Class V. Propositions already rejected by research

Finally, a fifth group is made up of a variety of theories, many of which are really appealing and fascinating, that were tested and rejected many times because they were not supported by the facts. Note that, one day, these theories were proposed, explained, and defended because somebody –often an important figure in the scientific community– believed they were plausible explanations or even brilliant solutions to the problem being researched.

This cemetery of hypotheses is not at all negligible or unimportant and should be studied in any field because it may help researchers to avoid making the same mistakes, and professionals to avoid harming or causing side effects to their patients. It is deplorable that misinformed professionals treat their patients with obsolete remedies that are known to be useless or even harmful, or simply less effective than other available treatments backed by serious research.

Not everything that seems to be a plausible explanation is true. Not everything that has an attractive, well-argued explanation is right. Instead, the opposite is true. It is common to find several interesting theories that strive to explain the same phenomenon. After a certain time, many such theories will not find the necessary support or, due to contradictory results after research, will join the cemetery of ideas that do not fit the facts.

A strongly advisable critical exercise when reading about a theory with claims of empirical science is to ask from time to time:

- Where are your data? What are the empirical pieces of evidence that support your theory?
- What real world studies have been performed to test the ideas presented?
- Which types of research designs have been used?
- Which kinds of statistical tests have been performed on data coming from which populations?

These simple questions may help us to discern the status of a theory or a proposition in empirical science. It is not that speculative thinking has no value, but it should be established to which of the abovementioned five groups each theory or main proposition belongs. The field of work of empirical science does not go beyond the empirically testable. The field of established scientific knowledge refers to already-tested propositions that have been repeatedly put to the test and not rejected by empirical evidence.

Therapeutic interventions as hypotheses

When researchers or practitioners propose a new treatment based on their knowledge and expertise, they are proposing a set of hypotheses that involve at least the following classes of propositions:

a. Disorder definition: There is a disorder/syndrome/illness/disease/condition that can be described by the following symptoms/indicators/measures/clinical characteristics;

b. Treatment definition: The treatment/therapy/intervention *consists of* the following therapeutic actions/steps/procedures involved/composition and should be administered in the following way/posology;

c. Indication statement: This treatment/therapy/intervention fits this disorder/syndrome/illness/disease/condition;

d. Success forecast: This treatment/therapy/intervention will provide better/at least the same therapeutic success as the best-established one/the commonly accepted one/a placebo;

e. Adverse effects. This treatment/therapy/intervention will produce acceptable/less adverse effects/secondary effects/contraindications compared to the best-established treatment/the commonly accepted treatment.

f. Restrictive indications: The proposed treatment/therapy/intervention will provide such results under the following general circumstances/patient properties or patient circumstances. In some cases, specific treatment instructions/posology instructions are required for different restrictive indications, such as disorder severity/age/patient condition, etc.

Type *a* propositions may be pre-assumed or commonly accepted in the field, thus making the disorder definition easier. However, in the case of psychological treatments, it is strongly advisable that type *a* propositions explicitly define how to identify the disorder, based on which standards or psychological assessment procedures and criteria. It is particularly important for treatments to clearly define the scope of their indications. This involves explicitly establishing the disorder in question (type *a* propositions) but also delimiting the circumstances under which the treatment is expected to be useful for this disorder (type f propositions): the patient's personal characteristics, disorder characteristics, and other general aspects such as comorbidity, contraindications, etc. that might affect the treatment indication.

The possible adverse effects of any treatment should also be studied. This is clear for all pharmacological treatments but it should also be taken into account in any research into psychological therapeutic treatments.

As with any other hypothesis, propositions about therapeutic interventions must be tested before they are accepted. Only therapeutic procedures, such as psychological interventions or psychological therapies, supported by empirical evidence obtained through special types of research designs specially indicated for this kind of research can be applied in common professional practice and satisfy ethical principles.

In some areas, for example, there is no conclusive evidence about some of the available interventions, so additional research is required. For instance, Barlow, Johnston, Kendrick, Polnay, and Stewart-Brown (2006), after reviewing the efficacy of group-based or one-to-one parenting programs in addressing child physical abuse or neglect, concluded that "further research is urgently needed" and "there is an urgent need for further rigorous evaluation of the effectiveness of parenting programs that are specifically designed to treat physical abuse and neglect, either independently or as part of broader packages of care". This conclusion comes from the "insufficient evidence to support the use of parenting programs to treat physical abuse or neglect", although the authors also identify that there is "limited evidence [...] that some parenting programs may be effective in improving some outcomes that are associated with physically abusive parenting". A look at systematic reviews on the effects of psychological therapies (see, for example, the Cochrane reviews) highlights the diversity in the degree of available evidence and often the need for research programs that apply randomized, controlled clinical trials.

In psychology, the *dodo bird perspective*[1] for psychological therapies is a popular, well-accepted stereotype (Beutler, 2006). It may come in different formats but basically it

---

[1] The dodo bird (*Raphus cucullatus*) is an extinct species of a large and heavy non-flying bird (up to 25 kg) that lived on the island of Mauritius until it was hunted for food to the point of extinction by Dutch colonization in the mid-17th century. Lewis Carroll, in his famous *Alice's Adventures in Wonderland* (chapter 3: "A Caucus-Race and a Long Tale"), introduced a dodo bird as a fictional character who organizes a Caucus-race in order to get dry. "What is a Caucus-race?" Alice asked. "The best way to explain it is to do it," said the Dodo. Participants have to run in a sort of a circle ("the exact shape does not matter" said the Dodo), but "they began running when they liked, and left off when they liked, so that it was not easy to know when the race was over". "But who has won"? The Dodo could not answer this question without a great deal of thought, and it sat for a long time. [...] At last the Dodo said: "Everybody has won, and all must have a prize". [...] Alice thought the whole thing very absurd, but they all looked so grave that she did not dare to laugh". [Highlights, omitting many interesting details, from *Alice's Adventures in*

establishes that all psychological therapies are more or less the same because all of them are equally effective or their efficacy is not related to the therapy itself but to the therapist's characteristics or abilities. Additionally, in the worst case, a psychological therapy may be ineffective, but never harmful. Both ideas have proven to be wrong. Not all therapies are equally effective, as can be seen by just reading review papers on almost any psychological condition. For example, "many of the earliest psychological treatments ultimately showed only limited efficacy in the clinic. Included in this group are early treatments for anxiety disorders, particularly phobias, such as Wolpe's systematic desensitization, and even early in vivo exposure-based procedures" (Barlow, Bullis, Comer, and Ametaj, 2013).

Some therapies have shown their efficacy whereas others remain untested, some are undergoing testing, and others have been shown to be ineffective. Finally, what is worse, some therapies have shown their effectiveness in harming patients, increasing their probability of suffering from what they are trying to solve or prevent, or producing adverse effects (Lilienfeld, 2007).

Because this kind of information does not seem to be very popular in some psychological training, some psychologists tend to believe that there is no evidence against any psychological therapy. Some even live in a happy wonderland where all therapies are always good and what the patient needs is just a good clinician. While many psychologists are prone to pointing out the well-known harmful and secondary effects of many psychopharmacological treatments, it appears that some of them have turned a blind eye to the possibility of secondary adverse effects from some psychological therapies.

However, it is easy to find reviews in which some therapies are backed by the evidence, whereas others are not; see, for instance, Becoña and Lorenzo's review (2001) on bipolar disorder. There are also reviews in which some therapies appear to be harmful or probably harmful. For example, after reviewing the efficacy of a single-session psychological debriefing in reducing psychological distress and preventing the development of post-traumatic stress disorder (PTSD) after traumatic events, Rose, Bisson, Churchill, and Wessely (2002) concluded: "Psychological debriefing is either equivalent to, or worse than, control or educational interventions in preventing or reducing the severity of PTSD, depression, anxiety and general psychological morbidity. There is some suggestion that it may increase the risk of PTSD and depression. The routine use of single session debriefing given to non-selected trauma victims is not

---

Wonderland by Lewis Carroll, first published in 1865 – A Book Virtual Digital Edition, 2000)

supported. No evidence has been found that this procedure is effective"; and also, "there is no evidence that single session individual psychological debriefing is a useful treatment for the prevention of post-traumatic stress disorder after traumatic incidents. Compulsory debriefing of victims of trauma should cease."

There are many other examples. For instance, "there is now a substantial body of empirical research demonstrating that efforts to avoid (or suppress) thoughts, emotions, or physiological responses actually result in increased physiological arousal, greater autonomic instability, and more stress-related symptoms, despite the desire to down-regulate arousal" (Barlow et al., 2013).

The conclusion on this point is straightforward: psychological treatments should be tested using proper research designs and the appropriate measures of the variables involved when required. A scientifically based perspective is the proper base for psychological interventions.

Operational definitions

Obtaining evidence and testing theories is often a hard job requiring many trials, the collection and analysis of field data, or the design of experiments.

To distinguish the status of a theory or hypothesis, it is imperative to differentiate the propositions of empirical content and the operational definitions of the terms they contain. As Grinnell (2018) clearly stated, an operational definition "identifies one or more specific, observable events or conditions such that any other researcher can independently measure and/or test for them." In an empirical science, the terms must be operationally defined, and the way of providing an operational definition of many psychological constructs passes through psychological measurement. For this reason, like any other empirical science, psychology has a compelling need for good psychological measurement.

Psychological measurement is one of the fundamental bases for psychological assessment and the register, analysis, and identification of therapeutic success.

Part II. Tests and Questionnaires: Some Problems Related to Classical Test Theory

Tests and questionnaires: optimal and typical performance

Most psychological measurement is carried out through the use of tests and questionnaires, a classic and well-developed way of objectively comprehending subjective variables. Tests and questionnaires are standardized ways of knowing about people, essentially by asking them questions.

The word test is more frequently used for measurement instruments that ask people questions that allow the tester, through test performance, to find out what people are able to do. This means that tests are primarily devoted to cognitive abilities, intelligence, attention, and so on; all of these are psychological variables on which examinees can pass or fail. The characteristic item on these tests can be either right or wrong, and examinees should do their best to solve them. These tests are therefore called *optimal performance* tests.

Educational exams also try to determine what people can remember or do in relation to a certain subject. Therefore, these exams are also optimal performance measures and share most of the properties and all types of psychometric analyses proposed for optimal performance tests.

When an optimal performance test is applied to determine what examinees know or are able to do, the test items do not ask them their opinion about what they know or are able to do. Instead of asking for the examinees' opinion about their knowledge, skill, or ability, an optimal performance test presents a sample of problems or questions related to the topic under measurement and encourages the examinees to solve them. That is, optimal performance tests do not yield a self-report or a self-description of knowledge, ability, or skill, but rather a test, a real test, of the matter under consideration.

It is easy to guess that, in circumstances such as an academic or work assessment where the consequences are valuable to the examinees, if we ask them whether they are the correct candidate for the job or if they have mastered an academic or job-related subject, many of them would answer affirmatively, regardless of their real opinion or their actual level of knowledge or ability.

This is a simple, rather obvious, but important and sometimes ignored lesson for any kind of psychological assessment: what is reported to the examiner by the examinee is not the same as what the examinee thinks, and what s/he thinks, based on inner self-perception, self-experience, and memory processes, is not necessarily the real state of things. Hence, in many circumstances, a self-report can be seen more as a sample of

actual behavior than as an accurate report describing real thoughts and memories or a trustworthy account of the examinee's behavior.

On the optimal performance tests, examinees could deceive the tester by intentionally producing fewer correct answers than they can actually solve but, if the test is well-designed, the probability of the examinee doing so by showing an improved result that is better than their real ability is quite small, and, in many cases, this probability can be known, estimated, and somehow controlled. If, as is usually the case in optimal performance assessment, examinees are interested in doing their best, the probability of intentionally deceiving by intentionally reducing the test score is presumed to be negligible.

Knowledge, abilities, and skills are fortunate areas. We can ask people about them in a way that avoids the blurring and distorting filters of self-perception, self-opinion, and memory bias, interested or manipulative self-descriptions, and the always misleading expectation of social desirability.

The other way to ask people is through self-report, which involves asking about opinions, preferences, attitudes, and even personality. Unfortunately, we generally do not have any way of really testing a person's true attitude. There is no way of knowing how somebody would think when faced with a certain situation, brand, service, problem, or social group. Thus, we ask for a self-report: "Do you like...?" "Which do you prefer...?" "How often do you...?"

It would probably be better if we could substitute or at least complement all these questions with direct observation of human behavior in real life situations but direct observation is almost always impossible. In many cases it would also affect actual behavior, and it is always very costly in terms of effort, time, and money. Therefore, so-called self-report questionnaires have become the universal approach for examining personality, attitudes, or opinions, and describing experiences, social situations, or social interactions.

Self-reports are susceptible to the many distortions introduced by memory, self-knowledge, and self-perception filters, as well as to intentional (and even planned) distortion introduced by the communication of self-reports to others as a deliberate act. Human communication is in itself a teleological activity. We communicate because we pursue a social goal, a purpose, or an interest. Thus, communication is rarely a pure description of our inner perception of a behavior, fact, or social situation. Instead, we build on our own experience, taking into account what we can expect from the other, what our perspective, purposes, or interests are, and how we can get the expected or desired behavior from our partners. Even many of the filters, accommodations, and distortions that we perform in everyday communication may remain unknown to us.

Tests and exams are called optimal performance because it is expected that motivated examinees show us their best performance. All kinds of personality, attitudes, opinions, and preference questionnaires are called typical performance because it is expected that a naive examinee –in some way well trained to observe him/herself, to remember his/her own opinions or behaviors, to average different thoughts, feelings, or behaviors within a period, and to tell us the truth– reports this mean, average, or typical aspect of his/her life in the matter under consideration. Of course, there is no hope of finding people like this under any circumstances, even though the types and intensity of bias may vary greatly. Therefore, typical performance is more a kind of intentionally reported image measurement than a representation of the examinees' typical thinking or behavior in their real lives. However, if these self-descriptions are useful in some psychological assessment practices, why not use them?

Despite the sharp differences between optimal performance tests and typical performance questionnaires from a psychological point of view, both share a substantial part of the psychometric core because many psychometric methodologies related to the determination of the test structure, the item and test analysis, the obtaining of norms, and many other psychometric procedures can be applied or easily adapted to both types of psychometric measures.

Although a more careful use of the word would differentiate between optimal measurement tests and typical measurement questionnaires, 'test' is sometimes used in a generic way to refer to any kind of psychological measurement. Thus, it is not unusual to speak about personality tests, or even use the term "test" as a generic class that also includes other psychological measurements that are quite different from optimal performance tests.

A test is essentially a set of questions (symbolic, worded, or manipulative) –a group of tasks to do in order to show how the person solves them or what his/her choices are.

Although in many cases manipulative or situational tests would probably provide a more like-life way of measuring, for many decades paper-and-pencil tests, and later computer-administered tests have been the dominant way of testing. All the classical and prominent tests in the areas of aptitudes or personality were developed as paper-and-pencil tests designed to be administered in a classroom-like setting –or in an office setting for those individually administered for clinical or educational purposes. This fact in itself means that tests have been developed as a symbolic representation of real life, where formal questions expressed in some kind of language try to capture real-life scenes.

Unfortunately, although many tests claim to measure cognitive or affective processes –like solving problems or emotional experiences– most focus on the final answers. Little or no reference to the cognitive or affective processes involved in the act of creating a new answer or selecting a preconfigured answer can be identified when reviewing traditional aptitude or personality tests.

An individual's answers, graded in the way the test describes, is what we call his/her test performance. Usually, the cognitive and affective processes are implicit, assumed, or inferred from the test performance by examining the scores, but they are rarely directly addressed. The common tradition of psychological test measurement is based on these kinds of psychometric inferences drawn from the test performance and referring to the underlying constructs, such as aptitudinal, attitudinal, or personality factors, which are the main objective of most psychological measurements.

Any test requires an effort to define the domain being measured, the concept and meaning of a defined construct or latent variable, and the factor or trait we are trying to measure. Additionally, most tests use closed questions –that is, questions with a limited set of answer options, predefined and offered to the examinee, whose task will be to choose the correct answer (optimal performance) or select the most representative or favorite answer (typical performance). This is a way to keep things easy that is suitable for professional purposes and appropriate for psychometric or psychological research.

Answer patterns and total scores on dichotomously scored items

Although on any psychological measurement, the set of questions focuses on a specific and relatively well-delimited area –such as a personality trait, a facet of social relationships, work-related order and safety, or a particular mental ability– and the possible answers are severely restricted to a closed format –such as multiple-choice, true-false, or Likert-like questions– the set of answers provided by any test is in itself difficult to manage.

Tests are made up of many questions, and even a short test with a dichotomous answer scheme can produce a large number of possibilities. For example, for a test made up of just 20 dichotomous questions, such as true/false, yes/no, or accept/reject, the number of possible answer patterns is astonishing. There is only one way to answer 20 noes and only one way to answer 20 yeses, and so things are easy at the extremes of the scale. However, as we approach a more balanced number of yes and no answers, the number of possible ways of answering a 20-item test increases substantially. There are 20 ways of answering one yes and 19 noes, 190 ways of answering 2 yeses and 18 noes, 1,140 of answering 3 yeses and 17 noes, 4,845 of answering 4 yeses and 16 noes, ..., 167,960 ways of answering 9 yeses and 11 noes, and 184,754 ways of answering 10 yeses and 10 noes. The other side of the possible patterns, involving more yeses than noes, shows a symmetrical structure: 167,960 ways of answering 11 yeses and 9 noes, ..., 4,845 of answering 16 yeses and 4 noes, ..., 20 ways of answering 19 yeses and 1 no, and 1 way of answering 20 yeses and 0 noes. To sum up, amazingly, for a simple 20-item yes/no test, there are 1,048,576 possible ways of answering, that is, 1,048,576 possible patterns of response –definitely a huge degree of complexity for a simple instrument.

For this reason, answers should be summarized (which in some way also means simplified) and what better way to sum up than by adding up? If every and all answers can be represented using numbers, these numbers can be summarized through a simple mathematical function, that is, by adding them.

In the simplest but most common case, this process involves two steps. First, each correct answer is represented by the number 1 –that is, a right answer on an optimal measurement test or an acceptance answer on any typical performance questionnaire is represented as “one point”. Second, the total test score is reduced to the counting or addition of points. This applies to many aptitude or knowledge tests, as well as some personality, attitude or experience tests.

After summarizing scores through the addition of points, the complexity around the answer patterns is drastically reduced. The original 1,048,576 possible ways of answering a 20-item test are reduced to 21 possible scores, from 0 (no yes answer) to 20 (all yes answers). However, it should be noted that the number of possible patterns represented by each of these 21 possible scores is quite different. Whereas scores 0 and 20 represent a unique state of answers, a score of 10 represents 184,754 possible patterns of responses, so a score of ten is really ambiguous in terms of which pattern of answers is being represented. Although rarely done in test theory handbooks, it is worth exploring the relationship between the limited number of possible total test scores –just 21 for a dichotomously scored 20 item test– and the possible patterns of answers behind these 21 scores –that is, the 1,048,576 possible response patterns for the same dichotomously scored 20-item test.

The total number of scores on this 20-item dichotomously scored test is 2^{20}=1,048,576, that is, two possible scores for each item to the power of 20 items. This number is based on the following reasoning: for the first item there are 2 possibilities, i.e., yes/no, pass/fail, accept/reject. Because the item answers are independent –that is, it is assumed that an item answer does not affect any other– for the second item there are also 2 possibilities, and the same for the third item, and for the fourth, and so on. Taking into account only two items, the number of combined possibilities is 2 x 2, that is, 4 possible answer patterns {pass, pass} {pass, fail} {fail, pass} {fail, fail}. Taking into account three items, the number of possibilities is 2 x 2 x 2 = 8. Thus, taking into account 20 items, the possibilities are 2^{20}=1,048,576. This is the number of different possible answer patterns, regardless of the likelihood of any of them.

The number of possible patterns for each total test score is obtained by applying the formula for combinations: C(m;n)=m!/[n! (m-n)!]. That is, the number of possible combinations of m elements taken in groups of n is m factorial divided by ((n factorial) times ((m minus n) factorial)).

For example, the number of patterns for a total score of 5 on a 20-item test is the number of combinations of 5 yes answers within 20 items; that is, the question can be understood as calculating how many ways we can get 5 yes or correct answers within a total number of 20 questions. This number is exactly the number of combinations of 20 elements taken in groups of 5, that is C(20;5)=20!/[5! (20-5)!)=15,540.

As the total number of possible answer patterns for a 20-item dichotomously scored test is 1,048,576, then, if all answer patterns are equally probable, a total test score equal to 5 would have a probability of 15,540/1,048,576 = 0.014786.

If all patterns are equally probable, (that is, if there is the same probability of a 1 score or a "yes" answer as a 0 score or a "no" answer to any item, and if this probability for

each item is independent of what happens to any other item), then column p in table 1 represents the probability of every total test score.

It should be pointed out that the "number of patterns" column represents the number of possible patterns with the same total test score, regardless of the probability of any of these patterns. However, column p represents the probability of each total test score, assuming that all patterns have the same probability, that is, that the 1,048,576 possible patterns are equiprobable. Of course, this is a restrictive assumption that might not be true, but it provides a simple way of modeling the simplest scenario for examinees' behavior.

Table 1. Number of response patterns for each total test score and probability of each total test score, assuming that all response patterns are equally probable, for a test of 20 dichotomously scored items.

Total Test Score	Number of patterns	Probability of the total score
0	1	0.00000095367
1	20	0.00001907349
2	190	0.00018119812
3	1140	0.00108718872
4	4845	0.00462055206
5	15504	0.01478576660
6	38760	0.03696441650
7	77520	0.07392883301
8	125970	0.12013435364
9	167960	0.16017913818
10	184756	0.17619705200
11	167960	0.16017913818
12	125970	0.12013435364
13	77520	0.07392883301
14	38760	0.03696441650
15	15504	0.01478576660
16	4845	0.00462055206
17	1140	0.00108718872
18	190	0.00018119812
19	20	0.00001907349
20	1	0.00000095367
Sum=	1048576	1

This table suggests the kind of distributions we can expect for total test scores in many circumstances. Of course, (maximum total score)/2 (=10 in this example) would not necessarily always be the most popular score, but the gradient of popularity of the various total test scores around a central value depicts the typical result for many tests.

Looking at the table, this gradient seems very clear. For example, the probability of a score of 5 is (a bit more than) 0.01; that is, more or less one test out of 100 would be expected to present a score of 5. A total score of 10 is much more likely (p=0.1762); in fact, the probability of a total score dramatically increases as it approaches the (maximum score)/2 point –that is, as it approaches a score of 10 on this 20-item test.

This table contributes some convenient uses, such as the estimation of a certain kind of confidence interval. For example, the five central categories accumulate 73.68% of the possible response patterns. This means that the probability of getting a score within the 8-12 interval is around .74. In other words, less than a quarter of the possible total test scores (i.e., 5/21) represent around three-quarters of the total possible answer patterns (i.e., .74) and, assuming equiprobable patterns, the scores from 8 to 12 would represent approximately 3/4 of the total examinees' scores.

It is easy to guess that the figures in table 1, with this nice symmetry around the center and ruled by the equiprobability of the 0 and 1 answers for each item, follow a definite law. For any independent item, the probability of a correct answer may be represented by p:

$$p_i=P(i_i=1)$$

That is, p stands for the probability P of any correct/acceptance answer ($i_i=1$) to item number i –that is, any item on the test.

Conversely, q represents the probability of any non-correct/rejection answer:

$$q_i=P(i_i=0)$$

That is, q stands for the probability P of a non-correct/rejection answer ($i_i=0$) to item number i –that is, any item on the test.

Because the number of possible correct answers (usually only one) plus the number of possible non-correct answers (usually only one for true/false items but some more, usually two, three or four, for multiple-choice items) exhausts the sample space (that is, all the item answers can be classified as either correct or incorrect) then the probability of a correct answer plus the probability of an incorrect answer is 1.

$$p_i+q_i=1$$

which means that p_i may be defined by q_i, and q_i may be defined by p_i.

$$p_i=1-q_i$$

$$q_i=1-p_i$$

For these dichotomously scored items –also called binary items– the mean or expected value is just p

$$\bar{X}_i = E(i_i) = p_i$$

and the variance is the product of p_i and q_i

$$s_i^2 = p_i \cdot q_i$$

The probability mass function *f* of this distribution, over possible outcomes k={1=correct/accept, 0=incorrect/reject}, is

$$f(k, p) = p^k q^{(1-k)} = kp + q(1-k)$$

When speaking about dichotomously scored {1, 0} optimal response items, p represents the proportion of correct answers, that is, the item difficulty.

If the probability of a correct/acceptance response is p=0.5, then q=0.5, the expected value or mean is 0.5, the variance is 0.25, and the standard deviation is 0.5. This would be the case for a random guessing response, assuming that there are only two possible answers, and both seem equally attractive to the examinee. This would also be the case for a medium-difficulty item, the kind of item that is passed by half of the cases.

An item correctly answered by half of the cases is a maximum discriminant item because its variance (0.25) reaches the maximum possible variance for a dichotomously scored item and then produces the maximum number of differences between the cases when compared in pairs. This is one reason why items with p=q=0.5 are a favorite class of items for many tests and questionnaires from a psychometric point of view.

For a two-item test, both with a 0.5 probability of a correct answer i_i={1}

$$p_1=p_2=0.5$$

the probability for the pattern i_1={1} and i_2={1} summarized as P{1,1} would be

$$P\{1,1\}=p_1p_2=0.5\bullet0.5=0.25.$$

For a three-item test, with $p_1=p_2=p_3=0.5$ the probability of {1,1,1} would be

$$P\{1,1,1\}=p_1p_2p_3=0.5\bullet0.5\bullet0.5=0.5^3=0.125.$$

Then, for a 20-item test, all items with p=.5, the probability of 20 correct answers is

$$0.5^{20}=0.00000095367$$

so, exactly 1/1,084,576. This is the probability of the only pattern with all correct responses. But what happens with all the other patterns where correct and incorrect answers are combined in a certain proportion?

In fact, when $p_i=q_i$ for all the items (which necessarily involves $p_i=q_i=0.5$ for each item), this is also the probability for any individual pattern of answers. For example, what would be the probability of {1,0} for a two-item test if $p_1=p_2=0.5$?

In this case, $q_1=q_2=0.5$, so the probability of a correct answer for the first item and an incorrect answer for the second one would be

$$P\{1,0\}=p_1q_2=0.5\bullet0.5=0.5^2=0.25$$

For a three-item test, with $p_1=p_2=p_3=0.5$, the probability of, for example, {1,0,0} would be

$$P\{1,0,0\}=p_1q_2q_3=0.5\bullet0.5\bullet0.5=0.5^3=0.125.$$

The same rule can be applied to any combination of correct and incorrect responses for an n-item test. Thus, for a 20-item test with all items accomplishing $p_i=q_i$, the probability of any pattern is

$$0.5^{20}=0.00000095367$$

And, in general, for an n-item test with $p_i=q_i$ for any item, it follows that

$$p_i^n = q_i^n = 0.5^n$$

So, if all items have the same mean or p_i value, the probability of any pattern of answers depends only on the test length n (or the number of items) and the constant p_i.

Of course, in the real world, dichotomously scored items do not always have a common and constant p_i. As p_i represents the proportion of correct answers for item i, it may be

calculated in any item as the number of correct answers A divided by the total number of cases N:

$$p_i=A/N$$

Because p_i is a proportion, it goes from $p_i=0$, which means that item i is so difficult that nobody has been able to answer it correctly, to $p_i=1$, which means that item i is so easy that everybody has answered it correctly. For this reason, p_i is known as the *index of difficulty*, or difficulty index –although the greater the p_i is, the easier the item is.

Because p_i is the index of difficulty, the classical optimal performance test, after one or two *starting items* designed to be solved by anybody, tends to arrange its items in such a way that it starts with easy items. In other words, it starts with the highest p_i values, has a more or less extended plateau where the items show pi values around 0.5, and, after that, more or less in the last third of the test, the pi indexes decrease steadily again to the minimum (though they never reach 0).

For example, an optimal performance 12-item test would ideally show a *p profile* like the following:

{1 0.9 0.8 0.7 0.6 0.5 0.5 0.5 0.4 0.3 0.2 0.1}

The first item in this *p profile* for a 12-item test is a *starting item* designed to be answered correctly by all examinees. It is useless as a measurement device because it does not help us to discriminate or differentiate the degree of an examinee's aptitude, but it may be necessary in order to show examinees that they have understood the instructions and encourage them to continue. It also helps the tester to check that everybody has understood the test instructions and is giving the answer in the correct way in the correct place.

After the starting item, the second item is a very easy one that is correctly answered by 90% of the examinees. The third, fourth, and fifth items gradually increase p_i – progressively increasing the difficulty. The following three items represent the zone where the test shows its maximum discriminant power. Each of these three central items is correctly answered by half of the cases, and so they fit the middle level of aptitude in the sample under measurement and produce the maximum number of possible discriminations or differentiations if we compare all the examinees in pairs. These three items also show the maximum possible variance for a dichotomously scored item: 0.25. After these three middle difficulty items, the rest of the test progressively increases the item difficulty. The last item shows $p_{12}=0.1$, which means that only 10% of the examinees, hopefully the best performers on the measured variable, are able to solve them.

The difficulty profile is a vector of p values, following the test order, usually represented in graphic form. Each p profile involves a q profile. For example, the 12-item p profile

{1 0.9 0.8 0.7 0.6 0.5 0.5 0.5 0.4 0.3 0.2 0.1}

involves the q profile:

{0 0.1 0.2 0.3 0.4 0.5 0.5 0.5 0.6 0.7 0.8 0.9}

Now it would be clear that the q index directly represents the item difficulty: as q increases, so does the item difficulty.

The p profile for a test can be determined after applying it to a sample. If the sample is big enough and can be considered representative of a certain population, then the sample p profile may be considered an estimator of the population p profile. This means that we can reasonably expect that this vector of p values represents the items' p values in this population.

Once the p profile is known, it can be used to estimate the probability of any pattern of answers. When p_i changes through the test items, the probability of the different response patterns also changes, as does the probability of the different possible total test scores based on these response patterns.

Imagine that we have a two-item test with a p profile {0.8 0.3}. With only two ordered items, the four possible response patterns are

{pass, pass} {pass, fail} {fail, pass} {fail, fail}

that is

{1 1} {1 0} {0 1} {0 0}

and their probabilities are, respectively

$$P\{1\ 1\}=0.8\bullet 0.3=0.24$$

$$P\{1\ 0\}=0.8\bullet 0.7=0.56$$

$$P\{0\ 1\}=0.2\bullet 0.3=0.06$$

$$P\{0\ 0\}=0.2\bullet 0.7=0.14$$

As the four events {1 1} {1 0} {0 1} {0 0} represent the whole sample space, then

$$0.24+0.56+0.06+0.14=1$$

to satisfy one of the basic axioms of probability.

These four possible events allow three possible total scores: 0, 1 and 2. In general, for a test of length n made of dichotomously (0 or 1) scored items, there are n+1 possible total test scores.

In this example, the probabilities of these three possible total test scores are:

$$P(X=0) = P\{0\ 0\} = 0.14$$

$$P(X=1) = P\{1\ 0\} + P\{0\ 1\} = 0.56+0.06=0.62$$

$$P(X=2) = P\{1\ 1\} = 0.24$$

Different p profiles would provide different distributions of the whole probability (always equal to 1) among the different total scores.

For example, a 5-item test shows the following p profile:

$$\{0.9\ 0.7\ 0.5\ 0.4\ 0.2\}.$$

What would the probability of 5 correct answers be?

The probability of the response pattern

$$\{1\ 1\ 1\ 1\ 1\}$$

would be:

$$P\{1\ 1\ 1\ 1\ 1\}=0.9\bullet0.7\bullet0.5\bullet0.4\bullet0.2=0.0252$$

Now, as the p values are not constant across the items, not all response patterns are equally probable. The probability of the response pattern

$$\{1\ 1\ 1\ 0\ 0\}$$

is

$$P\{1\ 1\ 1\ 0\ 0\}=0.9\bullet0.7\bullet0.5\bullet0.6\bullet0.8=0.1512$$

In fact, this is one of the two most probable patterns –the other is {1 1 0 0 0}, and both are equally probable because item 3 has the same probability of receiving a correct answer as an incorrect one.

Of course, the most probable patterns would be those where the easy items are passed and the difficult ones are failed. Where

$$P\{1\ 1\ 1\ 0\ 0\}=P\ \{1\ 1\ 0\ 0\ 0\}=0.1512$$

that is, the easy items are correctly answered, the difficult ones are failed, and the middle difficulty item (with p=0.5) is the same if it is passed or failed, the most improbable patterns are

$$\{0\ 0\ 1\ 1\ 1\} \text{ and } \{0\ 0\ 0\ 1\ 1\}$$

where the two easy items are failed, the two difficult ones are passed, and the middle difficulty item with p=0.5 is the same if it is passed or failed.

If the p profile is {0.9 0.7 0.5 0.4 0.2}, then

$$P\{0\ 0\ 1\ 1\ 1\} = 0.1\bullet0.3\bullet0.5\bullet0.4\bullet0.2= 0.0012$$

$$P\{0\ 0\ 0\ 1\ 1\} = 0.1\bullet0.3\bullet0.5\bullet0.4\bullet0.2= 0.0012$$

$$P\{0\ 0\ 1\ 1\ 1\} = P\{0\ 0\ 0\ 1\ 1\} = 0.0012$$

The items with p=0.5 do not affect the probability of a response pattern because they always contribute 0.5 to the product of probabilities. However, easy or difficult items with extreme p values make a difference if we consider them as passed or failed in a response pattern.

The pattern {1 1 1 0 0} is just one of the 10 ways of getting a total test score of 3 on a 5-item dichotomously scored test.

$$C(5,3)=5!/(3!(5-3)!)=10$$

Because the p value changes from item to item, the probabilities of each of the 9 remaining response patterns that also produce 3 as the total test score should be separately identified.

The case of graded response items

For a Likert scale, the procedure is not much more sophisticated. Likert scales constrain the possible answer to a limited set of graded states represented by short coined statements –for example, from "strongly agree" to "strongly disagree." Every statement is then coded with an ordinal number.

Hence, every examinee's answer showing a degree of acceptation or agreement is translated into an ordinal score. The total test score is just the addition or the average (that is, the addition divided by the number of answered items) of those item ordinal scores associated with the respondent's answers.

Because the procedure is very flexible and can be easily applied to very different psychological settings –always under the umbrella of typical performance measurements that are easy to understand and cheap to use– Likert scaling has become the most common way to measure psychological variables in some areas. It is omnipresent in social and organizational psychology for measuring attitudes, opinions, and experiences, and very frequent in other realms of psychology, such as educational psychology or clinical psychology, for obtaining student, patient or professional ratings on all kinds of assessments and self-assessments.

Likert presented his methodology as a kind of easy approach to the well-established scaling methods, though he was not against those methods. "It is feared that some will mistakenly interpret this article as an "attack" on Thurstone's methods. I, therefore, wish to emphasize in the strongest terms that I am simply endeavoring to call attention to certain problems of method, and that I am very far from convinced that the present data close the question" (Likert, 1932).

Although Likert explicitly recognizes the value of the Thurstonian scales –based on elaborate procedures imported from psychophysics– in reality, the Likert procedure implied a regressive step, abandoning all the progress in scaling methods that preceded Likert's contribution. At best, a Likert scale is a fast atheoretical method to roughly approach a scaling-based measure.

The original Likert scale has five options, including a central neutral category labeled "undecided." This original Likert scale works on a two-wing "disapproval-approval" scale, with two levels in each direction. The words following the ordinal number and describing the meaning of this possible answer are called *anchors*, and the numbers that

precede the anchors are called *anchor numbers*. The original 5-point Likert scale was: 1. Strongly disapprove; 2. Disapprove; 3. Undecided; 4. Approve; and 5. Strongly approve.

After Likert, the number of graded answers offered as response options varies from 3 to 11, but would typically be 3 to 5. An odd number of answer options usually involves the presence of a central category, defined as a neutral point and worded using anchors such as "neutral" or "middle", as well as those based on the idea of indecision. An even number usually represents an attempt to avoid undecided or evasive answers, forcing examinees to choose an answer in the positive or negative wings of the answer scale –or, perhaps in the worst case, to omit the answer.

Moreover, the original dimension "from strong disapproval to strong approval, passing through indecision as a central category" has been muted in a plethora of supposed scales designed to represent approval, frequency, degree, amount, satisfaction, preference, likelihood, etc.

It should be noted that any of these scales assumes symmetry and other convenient properties based more on lay intuition than on a sound empirical foundation. For example, the original answer scale assumes that approval and disapproval are symmetrical sides of a single dimension. This assumption involves at least three particular hypotheses: that there is the same number of degrees in each direction, that the approval and disapproval can be added on the same scale, and that the position and ordinal value of the stages preceded by the suffix "dis" are the same as those of these words without the prefix.

The presence of the central "undecided" category hinges on even more improbable assumptions. The category was originally named "undecided," but it is hard to accept that somebody who is "undecided" about how to rate an object or event occupies a category right in the center between the approval wing and the disapproval wing. It should be recognized at least that there are many ways of being undecided, some of which are related to a lack of knowledge, interest, or judgment about the evaluated object. Others may be undecided because they have a set of positive reasons and a set of negative reasons for all kinds of decisional situations, which probably means different types of thoughts and feelings about the evaluated object. Of course, after the respondent has chosen option "3", nobody knows without further inquiry what the real meaning of this convenient answer is.

The problem, of course, is not the presence of implicit assumptions that are in some way unavoidable but the lack of explicit procedures to test them. None of these assumptions is tested in the standard Likert procedure, and no testing methodology is provided or encouraged by the so-called Likert scaling. This is not surprising because the success of the Likert scale is based more on its simplicity in comparison with the Thurstonian-like scaling procedures than on a sound improvement in psychological measurement.

Likert scales are also called "summative scales" or "summated scales" because the total score for each respondent is usually obtained as the simple sum of the selected anchor numbers. A variant returns the total questionnaire score to the original anchor scale just by dividing the sum of points by the number of answered items. The latter will work if you divide by the number of items effectively answered –that is, not counting the missing responses.

If, in order to obtain the total questionnaire score as an average, the addition of points through the set of items is systematically divided by the test length –that is, by the total number of items– the result will be misleading for all the cases showing one or more missing answers, and the degree of the mistake will increase as the number of missing items increases and the test length decreases. This principle may not be valid for some optimal performance tests but it will be true for most typical performance questionnaires.

Obviously, the mean solution is a linear function of the summative solution except that, in the presence of missing answers, the denominator of the mean changes from case to case, subtly introducing into the total score the problem of the inappropriate items for the respondent under measurement, a different and undesired way of reflecting indecision, or the presence of several response sets and circumstances that stimulate the non-response behavior.

One of the most appreciated properties of graded item scales is the fact that they multiply the number of possible points on the final total questionnaire scale. Whereas a traditional yes/no 10-item scale that is dichotomously and binary scored {0, 1} produces an 11-point scale from 0 to 10, a 3-option scale with items scored from 0 to 2 will produce a 21-point scale from 0 to 20. If this three-option scale were scored as usual from 1 to 3, then the number of scale points for the total questionnaire score would also be 21, but from 10 to 30 (both included). A 4-point item scale would produce a 31-point total questionnaire score from 0 to 30 or from 10 to 40. The traditional 5-point item Likert scale produces, for a 10-item questionnaire, a 41-point total questionnaire score scale, whereas a 6-point item scale produces a 51-point total questionnaire score scale.

Therefore, in general, each point, anchor, or degree added to the item scale will produce an increase in the number of points on the total questionnaire scale equal to the test length. Increasing the total questionnaire's theoretical range –that is, the number of points on the total questionnaire scale– increases the possibility of a wider empirical range –that is, the difference between the maximum total questionnaire score and the minimum total questionnaire test score. The empirical range is a dispersion statistic associated with the variance. Therefore, increasing the empirical range increases the

possibility of a larger variance, which is usually considered a suitable property to show variability and further estimate co-variabilities between the total questionnaire score and other variables. The opportunity to enhance a 10-item questionnaire from a 21-point scale to a 41- or 51-point scale by increasing the number of anchors on each item scale is quite tempting since it is an easy and cheap way to increase the variance in the total questionnaire score.

The number of items on a test or questionnaire may be denoted as n. Scoring each item from 0 to v-1, with v being the number of anchors, and defining the total questionnaire score X as the addition of the item scores I_i,

$$X = \sum_{i=1}^{n} I_i$$

and the number of points on the total questionnaire scale or theoretical range is

$$n(v-1)+1.$$

Table 2. Number of points on the total questionnaire score scale for test lengths from 2 to 20 items and with 2 to 5 anchors.

	number of anchors			
test length:	2	3	4	5
2	3	5	7	9
3	4	7	10	13
4	5	9	13	17
5	6	11	16	21
6	7	13	19	25
7	8	15	22	29
8	9	17	25	33
9	10	19	28	37
10	11	21	31	41
11	12	23	34	45
12	13	25	37	49
13	14	27	40	53
14	15	29	43	57
15	16	31	46	61
16	17	33	49	65
17	18	35	52	69
18	19	37	55	73
19	20	39	58	77
20	21	41	61	81

The number of *possible patterns of answers* depends on the test length n –that is, the number of items– and the number of anchors v. It is exactly v^n.

For a given number of anchors v, this is an exponential function where the exponent is the number of items and the base is a constant v.

For a given test length, this is a power function where the variable number of anchors v is raised to a constant number of items n.

Table 3. Number of possible patterns of response for test lengths from 2 to 20 and with 2 to 5 anchors.

	number of anchors			
test length:	2	3	4	5
2	4	9	16	25
3	8	27	64	125
4	16	81	256	625
5	32	243	1024	3125
6	64	729	4096	15625
7	128	2187	16384	78125
8	256	6561	65536	390625
9	512	19683	262144	1953125
10	1024	59049	1048576	9765625
11	2048	177147	4194304	48828125
12	4096	531441	16777216	244140625
13	8192	1594323	67108864	1220703125
14	16384	4782969	268435456	6103515625
15	32768	14348907	1073741824	30517578125
16	65536	43046721	4294967296	1.52588E+11
17	131072	129140163	17179869184	7.62939E+11
18	262144	387420489	68719476736	3.8147E+12
19	524288	1162261467	2.74878E+11	1.90735E+13
20	1048576	3486784401	1.09951E+12	9.53674E+13

As can be expected, the number of response patterns that are able to produce each possible total questionnaire score is not the same for all total scores.

For example, analyzing a very simple case, for a 2-item questionnaire (n=2) with a 3-point answer scale (v=3) from 0 to 2 –i.e., [0 1 2]– the number of points on the total questionnaire scale is n(v-1)+1=2(3-1)+1=5 –i.e., [0 1 2 3 4]– and the total number of possible answer patterns is $v^n=3^2=9$. However, these 9 response patterns are not equally distributed on the 5 possible total scores. In other words, the number of patterns for each total score is not the same.

Table 4. Patterns of response for a 2-item questionnaire (n=2) with a 3-point answer scale (v=3) and total questionnaire score produced by each pattern.

	item 1:		
item 2:	0	1	2
0	{0 0}=0	{1 0}=1	{2 0}=2
1	{0 1}=1	{1 1}=2	{2 1}=3
2	{0 2}=2	{1 2}=3	{2 2}=4

Whereas some total scores can be produced only by one response pattern, others can be obtained from many of them.

Table 5. Total questionnaire scores for a 2-item questionnaire (n=2) with a 3-point answer scale (v=3) from 0 to 2 and the set of patterns of response that produce each total score.

Total:				number of patterns:
0 =	{0 0}			1
1 =	{1 0}=1	{0 1}=1		2
2 =	{2 0}=2	{1 1}=2	{0 2}=2	3
3 =	{2 1}=3	{1 2}=3		2
4 =	{2 2}			1
	Total number of patterns =			9

The number of patterns that may produce a total score dramatically changes from the extremes of the distribution to the center. The table 5 presents the number of patterns of response for a 2-item questionnaire (n=2) with a 3-point answer scale (v=3) from 0 to 1, In the this example, from the extreme scores 0 and 4 to the middle score 2 the number of patters increases from 1 to 3.

In any scale, there is only one way to produce either of the two extreme scores. For example, if the total questionnaire score equals 0, then necessarily all the items have to be scored 0. In the example of the table 5, to score 0 an examinee only can answer the pattern {0 0}.

For a 2-item questionnaire (n=2) with a 3-point answer scale (v=3) from 0 to 1, the only way to get the maximum score 4 is by scoring both items with the maximum anchor score, i.e., 2 showing the pattern {2 2}.

The next sub-extreme total scores, approaching the center from each wing of the distribution are in this example the total scores 1 and 3. For the sub-extreme total scores the number of patterns equals the number of items n.

For example, here there are 2 ways of getting a total score of 1. Why? Because a total score of 1 means that only one item can score 1, while the rest remain 0, and so, in this

case, with only two items, there are only two ways of producing a pattern that generates a total questionnaire score equal to 1.

The same thing happens on the other side of the distribution, for the sub-maximum score 3. Because 3 is the sub-maximum score, that is, the maximum questionnaire score of 4 minus 1, all the patterns producing this maximum minus one score must be made using the maximum anchor score for all items except one. In this case, all the response patterns that are able to produce 3 as the total questionnaire score have to be composed of the maximum anchor score v_{max}=2 for all the items except one. This exceptional item non-scoring the maximum anchor, must score the maximum anchor score minus 1, in this case

$$v_{max}-1=2-1=1.$$

Then, as this only "maximum score minus one" answer can move from item to item occupying only one position, the number of patterns that can produce the submaximal total questionnaire score of 3 also equals the number of items n, that is, 2 in this example.

For this simple 2-item questionnaire (n=2) with a 3-point answer scale (v=3) from 0 to 1, the central possible total questionnaire score is 2. With two items, there are three ways to obtain this total. First, either of the two items can be equal to 1, and so the pattern {1 1} produces a total equal to 2. Second, we might place a maximum anchor score of 2 on only one of the two items, letting the other be 0, that is, the patterns {2 0} or {0 2} also produce a total questionnaire score equal to 2.

Although the example is rather simple –researchers or practitioners rarely use a 2-item measure (though there are exceptions)– it shows the dynamic of the relationship between the total questionnaire score and the number of patterns associated with each total questionnaire score.

The distribution of the number of response arrangements that can produce a certain total questionnaire score is always symmetrical and grows fast from a frequency of 1 at both extremes to the highest frequency for the central total questionnaire score.

This means that, if all response patterns were equally probable, the probability of a total questionnaire score T would increase quickly as we approach the center of the total questionnaire scores T/2.

Although the detailed analysis becomes exponentially more complex as we increase the number of items, these general principles remain.

The following tables present the detailed analysis of a 3-item questionnaire (n=3), where each item has 3 anchors, that is v=3, scored [0 1 2]. A questionnaire of this type would have a total questionnaire score with 7 different possible values –i.e., n(v-

1)+1=3(3-1)+1=7– ranging from 0 to 6 –i.e. [0 1 2 3 4 5 6]– and a total number of possible answer patterns equal to $v^n=3^3=27$.

Table 6. The 27 patterns of response for a 3-item questionnaire (n=3) with a 3-point answer scale (v=3) and the total questionnaire score produced by each pattern.

When item 1 scores 0				When item 1 scores 1				When item 1 scores 2			
	item 2:				item 2:				item 2:		
item 3:	0	1	2	item 3:	0	1	2	item 3:	0	1	2
0	{0 0 0}=0	{0 1 0}=1	{0 2 0}=2	0	{1 0 0}=1	{1 1 0}=2	{1 2 0}=3	0	{2 0 0}=2	{2 1 0}=3	{2 2 0}=4
1	{0 0 1}=1	{0 1 1}=2	{0 2 1}=3	1	{1 0 1}=2	{1 1 1}=3	{1 2 1}=4	1	{2 0 1}=3	{2 1 1}=4	{2 2 1}=5
2	{0 0 2}=2	{0 1 2}=3	{0 2 2}=4	2	{1 0 2}=3	{1 1 2}=4	{1 2 2}=5	2	{2 0 2}=4	{2 1 2}=5	{2 2 2}=6

Table 7. The 7 possible total questionnaire scores for a 3-item questionnaire (n=3) with a 3-point answer scale (v=3, from 0 to 2), and the set and number of patterns of response that produce each total score.

Total score	Patterns when item 1 = 0			Patterns when item 1 = 1			Patterns when item 1 = 2			patterns when i1=0	patterns when i1=1	patterns when i1=2	total number of patterns
0 =	{0 0 0} = 0									1			1
1 =	{0 1 0} = 1	{0 0 1} = 1		{1 0 0} = 1						2	1		3
2 =	{0 2 0} = 2	{0 1 1} = 2	{0 0 2} = 2	{1 1 0} = 2	{1 0 1} = 2		{2 0 0} = 2			3	2	1	6
3 =	{0 2 1} = 3	{0 1 2} = 3		{1 2 0} = 3	{1 1 1} = 3	{1 0 2} = 3	{2 1 0} = 3	{2 0 1} = 3		2	3	2	7
4 =	{0 2 2} = 4			{1 2 1} = 4	{1 1 2} = 4		{2 2 0} = 4	{2 1 1} = 4	{2 0 2} = 4	1	2	3	6
5 =				{1 2 2} = 5			{2 2 1} = 5	{2 1 2} = 5			1	2	3
6 =							{2 2 2} = 6					1	1
	Total number of patterns =									9	9	9	27

Again, it is clear that the number of patterns that are able to produce a certain total questionnaire score dramatically changes from the extremes of the distribution of the total questionnaire score to the center.

In this example, the central score is $T_{max}/2=6/2=3$, and this total questionnaire score T=3 is the score with the highest number of patterns associated –this total score 3, may be produced by 7 response patterns.

If all anchors were equally probable on every item, then all 27 patterns would be equally probable. If all anchors were equally probable on every item, then for this 3-item questionnaire, the expected total questionnaire score would be 3 because the distribution of patterns is symmetrical around 3 –with the total score 3 being the central value and showing the large number of response patterns.

The discussion of how a limited set of discrete positive integers produces a certain sum is a nice mathematical problem analyzed under the topic of compositions, which is related to combinatorics.

> Following Eger (2013), an integer *composition* of a nonnegative integer T with v summands, or parts, is a way of writing n as a sum of v nonnegative integers, where the order of the parts is significant. We call the integer composition S-restricted if all the parts lie within a subset S of the nonnegative integers. In classical combinatorics, the number of S-restricted integer compositions of T with v parts is given by the coefficient of x^T of the polynomial or power series (Eger, 2013). However, these procedures, such as the generalized results and mathematical proofs provided by Eger and others, are beyond the scope of this introductory text.

In general, when a variable X_T –as a total questionnaire score or a total test score– is the sum of a set of variables X_i,

$$X_T = X_1 + X_2 + \ldots + X_i + \ldots + X_n$$

$$X_T = \sum_{i=1}^{n} X_i$$

then the mean of the sum is equal to the sum of the means:

$$\bar{X}_T = \bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_i + \ldots + \bar{X}_n$$

$$\bar{X}_T = \sum_{i=1}^{n} \bar{X}_i$$

And the variance of the sum can be obtained as the sum of all the variances of the variables involved as addends plus all their covariances:

$$s_X^2 = \sum_{i=1}^{n} s_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}^2$$

This last term is subject to the restriction:

$$i \neq j$$

That is, the variance of the composite X_T equals the sum of all the n^2 terms of the variance-covariance matrix.

When X_T is just the sum of two addends, as a total questionnaire score or a total test score composed of two items

$$X_T = X_1 + X_2$$

the general formulas simplify to the following expressions:

$$\bar{X}_T = \bar{X}_1 + \bar{X}_2$$

$$s_X^2 = s_1^2 + s_2^2 + 2s_{12}$$

because the covariance of X_1 with X_2 is the same as the covariance of X_2 with X_1:

$$s_{12} = s_{21}$$

If for a single item all anchors were equally probable, the item score distribution would be a discrete uniform distribution defined in the interval

$$[I_{min}, I_{max}]$$

where I_{min} means the minimum anchor number and I_{max} means the maximum anchor number. For example, for an item with a 3-point answer scale (v=3), scored [0 1 2], I_{min} =0 and I_{max} =2.

If the item shows a discrete uniform distribution, then

$$P(I=0) = P(I=1) = P(I=2) = 1/3 = 0.\widehat{3}$$

A discrete uniform distribution is a symmetric probability distribution where the several discrete values of the variable are equally likely to happen, and each of the v values has an equal probability 1/v.

In these distributions, I_{min} and I_{max} define the interval $[I_{min}, I_{max}]$ and are considered the main parameters.

For such a distribution, the mean is:

$$\overline{I} = \frac{I_{min} + I_{max}}{2}$$

The median has the same value as the mean:

$$Md(I) = \frac{I_{min} + I_{max}}{2}$$

The variance is:

$$s_I^2 = \frac{(I_{max} - I_{min} + 1)^2 - 1}{12}$$

And the cumulative distribution function (CDF):

$$F_{(k;I_{min},I_{max})} = \frac{k - I_{min} + 1}{I_{max} - I_{min} - 1}$$

for any k, such that

$$k \in \left[I_{min}, I_{max}\right]$$

This distribution has the particularity that, if it is known that an item shows a discrete uniform distribution defined in the interval [I_{min}, I_{max}], then its mean and variance may be calculated without knowing the empirical distribution.

For example, for a discrete uniform distributed item with a 3-point answer scale (v=3), defined in the interval [0, 2], the mean (and also the median) is (0+2)/2=1.

The variance is (((2-0+1)^2)-1)/12=0.66667 and the standard deviation is 0.81649658. For a 3-item questionnaire whose total questionnaire score is the sum of its items, the mean of the total questionnaire score would be 3, because 3 is the sum of the means of the 3 items.

If we modify the test length by adding more and more uniform distributed items [0 2], the expected total questionnaire score is just the test length because the expected mean for each item is 1.

As can be intuitively deduced from our detailed analysis of the response patterns for a 3-item test, the addition of two discrete uniform distributions is no longer a discrete uniform distribution because not all the sums have equal probability.

The sum of a few discrete uniform random variables –such as the items with v=3 and equiprobable anchors– does not produce a new uniform random variable or a normal distribution. The distribution is asymptotically normal –that is, it tends toward a normal distribution when the number of independent uniform random variables defined over the same interval tends to infinity– otherwise, it can be described by a normalized extended binomial coefficient (Grinstead and Snell, 1997).

In more practical terms, this means that if we add a large enough number of uniform random answered items, the expected result is a one-peaked symmetric distribution that increasingly resembles a normal distribution as the test length increases.

Note that to obtain this normal-like distribution for the total questionnaire score, we are adding independent uniform random answered items (that is, items where all anchors have the same probability of being chosen and items that are statistically uncorrelated) to form a questionnaire where all response patterns have the same probability. Given these results, the common psychological conjecture assuming that many attitudes, personality traits, and other variables measured using this type of questionnaire have a normal distribution, or at least a normal-like distribution, might well be an artifact resulting from the way of measuring.

Reverse-worded items

It should be noted that on Likert scales, it is not unusual to introduce some items –sometimes half of them– reverse-worded, that is, measuring the other way around. For example, on a psychological health scale, if on most of the items a high anchor means a healthy state or behavior, a reverse-worded item measures in such a way that a high anchor means an unhealthy state or behavior. These intentionally reverse-worded items should be reverse-scored before they are added to the total questionnaire score. If they were not reverse-scored, their correlations with the mainstream items would be negative, whereas after reversing the score, their correlations with the mainstream items become positive. Score reversion is an easy process based on a linear transformation.

For a reverse-worded item I_r with anchors scored in the interval $[I_{min}, I_{max}]$, the reverse-scored item or unreversed worded item I_u would be:

$$I_u = (I_{min} + I_{max}) - I_r$$

The role and positive and negative properties of the reverse-worded items have been a matter of research and discussion. Originally, introducing a combination of positively and negatively worded items within the same Likert scale was proposed as a way to reduce or eliminate some response bias, especially acquiescence and disacquiescence, as well as some forms of random guessing. However, some research does not support the contribution of reverse-worded items to the reduction or elimination of response bias, while confusion is often introduced in respondents' answers.

Item homogeneity

When item answers are the result of random guessing, with all the anchors having the same probability, we can expect the previously described results. However, items can be added to a total questionnaire score because it is expected that they all measure the same psychological construct or represent the same type of behavior. If a set of items does not measure the same thing, why add or average them to form a unique composite?

Because all the items making up the same total questionnaire score are designed to measure the same psychological variable, it is expected that they will all positively correlate with each other. As different items often represent different facts, aspects, nuances, occasions, versions, or details of the same construct, a perfect correlation or a quasi-perfect correlation is not expected but, in general, a moderate positive correlation is.

The relationships between the set of items that belong to the same total questionnaire score can be analyzed using correlations. The *item matrix of correlations* represents a first approach to understanding the patterns of relationships among several subsets of items. In general, items belonging to the same total questionnaire score are expected to show positive and moderate correlations.

From the point of view of item consistency, the scrutiny of the item correlation matrix might be complex and inconclusive. If the test length is n, then the item correlation matrix will show n(n-1)/2 relevant correlations to interpret. This number [(n(n-1))/2] is the result of $n^2 - (n + n(n-1)/2)$ where n^2 is the total number of correlations in a correlation matrix, n is the number of correlations on the main diagonal, all of which are equal to 1, and n(n-1)/2 is the number of correlations in the upper triangle matrix or in the lower triangle matrix. The upper triangle matrix and in the lower triangle matrix will always show the same results because $r_{ij} = r_{ji}$. For example, if n=5 then $n^2 - (n + n(n-1)/2) = 5^2 - (5 + 5(5-1)/2))=10$, i.e., [(n(n-1))/2] =5(5-1)/2=10.

Any particular item presents n-1 correlations with the rest of the items but none of these n-1 correlations are unique indicators for a particular item. All the elements of this vector of correlations are shared systematically with the n-1 vectors of each n-1 correlation that belongs to the other n-1 items.

From the point of view of its magnitude, some of these item correlations may seem acceptable but others may not and it is rather arbitrary to consider that a correlation has an adequate size, thus showing adequate item consistency, or an inadequate size, thus not showing enough item consistency.

The level of statistical significance may not be helpful because, if for some reason the sample size is small, it might be difficult to find statistically significant correlations. However, if the sample is large enough –the usual case in questionnaire analysis– many of the correlations appear to be statistically significant. Roughly speaking, a correlation under .2 becomes significant if the sample size is greater than 100. Most of the time our samples are going to be greater than 100, so we expect positive item correlations larger than .2.

A second way to study item consistency hinges on the relationship between the item scores and the total questionnaire scores. When Likert-like scales are analyzed, it is usually assumed that both variables (the item scores and the total questionnaire scores) can be treated as quantitative variables, and so the Pearson correlation coefficient is the usual statistic of choice. The Pearson correlation between an item and the total questionnaire (or test) score is known as the *item coefficient of homogeneity*. Of course, the value of the coefficient of homogeneity may be different for every item that makes up a scale. Although there is no strict rule to interpret these correlations, a moderate and positive correlation is expected between the item and the total scale score that the item contributes to configuring.

Item homogeneity has a foundational flaw. Because it is the correlation between the item and the total scale score, given that the total scale score is created by adding all the items, this correlation contains the item information on both sides of the Pearson coefficient, first as an isolated item, and second as part of the total test score. This problem is not serious if the scale is made up of a large number of items but it may be considered a serious contamination when the questionnaire (or test) length is short. In practical terms, this redundancy becomes less important as the number of items increases. For example, it is unimportant for scales with over 20 items but quite important for scales with only a few items (e.g., 4, 6 or 8), which are very popular in numerous areas of research or professional practice.

If a scale is made up of 4 items, all of which are equally scored with the same v anchors, in general it may be presumed that any of its 4 items contributes 1/4 to the total scale score. This means that 25% of what happens in the total scale score may depend on a certain item. The conclusion is that an item homogeneity coefficient for any item on this scale may be inflated by 25%.

The so-called corrected homogeneity coefficient is an attempt to solve the statistical contamination of the homogeneity coefficient. The bases for this new item coefficient are straightforward. If the problem comes from the presence of the correlated item again in

the total scale score, let us remove the item from the total scale score; that is, why not correlate the item with a total scale score where all the items are added except the item under analysis?

There are two practical ways of obtaining this "total scale score without a given item". The first is simply by creating a new total score by adding the rest of the items but not the item under consideration. The second is by subtracting the item under consideration from the regular total scale score. Of course, both should give exactly the same result. The second is perhaps less tedious if you are performing calculations with a calculator and perhaps a bit easier to set up if you are using a spreadsheet such as EXCEL or Numbers. Statistical packages such as SPSS and Stata have predefined procedures for item analysis so that you do not need to pay attention to the details of calculation but only understand the psychometric implications of the two coefficients –homogeneity and corrected homogeneity– in order to interpret the results correctly.

In any case, the index of corrected homogeneity for item 1 is obtained from the correlation of the item 1 score with a total test score made up of the addition of items 2 to n –but not item 1. The index of corrected homogeneity for item 2 is obtained from the correlation of the item 2 score with a total test score made up of the addition of items 1 and 3 to n –but not item 2. And the same is true for any item on the questionnaire or test.

It is true that the corrected item homogeneity solves the problem of statistical inflation due to the presence of the item on both sides of the Pearson coefficient. But it is also true that the corrected homogeneity coefficient returns the correlation with a fictitious scale, one made up of all the other items but not the item under consideration and, even worse, this fictitious total scale score of reference changes from item to item.

Again, if the number of items is large, the difference between the real total scale score containing all the items and any of these n total scale scores, each consisting of a different set of n-1 items, may be negligible. However, the effect of subtracting an item from the total scale score becomes more and more important as the test length becomes more and more reduced.

For instance, a short questionnaire made up of 4 items requires calculating four different total scores to calculate the corrected homogeneity coefficient of its 4 items. Any of these 4 different total scores is composed of only 3 items, and so any of these 4 different total scores contains a different subset with a different 75% of the scale information.

As either of the two coefficients (homogeneity or corrected homogeneity) is a perfect indicator of item homogeneity, a practical compromise solution is to estimate both for every item and interpret any difference. For example, if an item from a short scale shows a coefficient of homogeneity equal to .4, whereas its corrected coefficient of homogeneity is .1, this means that the contribution of this item to the total score is really

important, and its relationship with the total scale score is mainly a result of the statistical inflation produced by having the item on both sides of the Pearson coefficient. This item is hardly related to a simple composite made up of the other three items, so perhaps the scale composition or the role of this item within this scale should be reviewed. Perhaps the item is not part of this set. Maybe more items measuring the same psychological facet of the intended construct should be added. Perhaps there are other kinds of problems, from mistakes in the wording of the item stem to several types of response sets –e.g., acquiescence, random answering, etc.– that may distort the item answers. In any case, a more thorough qualitative and quantitative analysis of this item is required.

Item homogeneity is not the only way to study item consistency. Another approach is to estimate the item's multiple correlations with the rest of the items. Perhaps with less psychometric tradition –maybe due to the difficulty of estimating multiple correlations without the assistance of a computer when the test length is moderate to long – the multiple correlation between an item and the rest of the items is a natural solution for studying whether a single item is consistent with the rest of the scale.

Although this statistic does not require the calculation of any total scale score, as in the case of the corrected homogeneity, the main flaw is that, in fact, the set of correlated items changes from item to item. Because the item multiple correlation involves the item and the rest of the items, this "rest of the items" is a different set for each item. Again, this statistic is not expected to provide very high results. Because items often represent different facets or aspects of the same construct –but not exactly the same content– items are expected to show moderate multiple correlations with the rest of the items.

The alpha coefficient

All these solutions –the coefficient of homogeneity, the corrected coefficient of homogeneity, and the item multiple correlation– address the issue of homogeneity from the point of view of each individual item and they are considered part of the item analysis process. As important as they may be in developing a reliable and valid measurement instrument, they cannot be compared to the global approach for testing consistency based on the joint consideration of the whole set of items simultaneously.

This latter approach is well represented by the alpha coefficient, which is perhaps the most successful and popular psychometric index. From the point of view of psychological research, alpha is the most widely used way to estimate test reliability (using the general meaning of the word test, that is, a test, questionnaire, or scale) and is reported for any test applied, for any dataset, as a general test quality indicator. From the point of view of the practitioner, alpha is probably the most common way to indicate the amount of reliability necessary to make all types of patient- or client-related

decisions. From a psychometric point of view, alpha is a crossroads, a key formula that has been deduced under different assumptions with different purposes.

Here we will omit all of this rather technical information to concentrate on a view of alpha as a global indicator of test consistency for a set of items. As the set of items added together to make up a total scale score is supposed to measure the same construct, it is expected that, as a set, they are consistent, that is, that they show a combined consistency.

The coefficient alpha can be expressed as

$$\alpha = \frac{n}{n-1} \cdot \left[1 - \frac{\sum s_i^2}{s_X^2} \right]$$

In this formula, n is the test length –that is, the number of items. The second numerator is the sum of the item variances –there is one variance for each item– and the second denominator is the variance of the total scale score.

This is the most popular form of presenting the alpha coefficient, which is also called Cronbach's alpha.

The result is a number that ranges (when everything is correct) between 0 and 1.

Alpha is less than or equal to (in most cases, less than) the coefficient of reliability and for this reason is taken as an estimator of test reliability. If alpha equals 1 (a really improbable result), the test is perfectly reliable; that is, the test measures without any error of measurement. An alpha of 0 indicates an absolutely unreliable test.

In practical terms, alpha coefficients such as 0 or 1 are never observed in real practice. In general, alpha is less than the coefficient of reliability but reaches the value of the coefficient of reliability when items are at least *congeneric* measures.

In general, the higher the alpha coefficient, the more reliable the test is. For research purposes, an alpha larger than 0.6 or 0.7 is required. For professional practice, when the psychologist is to make decisions based on test scores, an alpha close to 1 may be required, while a threshold of 0.9 or even 0.95 is the usual rule of thumb.

Alpha is basically a global indicator of the consistency of the items. Roughly speaking, it shows whether the items work together, whether they run in the same direction, and whether their scores are mutually consistent. This is one of the bases of any reliability procedure, i.e., testing whether different measures –in this case, different items– that are trying to measure the same variable contribute consistent results.

The alpha coefficient and the standard error of measurement

To understand alpha as an indicator of reliability, it is worth introducing the simple absolute indicator of test reliability, the standard error of measurement. The classical test model assumes that each observed score, the simple result of an act of measurement (such as a total test score), which is usually represented as X, is an approximation to the true score, represented by T. The difference is called an error of measurement and is represented by E. An E is the realization of a random process. It may take positive or negative values, and its expectation is 0; that is, there is the same probability of random positive errors (where X overestimates the true score T) as of random negative errors (where X underestimates the true score T).

As the sum of positive errors tends to be the same as the sum of negative errors (that is, the mean of E is 0), if the sample is large enough, the expected error cannot be used as a statistic to summarize the global amount of measurement error. Rather than the mean of the error, the standard deviation of the measurement errors is used for this purpose.

The standard deviation of the measurement errors is represented as s_E and called the standard error of measurement. Roughly speaking, the standard error of measurement is a kind of estimation of the typical amount of error behind the test measurement for the standard or typical case. It is stated in the same units of measurement. For example, if we measured in meters, the standard error of measurement would also be in meters, thus making it easy to use and interpret.

The standard error of measurement is a really important contribution to the classical test model, especially because it helps us to understand that the whole score is not the true measurement, that there is always a certain error expected behind each measurement, and that we really need reliable tests to reduce the amount of error.

Although conceptually important, the formula of the standard error of measurement as the standard deviation of the errors of measurement is far from practical. It would require estimating the true scores for the entire sample and then calculating the measurement errors and their standard deviation. Fortunately, psychometricians have developed a simple formula for the standard error of measurement based on the standard deviation of the test s_X and the coefficient of reliability r_{XX} –both observable statistics readily available for any psychological measure.

$$s_E = s_X \sqrt{1 - r_{XX}}$$

Because the coefficient of reliability is a positive number between 0 and 1, that is, $0 \leq r_{XX} \leq 1$, the square root term of the standard error of measurement formula also ranges between 0 and 1. The square root term of the standard error of measurement formula is 0 when the coefficient of reliability is 1, and the square root term of the standard error of measurement formula is 1 when the coefficient of reliability is 0. Then, the standard

error of measurement is equal to the standard deviation of the total scale score when the coefficient of reliability is 0, and the standard error of measurement is equal to 0 when the coefficient of reliability is 1, regardless of the value of the standard deviation of the total scale scores.

This interpretation of the standard error of measurement as related to the coefficient of reliability is theoretically interesting because it shows both ways of estimating the test reliability (called the absolute way for the standard error of measurement and the relative way for the coefficient of reliability) and the limits of the standard error of measurement.

From this theoretical analysis we can conclude that the standard error of measurement is an absolute indicator of test reliability, expressed in the units of the total scale score, that ranges from 0 (when the reliability is perfect, that is, when the coefficient of reliability equals 1) to the standard deviation of the total scale score. The standard error of measurement reaches its maximum (the standard deviation of the total scale scores) when the measure is absolutely lacking reliability, that is, when the coefficient of reliability is equal to 0.

Of course, in practical terms there is no such thing as a perfectly reliable test or an absolutely unreliable test. These two extremes of the standard error of measurement scale are of theoretical interest only. However, this analysis shows a really important property: that the standard error of measurement is a fraction of the standard deviation of the test. In other words, the standard error of measurement can be interpreted as the part of the standard deviation of the test that is due to unreliability.

Because the coefficient of reliability is a positive number in the closed interval [0 1], the term “square root of (one minus the coefficient of reliability)” is also a positive number defined in the closed interval [0 1]. This square root term is multiplied by the standard deviation of the total scale score to obtain the standard error of measurement. This means that this square root term may be interpreted as a proportion; that is, the square root term expresses which proportion of the standard deviation of the total scale score is the standard error of measurement. This is a rather unusual but nice and revealing interpretation.

$$\sqrt{1 - r_{XX}} = \frac{s_E}{s_X}$$

This view of the standard error of measurement as the proportion of the standard deviation of the total scale score that is due to unreliability is parallel to the formal definition of the coefficient of reliability as the proportion of the variance of the observed scores due to the variance in the true scores

$$r_{XX} = \frac{s_T^2}{s_X^2}$$

often used to express the coefficient of reliability as a function of the variance in the errors of measurement

$$r_{XX} = 1 - \frac{s_E^2}{s_X^2}$$

as a result of the basic decomposition of the variance in the total score variance

$$s_X^2 = s_T^2 + s_E^2$$

dividing both sides by the variance of the total scale scores

$$1 = \frac{s_T^2}{s_X^2} + \frac{s_E^2}{s_X^2}$$

In fact, if

$$\sqrt{1 - r_{XX}} = \frac{s_E}{s_X}$$

then, squaring both terms:

$$1 - r_{XX} = \frac{s_E^2}{s_X^2}$$

which also defines the coefficient of reliability as

$$r_{XX} = 1 - \frac{s_E^2}{s_X^2}$$

Although the alpha coefficient is a lower bound of the test coefficient of reliability, it is commonly used as the estimation of the test coefficient of reliability. Using the alpha coefficient as an estimation of the coefficient of reliability means that we can substitute it in the standard of error formula

$$s_E = s_X\sqrt{1-\alpha}$$

and so, the term

$$\sqrt{1-\alpha}$$

expresses the proportion of the total scale score's standard deviation explained by the error of measurement:

$$\sqrt{1-\alpha} = \frac{s_E}{s_X}$$

Interpreting alpha as an indicator of homogeneity between test halves

Although alpha is commonly used as an estimator of the coefficient of reliability, it should be borne in mind that there are several approaches to the estimation of test reliability, and that each of them assesses a different facet of test reliability.

Classical test theory identifies three procedures for estimating the coefficient of reliability based on the cornerstone definition of parallel measurements (two measurements are parallel if they provide the same true score for every case and the same standard error of measurement for both measurements). Each of these three procedures also involves a different type of reliability.

The first, called parallel tests or parallel forms, evaluates the equivalence between forms, and the coefficients of reliability provided are called coefficients of equivalence. The parallel test procedure is based on the Pearson correlation between the total test scores of two forms of a test under the assumption that these forms satisfy the definition of parallel measurements.

The second procedure, called the test-retest method, provides coefficients of reliability called coefficients of time stability. This method is based on the Pearson correlation between the total scores on the same test applied twice.

The third classic procedure for estimating the test reliability defined according to the concept of parallel measurements is based on the relationship between halves of the test and provides coefficients of reliability identified as coefficients of homogeneity.

The two halves of a test can be identified by many procedures, the three most common of which are "first part versus second part", even-odd items, and the ad hoc method. Whichever procedure is used to identify the two halves within the test items, there are

three methods for obtaining the test reliability in this situation. That is, the test coefficient of reliability under the split-half situation can be estimated using three methods: first, the Pearson correlation between the two halves of the test, followed by a Spearman-Brown correction for the case of double length –that is, R=2r/(1+r), where R is the coefficient of reliability of the test and r is the Pearson correlation between the total scores of two separated halves. The second method involves applying the Rulon formula and the third method involves using one of the Guttman or Guttman-Flanagan formulas for the split-half situation.

The alpha coefficient can be considered a development and a better estimation of the homogeneity coefficient of reliability but homogeneity does not necessarily involve equivalence with other forms of the test or stability through many measurement attempts.

More specifically, the alpha coefficient can be demonstrated as the average of the Rulon formula applied to all the forms of splitting a test into two halves.

There are

$$\frac{C_n^{n/2}}{2} = \frac{n!}{2 \cdot (n/2)!(n-(n/2))!}$$

ways of splitting a test of test length n into two parts of equal lengths. For example, there are 126 ways of splitting a 10-item questionnaire into two halves, and any of these ways of splitting the test into two halves would produce a different Rulon result (usually slightly different, but different).

This simple fact shows how the alpha coefficient, which provides only one result for the evaluation of the homogeneity between parts, implies a big advantage. The alpha coefficient as the average of the Rulon estimations of all the ways to split the test into two halves therefore best summarizes this approach to test reliability.

Interpreting the alpha coefficient as an indicator of internal consistency

The alpha coefficient is a lower bound of the test coefficient of reliability and also the mean of all possible estimations using the split-half methodology, but it is especially known as the main indicator of internal consistency. The internal consistency, roughly speaking, is the degree to which all the items work together and are more or less closely related.

To fully understand this key concept of the alpha coefficient, we need to go two steps back to remember how the total test variance –that is, the variance of the total test

scores– can be decomposed into the sum of the item variances plus the sum of all the item covariances

$$s_X^2 = \sum s_i^2 + \sum s_{ij}$$

It should be noted that the variance in the total test score is an outstanding statistic that summarizes people's variability on the psychological variable of interest. Intuitively, it is easy to see that explaining a variable –for example, a psychological trait– means explaining how some people score high, whereas others score low. That is, explaining a variable –any variable– means explaining how it varies, and explaining how a variable varies means explaining its variance because the variance just summarizes the variable's variation. For this reason, methodologists and scientists in general, and psychologists and psychometricians in particular, are so obsessed with explaining variances –for example, by decomposing a whole variance into the parts that make it up.

Following this line of reasoning, to interpret the last two terms of the decomposition of the variance in the total test score into variance due to the items' variances and variance due to the items' covariances, we will transform all the expressions into proportions.

To interpret the last two terms as proportions of the variance in the total test score, we divide both terms by the variance in the total test score

$$\frac{s_X^2}{s_X^2} = \frac{\sum s_i^2 + \sum s_{ij}}{s_X^2}$$

and operate in the following way

$$\frac{s_X^2}{s_X^2} = \frac{\sum s_i^2}{s_X^2} + \frac{\sum s_{ij}}{s_X^2}$$

$$1 = \frac{\sum s_i^2}{s_X^2} + \frac{\sum s_{ij}}{s_X^2}$$

The latter expression shows that the total test variance can be split into two proportions: the proportion of the test variance due to the item covariances and the proportion of the test variance due to the item variances.

In this formula, the proportion of the variance in the total test scores due to the covariances among the items is the part that summarizes the idea of item consistency. A set of items is consistent if they consistently and positively covariate. The covariance is

just one of the statistics indicated to assess the degree of linear relationships between two variables.

If we isolate the term that expresses the proportion of the test variance due to the items' covariance, we get

$$\frac{\sum s_{ij}}{s_X^2} = 1 - \frac{\sum s_i^2}{s_X^2}$$

This expression makes it clear that the second term in the usual formula for the alpha coefficient is just a proportion –the proportion of the covariance in the total test scores due to the whole set of item covariances.

The coefficient alpha evaluates the internal consistency of the test and it is now a little clearer how this is reflected in the usual coefficient alpha formula.

However, to fully understand the isolated contribution of the items' consistency to the alpha coefficient formula, it is better to transform the alpha coefficient into a less frequently used form that would nevertheless be essential for a sound interpretation of the results of the coefficient alpha. This new form of alpha is called the Hoyt formula.

As alpha is defined as

$$\alpha = \frac{n}{n-1} \cdot \left[1 - \frac{\sum s_i^2}{s_X^2} \right]$$

and

$$\frac{\sum s_{ij}}{s_X^2} = 1 - \frac{\sum s_i^2}{s_X^2}$$

we can rewrite the alpha coefficient as

$$\alpha = \frac{n}{n-1} \cdot \frac{\sum s_{ij}}{s_X^2}$$

The second numerator, that is, the sum of the item covariances, may be written as a function of the average of the item covariances

$$\sum s_{ij} = n(n-1)\overline{s_{ij}}$$

which allows us to again rewrite the alpha coefficient based on the mean of the item covariances:

$$\alpha = \frac{n}{n-1} \cdot \frac{n(n-1)\overline{s_{ij}}}{s_X^2}$$

This form of alpha can be simplified into

$$\alpha = n^2 \frac{\overline{s_{ij}}}{s_X^2}$$

This latter formula is an elegant expression of the alpha coefficient and is known as the Hoyt formula.

The Hoyt formula is not really useful for a practical estimation of the alpha coefficient –it would require estimating all the n(n-1) covariances to get their average, as well as calculating the variance in the total test score.

However, the Hoyt formula is valuable for understanding how the alpha coefficient may be interpreted. It decomposes the alpha coefficient into two parts, revealing the two main factors that affect it.

The first term is the test length squared, that is, the number of test items squared. This term reveals the great importance of test length in the estimation of the test reliability. This term is always an integer number that grows as the square of the number of items – for instance, if we have a 4-item test, this factor is 16. Similarly, it is 100 for a 10-item test, 400 for a 20-item test, and 1,600 for a 40-item test. It should be noted that the influence of test length on the alpha coefficient comes through the square of the number of items, which has a really strong influence. This term has no maximum.

The second term is much more modest in its amount. In fact, the second term is a proportion, that is, a number in the interval [0 1]. This second term expresses what proportion of the variance in the total test scores is due to the average of the item covariances. That is, the second term expresses the contribution of the item consistency to the total test variance. If the items are closely related, showing high covariances, then the average of the item covariances is large, and the resulting proportion is also large. But this number, since it is a proportion, can never be greater than 1.

In fact, the proportion of the variance in the total test scores due to the mean of the item covariances must always be considerably less than 1 because the alpha coefficient is also a number in the interval [0 1], and this second term has the role of reducing the first term –an integer (usually a big one)– to the [0 1] scale of alpha. It seems a bit paradoxical at first sight but the second term of Hoyt's formula has to be increasingly

smaller as the test length increases. For example, for a 4-item test with an alpha coefficient equal to .9, the second term of the Hoyt formula is 0.05625. However, for a 10-item test with the same alpha coefficient, the second term is 0.009. In general, the proportion of the variance in the total test score due to the average of the item covariances is

$$\frac{\overline{s_{ij}}}{s_X^2} = \frac{\alpha}{n^2}$$

Why does the proportion of the variance in the total test scores due to the average of the item covariances decrease so fast as the test length increases? If we recall that the total variance can be decomposed into the proportion due to the sum of the item variances and the proportion due to the sum of the item covariances

$$\frac{s_X^2}{s_X^2} = \frac{\sum s_i^2}{s_X^2} + \frac{\sum s_{ij}}{s_X^2}$$

there are two reasons. First, as the number of items increases, the number of item variances increases, and so the term

$$\sum s_i^2$$

increases in a way that is approximately proportional to the number of items (assuming the items have similar variances) because this term can also be decomposed into

$$\sum s_i^2 = n \cdot \overline{s_i^2}$$

Second, as the number of items increases, so does the sum of the item covariances precisely by a n(n-1) factor because

$$\sum s_{ij} = n(n-1)\overline{s_{ij}}$$

The proportion of the total test variance due to the item covariances can be represented by the term

$$\frac{\sum s_{ij}}{s_X^2}$$

which is equal to

$$\frac{\sum s_{ij}}{s_X^2} = \frac{n(n-1)\overline{s_{ij}}}{s_X^2}$$

But, in the form

$$\frac{\overline{s_{ij}}}{s_X^2}$$

the influence of the number of items n(n-1) has been removed. Thus, while the number of items increases according to n^2, the proportion of variance of the total test score due to the average of the item covariances must reduce its value to accommodate its contribution to the alpha coefficient.

Incidentally, the Hoyt formula is an easy way to obtain the average of the item covariances

$$\overline{s_{ij}} = \alpha \frac{s_X^2}{n^2}$$

and from this expression it is fast and easy to obtain the sum of the item covariances, again

$$\sum s_{ij} = n(n-1)\overline{s_{ij}}$$

To summarize this point, the alpha coefficient may be interpreted as an indicator of the internal consistency of the measurement under analysis. Looking at the usual formula for the alpha coefficient, we can see that alpha depends on the proportion of the total test variance due to the sum of the items' covariance –the term that expresses item consistency– and the test length, under the form n/(n-1).

Looking at Hoyt's formula, we can see alpha as the product of a first factor associated with the test length –exactly n^2– and a second factor that expresses what proportion of the variance in the total test score can be attributed to the mean of the item covariances (this second term expresses a purer form of absolute internal consistency with no influence from test length). In any case, the influence of test length on the alpha coefficient and, hence, on the test reliability, is substantial.

Test length and reliability: the Spearman-Brown prophecy

Based on the assumption that all test items are parallel measures, the relationship between the item statistics and the total test score statistics allows us to deduce the exact relationship between the coefficient of reliability and the test length. This relationship is summarized in the Spearman-Brown formula:

$$R = \frac{nr}{1+(n-1)r}$$

where R is the coefficient of reliability of a test with f items, r is the coefficient of reliability of the same test with i items, and n is the relationship between the final and initial number of items, exactly n=f/i.

This formula allows us to estimate the coefficient of reliability of a test after increasing or reducing the test length. For example, if a 20-item test has a coefficient of reliability equal to 0.8, then if we introduce 10 additional items (all of which are parallel measures, as the original 20 items are assumed to be), the expected coefficient of reliability R is be estimated as follows:

$$n = \frac{f}{i} = \frac{20+10}{20} = 1.5$$

$$R = \frac{nr}{1+(n-1)r} = \frac{1.5 \cdot 0.8}{1+(1.5-1)\cdot 0.8} = 0.857142$$

Therefore, if we increase the test length by a factor of 1.5, the coefficient of reliability increases from 0.8 to 0.857.

The formula also works for those cases where we may be interested in reducing test length –usually for practical reasons associated with the time and cost of the measurement process in large group assessments– or when we try to estimate the change in test length required to achieve a predefined coefficient of reliability.

As a particular case, when the Spearman-Brown formula is applied to the case of the split-half method, the formula is conveniently simplified. Thus, in this case

$$n = \frac{f}{i} = \frac{2}{1} = 2$$

then

$$R = \frac{nr}{1+(n-1)r} = \frac{2r}{1+(2-1)r} = \frac{2r}{1+r}$$

Response patterns and test consistency

Items may show homogeneity and tests may show internal consistency because respondents answer the tests in a coherent way. Usually, all the items that belong to a total scale score are assumed to measure the same construct, that is, the same psychological variable, although there are also some small differences related to the aspect, facet, or perspective of the questions. It can be said that all items measure the same construct, but introduce some slight particularities.

Based on this assumption, not all response patterns are equally probable. Some patterns are highly probable, whereas others are difficult to accept as a coherent way of answering the test or questionnaire.

Some response patterns can be considered consistent with the expected results, whereas others can be considered *inconsistent response patterns* because their meaning goes against the design of the measure.

For example, let X be the total test score on a simple 4-item questionnaire, worded as a Likert scale with 5 anchors from 1 to 5. All items measure in the same direction –that is, there are no reverse-worded items– so that a score of 5 for any item means "strongly agree" and shows the maximum acceptance or agreement of the object under measurement, whereas an anchor number of 1 means "strongly disagree" and represents the maximum disagreement or rejection. The 4 items have been defined following the original 5-point Likert scale –that is, 1. Strongly disapprove; 2. Disapprove; 3. Undecided; 4. Approve; and 5. Strongly approve. The purpose of the items may be to measure a certain social object, such as, for example, the mobile phone application X for keeping in touch with friends and relatives. The basic stem for the 4 items is "Do you like application X?" All four items are variants of this basic question related to different facets of the topic. All respondents are users of these kinds of applications and have had at least a fixed number of experiences using application X. The responses of case 1 to the 4 items are respectively {1 1 2 1}, and so it is clear that respondent 1 strongly disapproves or disapproves of application X. Respondent 2's answers are {4 5 5 4}, and so it is clear that respondent 2 approves or strongly approves of application X. Respondent 3's answers are {2 3 4 3}. In this case there is a lot more indecision; respondent 3 chooses "3. Undecided" for items 2 and 4, disapproves of item 1, and approves of item 3. The response pattern of respondent 3 is more complex, around the neutral point, but is still reasonable. The first three respondents show coherent answers,

compatible with the meaning of the questions. Now, respondent 4 presents the following set of answers {1 5 1 5}. This may be a rather contradictory set of answers. Given that all questions are different ways of asking the degree of acceptance of application X, this may be an unexpected set of answers. These answers may be identified as incoherent or contradictory and probably require a detailed analysis of case 4's answers to the rest of the test –if there are other questions– or may even require an individualized interview to understand the reasons for this set of answers.

Detailed analysis of the response patterns may provide a great deal of clinical or qualitative information that may be useful to psychologists in many ways, from data control quality to the identification of unusual ways of thinking or clinical symptoms. Hence, in general, do not analyze your test scores mechanically just by creating aggregates without carefully studying the possibilities of the analysis of the response patterns.

Of course, composites are a way of summarizing the complex information of dozens or hundreds of possible response patterns but these composites should be based on item homogeneities and test consistency as well as on the analysis of response patterns from the point of view of the respondent's coherence.

Some questions on implicit assumptions and the complexity of human behavior

The basic and traditional way of approaching psychometric analysis involves scoring item anchors in a simple way and summarizing these item scores into a composite score –the total test or questionnaire score– by adding or averaging the item scores.

Many questions arise immediately from these well-accepted, simple, almost universal ways of coding and scoring.

A first concern is the degree of equivalence between answers that are scored equally. Are the answers to different items really equivalent or at least equivalent enough to be coded with the same values? For example, should the correct answer to item 1 be graded the same as the correct answer to item 20?

Bear in mind that, when, for example, we score 1 for every correct answer to any item on an aptitude test, we are assuming that all of the answers represent the same amount of whatever we are measuring –a psychological trait, a cognitive process, etc.

This assumption of "equal anchoring values" might be seen as an empirical question, that is, something that can be checked in a dataset.

Even more complex might be the question of how to score the incorrect answers. Are the wrong or missing answers simply showing a lack of aptitude or are they a valuable indication of other ways of processing or thinking?

In traditional psychological measures, the incorrect answers –no matter how rich they may be from a psychological point of view– tend to be ignored.

Tests and questionnaires are usually correct-answer-focused, and so all kinds of incorrect answers tend to be scored as 0 or ignored as missing data. However, should an incorrect answer to item 1 be graded the same as an incorrect answer to item 20? Or, looking in more detail at a set of distractors (the possible wrong answers inside one item) are all the possible types of incorrect answers equivalent enough to be scored using the same value? The use of mistakes, incorrect answers, and not-focused-on-the-trait answers should be enhanced as a source of relevant psychological information.

These questions might be seen as forms or facets of the content validity process. Do the items elicit the kind of psychological processes intended? Even in the simplest cases, this is not warranted. For example, if a simple mathematical question may be solved by guessing, then the correct answer (and many possible forms of incorrect guessing) may not represent the mathematical reasoning we are trying to measure.

These questions and many others of the same kind that can be formulated concern the relationship between the numbers and the functions we use to represent samples of human behavior (usually under the form of test or questionnaire answers) and their psychological meaning. These kinds of questions are mainly related to the process of scaling (how to assign numbers to facts in order to create a psychological scale) and the process of content validation (does every item/answer represent properly what we intend to measure? Is this set of answers –as a whole– representative enough of the full domain or construct we are trying to measure?).

Measuring is when we isolate a property in order to obtain a number representing only the amount of this property in the object under measurement, and not a mixture of this property with other properties from either the object, the setting, or the act or instrument of measurement. It may seem easy for some properties but in fact it is not. Even for a single apparently simple property such as physical length, for a precise measurement we should take into account other properties such as temperature because the temperature of the object and the temperature of the measurement instrument may affect a precise measure of length for some objects. Although isolating physical properties may be difficult and require special techniques or apparatus, isolating psychological properties might be a challenging task.

Because the second variable or intrusive factor can rarely, if ever, be eliminated from the object or from the setting, isolating the measured property from the effect of an intrusive factor (that is, for example, isolating the length measurement from the temperature effect) might be done by keeping the intrusive factor constant (for example, by performing the measure at a certain conventional temperature) or knowing the

precise function relating the main variable and the intrusive factor, registering the intrusive factor, and considering its effect from the known relationship. Unfortunately, many times, none of these options is available for all the possible intrusive factors that can affect a psychological measurement.

It is true that tests and questionnaires are applied under standardized conditions and scored using objective functions. At best, these standardized conditions seriously applied to the testing process are able to maintain constant the factors that stem from the act of measurement and many that come from the psychological tester. At best, the standardized test using closed stimuli and –optionally but usually– closed answers keeps constant all the possible intrusive factors coming from the test itself. And finally, at best, the well-defined objective procedures used for scoring the answers maintain constant any factors stemming from the processes of translating behaviors into numbers.

However, less effort has been made to avoid the possible influence of intrusive factors coming from the individual under assessment. It is known that measurement results are affected by health, stress, fatigue, motivation, and many other factors such as certain personality traits. Although it is obvious to recommend not measuring people under heavy stress, illness, or fatigue, some stress and fatigue are inherent to answering tests, especially if the results might be relevant for admittance, selection, classification, promotion, or similar situations. Being assessed is actually a stressful situation. This should be taken into account when interpreting test results. Incidentally, these limitations could be even worse for other kinds of assessment procedures, such as direct observation or interviews. In the case of direct observation and interview, not only is it impossible to isolate some individual factors but many other uncontrolled factors stemming from the psychologist or the situation can also appear.

Human behavior is a complex flow where even defining some temporal cutoffs, separating the notion of *act*, is a somewhat difficult task with no easy criteria.

Even if we are able to separate simple acts –as we try to do with the answer to a single item– it is easy to understand that even the simplest act is the result of a complex chain of psychological processes influenced necessarily by many different human properties – some under the umbrella of what we call traits, human traits, or psychological traits. Hence, acts are slices of the human-behavior-flow created under convenient but somewhat arbitrary time periods, resulting in complex sequences of simultaneous, sequential, or interconnected processes, each of which is influenced by a set of traits and other factors.

To further complicate the scenario, the way these traits and factors affect the processes underlying the acts may be quite different. Some properties might be a precondition of

the act –for example, do I feel healthy enough to start/continue with this test? Am I open enough (as opposed to too shy) to answer this question or too shy to skip it? Other factors may affect the ability to find an answer or the quality of the answer. For example, are the examinees intelligent enough to understand the instructions, follow them, and solve the problem? Do they have enough knowledge or enough motivation? Also, many kinds of social factors related to the examinee and related to the situation as interpreted by the examinee may affect the answer positively or negatively. It is clear, therefore, that virtually any individual act –and hence, any test answer– might be affected at the same time by intelligence, motivation, state of health and numerous personality traits, to mention only some of the most obvious factors involved. There is no way to measure without having these factors in play. There is no act of psychological measurement without all these factors affecting the measure to a certain degree –just as there is no length measurement without temperature. However, in psychology, we do not know the function that connects these factors to the variable we are trying to measure and we may suspect that different configurations of the situation may introduce huge changes in the relative importance of many of these intrusive factors.

Trait-related assumptions

Traits and states have a long tradition in Psychology. Traits are usually described as stable characteristics of individuals conceived in such a way that all individuals from a population share the presence of the same set of traits but differ in their amount or intensity. This is a convenient way to describe individuals: once the set of relevant traits is described and workable tests are available for them, the categorization of human behavior or human beings becomes relatively easy.

States are understood as temporary feelings, thoughts or experiences, not necessarily based on a sensorial or social experience or necessarily connected to traits. Thus, based on this distinction, an individual may experience anxiety (state) without being anxious (trait), although if you are anxious (trait), you are more likely to experience anxiety (state) when dealing with a certain situation.

This relationship between states and traits might not be obvious, especially taking into account that dozens of traits have been seriously proposed but only a few have a state counterpart.

Although it is clear that our mind changes its ability to solve difficult problems from time to time and sometimes we feel clever and sometimes dumb, as far as I know, nobody has used the term "brilliant" to describe a positive state of the "intelligence" trait. For some reason, researchers have been very conservative about proposing new states but very generous in proposing new traits.

On the other hand, an individual can be described by dozens of traits but only by one or a few states. That is, somehow the nature of the state means that the state, as a prevalent variable, tends to cover the whole situation, whereas traits, as silent properties of individuals, might be together, sharing the individual description.

The question "when does a state become a trait?" reveals that perhaps the main difference between states and traits is simply the temporal horizon of the test question – assuming that the test takers are able to correctly memorize and recall when they have experienced the content of the item.

Viewing human behavior as a continuous flow where feelings, thoughts, and experiences interconnect and continuously change, the trait-state difference may be conceived as a matter of degree.

In practical terms, a trait is a result of applying a factor analysis-like technique to a set of related answers from a large enough sample. Any set of related (usually correlated) variables will provide a set of factors. Regarding human behavior, a set of related variables appears each time we produce a group of related questions.

Questions might be related for many reasons. In general, similar questions produce similar or correlated answers. Questions may be similar due to their contents, that is, because they present similar problems or ask for similar things. But questions can also be similar because they share the same form. Of course, question similarity is a matter of degree. Some personality questionnaires present questions that are so closely related that they may be understood as formal variations of the same question. Some aptitude tests might present a delicate variation of the same contents and processes, exploring a certain domain in detail. It is easy to produce similar questions. What is not so easy is to produce a set of questions that is sufficient to sample a full pre-defined domain (usually the problem starts with the very definition of the domain).

Because it is easy to generate similar questions, both in form and content, it is easy to get new traits using any factor analysis-like technique. In fact, it is not only easy, but also economical: just one test application is enough. Simplicity and economic reasons often go hand in hand with the success of a methodological procedure. For decades, some areas of psychology seemed to be in a rush to produce more and more *à la carte* factor-analysis-based constructs. Why not if it is that easy? However, factor analysis techniques cannot identify substantive traits from a psychological point of view. As a "blind" methodological procedure, factor analysis techniques simply combine the items whose answers show correlations. This has many important consequences.

First, from the point of view of factor analysis techniques, there is no difference between causes, consequences, or covariates. If two variables are correlated enough, they tend to appear in the same factor, even if they refer to separate entities. The match, the oven, and the roasted chicken might appear in the same factor just because they appear to be

associated. Because the correlations are by themselves unable to differentiate causes, effects, and covariates, the factor analysis techniques –based on correlations– mix them up inside the same factors. For this reason, factor analysis techniques might be useless for disentangling sets of closely related causes, consequences, and covariates. Items or tests measuring different things might appear to be part of the same factor just because these things tend to be associated.

Second, the amount of correlation necessary to include two variables –or items– in the same factor is a relative matter.

Many theories based on the idea of looking for a simple structure for the human mind describe traits as orthogonal variables (that is, separate and uncorrelated). However, some theories describe a complex scenario of oblique variables, sometimes with second-order orthogonal or oblique traits, more as the result of oblique rotations to try to improve the fit of the factor structure than as the result of a sound psychological theory.

Nonetheless, both views of traits, as orthogonal or oblique entities, and also the common view of states, envisage them as additive structures where item answers describing particular pieces of behavior are summed up. Using a geometrical metaphor, all these theories conceive traits and states as straight lines where items are equal unit segments available for concatenation in any order. In other words, all of the various items are considered the same size bricks that can be put together in any order. Trait theories presume a simple straight-line geometry. Inspired in the measurement of physical properties such as weight or height, where each gram or centimeter is worth the same as any other, trait measurements assume that different complex behaviors and processes associated with the different items can be concatenated. No matter how simple it seems, it has been a good try. Often, the best theories have to start by making simple assumptions.

It should be noted that there is no way to come up with a psychological theory without making underlying assumptions, and translating behavior into numbers is no exception. "All numerical analysis of test scores rest on assumptions. The assumptions generally are false to some degree, because they treat the world as simpler than it is. 'Violation of assumptions' sounds bad, but we live with violations much of the time. We plan a trip for example, with a map that assumes the world to be flat. That could cause trouble on a long voyage, but not otherwise." (Cronbach, 1990). Remember the now classic George Box aphorism: *all models are false, but some are useful*. Explained in Box's words: "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an 'ideal' gas via a constant R is not

exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules. For such a model there is no need to ask the question 'Is the model true?'. If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is 'Is the model illuminating and useful?'" (Box, 1979). Simplicity, following *Occam's razor*, is not necessarily a bad thing. On the contrary, it should be pursued in any science. And how well do the psychological trait models do? Well, that is not an easy question to answer because there are multiple psychological trait theories, and there is a huge body of research that is difficult to summarize. All these trait theories share a common body of psychometric assumptions, but they present considerable differences in their psychological knowledge. Some of these theories can be contradictory or partially contradictory in their psychological assets. However, all in all, "the assumptions common in psychometrics work well enough most of the time. The more one knows of the assumptions, the more aware she will be of the circumstances where they lead to seriously wrong conclusions." (Cronbach, 1990)

Why should we take item difficulty into account when scoring tests?

Thurstone developed a series of methods based on psychophysical procedures to scale psychological variables without a physical counterpart, such as opinions and attitudes. These methods usually assume that a set of items represents a factor or dimension, but different items with different content could be located along the dimension, expressing different degrees or amounts of the same dimension. In some way, this methodology is fully congruent with the idea of the psychological dimension, where individuals may be identified as showing more or less of it. Regardless of whether the dimension is conceived as monopole (that is, from 0 to a positive number greater than 0, for example, from 0 to 100) or bipolar (that is, with negative and positive poles, for example from -3 to +3), the same idea of dimension involves at least a rank of states, behaviors, experiences, ideas, abilities, or feelings where different points or levels represented by different numbers are associated with different positions in these ranks. Thurstone and many others working from the psychological scaling perspective fully identify this idea that test or questionnaire items are properly scaled –namely, the position of each item on the scale is identified– and, lately, this is taken into account when measuring people.

The case for traditional aptitude or intelligence tests is paradoxical. These tests are also predicated on the idea of dimension, which certainly involves a rank of progression, usually expressed through a monopole scale –for example, from 0 to 100– and they also recognize that different items represent different points on this scale, from very easy (the class of items most people are able to solve) to very difficult. However, for classical tests that are classically scored, the usual procedures used for scoring tests ignore the item difficulties and score all items as if they represented the same point on the

dimension. As a general rule, regardless of their difficulty, all optimal measurement items are scored 1 for the correct answer.

When all the items have similar difficulty, or when the difficulty profile shows great variation (as is usually the case) but all the cases answered show patterns compatible with the difficulty scale, this way of scaling may be a reasonable approach. However, if some examinees show special patterns of answers, especially regarding the items they choose to respond to or leave unanswered, producing missing values, then taking item difficulty into account would make a difference.

In the following example, we will see how the item difficulty contributes to better scoring just by using a simple weighting scheme. Let us assume that we have a test with four optimal performance items whose p values or indexes of difficulty are, respectively, (0.9 0.8 0.2 0.1). The index of difficulty or p value is a representative statistic of classical test theory that expresses the proportion of cases that correctly answer an item. Therefore, items 1 and 2 are very easy (most people answer them correctly) whereas the last two items are difficult (most people fail them). These four items do not have the same value in terms of the aptitude under measurement but when scoring them in the traditional undifferentiated way, the total test score does not reflect these differences. If all items are scored [0 1], where 0 is the score for a wrong answer and 1 is the score for a correct answer, and the total test score is just the sum of points, then two examinees, the first with the response pattern {1 1 0 0} and the second with the response pattern {0 0 1 1}, would have the same total score of 2. However, if they are scored using a weighting scheme based on the p

$$X_T = \sum_{i=1}^{n} p_i X_i$$

their scores would be 0.9+0.8 and 0.2+01, respectively, with both scores on the p scale. The index of difficulty may provide a simple way of scaling optimal performance tests.

Validity and validation

Many old and renewed psychological theories are based on creative and imaginative thinking built from amazing in-depth interpretations of one or a few client reports – often from vague or ambiguous oral expressions generated in more or less open or unstructured interviews. Of course, many of these theories seem to be captivating, nice, and interesting psychological explanations covering the emptiness of our ignorance. They are the kind of thing lay people are prone to identify as good psychology. If the analytic discourse seems coherent because it provides a mesmeric view of the inner psychology, that seems to be enough to accept the theory. In fact, some of these theories implicitly defend coherence as a criterion for truth –as if they were mathematics!

Coherence may indeed be the true criterion in mathematics or symbolic logic, but it cannot be the only criterion in empirical sciences.

These psychological theorists do not feel the need to base their theories and interpretations on serious research. From time to time, these theories provide a valuable hypothesis for scientific inquiry. However, we should remember that there are many captivating hypotheses but only a few have been tested and reasonably non-rejected. As Cronbach (1990) said, sometimes "deeper interpretations rest on complex theories about the wellsprings of behavior, and few such theories have been substantiated."

Faced with these pseudo-scientific views of psychology –pseudoscientific means not scientific, let us be clear– psychometrics defends the humble validation of the test interpretations. That is, we cannot believe a test interpretation just because it is deep, captivating, or coherent. Test inferences must be validated –which means that test interpretations must be substantiated in previous research. Test interpretations must be empirically tested, verified in the court of empirical research, and based on real-world data.

Selecting appropriate tests

Psychologists have the main responsibility for the tests they apply to their clients. Knowing the available tests, carefully reading test reviews and test manuals before test selection, and carefully following test instructions are some of their main duties.

There are many lists of do's and don'ts for psychological testers. One of the most well-known is the list for selecting appropriate tests as a responsibility of test users from the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988), published in the Lee J. Cronbach handbook *Essentials of Psychological Testing* (1990). The main general idea is that test users should select tests that meet the intended purpose and are appropriate for the intended test-taking group. Appropriate means that the test instructions, the level of difficulty (if it applies), the validation evidence, and the score interpretation tables (such as percentile tables) correspond to the language, age, educational level, and other main characteristics of the group being measured.

According to the Code of Fair Testing Practices in Education, test users should:

- -Define the purpose of testing and the main characteristics of the group to be tested (mother language, age, gender, educational level, special needs or special characteristics, etc.)

-Review the available information about tests for that purpose and population and select the test or tests that fit best. Read independent evaluations of the tests. Examine specimen sets, disclosed tests, or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test. Look for evidence supporting the claims of test developers –especially those related to the kind of score interpretation intended for the group under measurement.

-Ascertain whether the test content and normative group(s) or comparison group(s) are appropriate for the intended test takers.

-Carefully study the test manuals and other materials provided by the test developers or the test publishers. Sometimes, after these readings, psychological testers should discard a preselected test –for lack of concordance with the intended group, incomplete information, or other possible issues.

Part III. Basic issues of Item Response Theory

Item response theory and test theory

Psychometrics, the field related to psychological measurement, has developed into two main branches: *the theory of measurement and psychological scaling*, on the one hand, and *test theory*, on the other.

Test theory can be subdivided into two large theoretical bodies: *classical test theory* and *item response theory*. The latter family of models is representative of a more general class of models called *analysis of the latent structure*.

Classical test theory can be seen as a historical antecedent of item response theory, contributing a common background on test scoring and analysis and many of the issues that item response theory tries to solve. However, classical test theory cannot be considered obsolete or surpassed by item response theory. Although item response theory hardly makes use of the results of classical test theory, it does not contradict or invalidate them (Lord, 1980). On the contrary, under certain assumptions, there are clear connections between the statistics and parameters of classical test theory and those of item response theory.

As Steinberg and Thissen (1996) point out, item response theory "is not really a theory; it should be called a collection of statistical models and methods to make sense of data in the context of psychological measurement" (p. 832).

A model is a representation of a real system that represents how we believe that a set of factors or variables involved in this real system interact. A mathematical model translates these beliefs into a mathematical form, usually using mathematical functions (Lawson and Marion, 2008).

All models involve assumptions. Item response theory, as a collection of models and related methods, involves a series of basic general assumptions, mostly common to other test theories:

(1) Any test is a standardized device that attempts to measure individual differences in an unobservable construct, generally symbolized by the Greek letter θ in the realm of item response theory.

(2) This underlying unobservable latent variable θ that we intend to measure can be inferred from the covariation of the responses to the items.

(3) The purpose of item response theory is the elaboration of an item response model that statistically accounts for the likelihood of the subjects' responses given θ (for example, the probability of a correct answer given the respondent's position in the trait).

Before introducing the item response models, it is convenient to frame them in a more general class of models called latent structure models.

Latent structure models: concept and classes

Statistical models in which the dependence between a set of observed variables can be explained (and, therefore, suppressed statistically) by introducing one or more unobserved variables are called *latent structure models* (Andersen, 1980).

If the dependency structure among a set of variables can be described by their common dependence on an unobservable variable θ, we say that θ is a latent variable. Given θ, the observable variables are independent of each other. In this way, the concept of local independence defines the concept of latent variable.

In practice, most latent structure models imply that the latent variable θ represents a parameter of the cases, so that for each individual in the sample there is an associated value of the latent variable θ_j that must be estimated. This latent variable (for unidimensional models) or these latent variables (for multidimensional models) express individual differences that are used to explain the dependencies between response variables under certain conditions that characterize each model.

The statistical analysis of a latent structure model is called latent structure analysis and can present two forms depending on whether the latent variable (or variables) is considered discrete or continuous. If the latent variable is assumed to be discrete, the model is called a *latent class model*. If the latent variable is assumed to be continuous, it is a *continuous latent structure model*, known in the framework of psychometrics as an *item response model*.

Measurement models, that is, those statistical models that allow functional relationships between their parameters and include measurement error, can be classified according to the continuous or discrete nature of the observed variable (or variables) and the latent variable (or variables), producing two axes that give rise to four large families of models, according to the following classification by Hershberger (1994).

Table 8. Hershberger's classification of measurement models by the discrete or continuous nature of the observed and latent variables.

	Manifest Variables	
Latent Variables	Continuous	Discrete
Continuous	Factor Analysis	Analysis of latent trait
Discrete	Analysis of latent profile	Analysis of latent class

Assuming that the concept of "latent structure" refers to any measurement model with latent, continuous, or discrete variables –that is, virtually any measurement model– then the four cells in the previous classification represent several types of *latent structure analysis*.

According to Traub and Lam (1985), the objective of latent structure analysis is to obtain from the observed response patterns of a sample of examinees:

(a) estimates of the quantities that characterize the items on the test or

(b) estimates of the quantities that characterize the examinees or

(c) estimates of both.

The estimates of the quantities that characterize the items are called item parameters. The estimates of the quantities that characterize the examinees are the latent traits; only one latent trait for unidimensional models is expressed as θ.

All these quantities are the parameters of the function to estimate the conditional probabilities

$$P_j(x_{ij} = x \mid \theta_i).$$

These conditional probabilities are what the item response functions intend to estimate. The graphic representation of these functions is the item operative curve (Samejima, 1998) or item characteristic curve.

In fact, these conditional probabilities or their operative curves are more obvious concepts than they may seem. The item operative curves can be defined as the conditional expectations of the item score $E(X_i \mid \theta)$ (Sijtsma and Junker, 1996).

For binary items, the conditional probability can be written as

$$E(X_i | \theta) = P(X_i = 1 | \theta)$$

i.e., the probability of a correct answer to a certain item i given the θ level of the respondent.

For polytomous items, the conditional probability can be written as

$$E(X_i | \theta) = P(X_i = x | \theta)$$

which has a similar interpretation, i.e., the probability of a certain answer x to the item i given the θ level of the respondent.

The expected score is an 'excellent option' for describing the respondent's behavior due to the trait (Sijtsma and Hemker, 1998). Thus, it is clear that item response models can be understood as regression models (non-linear, usually, though not necessarily logistic) with one (one-dimensional models) or more (multidimensional models) latent variables.

It was precisely the verification that the factorial analysis was inadequate for observable binary responses that led Lazarsfeld (1950) to consider other options for analyzing latent structure. However, whereas the traditional objective of linear factorial analysis has been to establish the structure and number of factors, latent trait theory, generally non-linear, pursues the measurement of a predefined number of traits.

McDonald (1989) drew the connections between the factor analysis models and the item response models: "the oldest and best known latent trait model is the linear analysis of common factors, and there is no reason to distinguish between latent and common factors [...] The common linear factor model is, however, appropriate for quantitative test scores rather than responses to discrete items" (McDonald, 1989, p.206). All these models are characterized by referring to latent variables, and the concept of a latent variable is necessary and sufficiently defined by the local independence.

However, it is common to classify the latent structure models by considering only those with discrete manifest variables, so that latent structure analysis is considered a broader set of models that includes item response theory and latent class analysis. This convention is the most widespread in the literature (Langeheine and Rost, 1988).

Because item response theory assumes a latent, underlying, and continuous feature of an unobservable nature, item response theory can be classified as one of the two major branches of the latent structure models or latent structure analysis for discrete observable variables. According to Traub and Lam (1985), "a distinction can be drawn between two types of latent structure models, those of the item response theory and

those of the latent class analysis. The distinction lies in the assumed nature of the distribution of θ in the population of examinees. This distribution is taken to be continuous for the item response theory and discrete for the latent class analysis. Thus, the item response theory is considered appropriate for the psychological tests of ability or attitude, in which the examinees of a population are seen as continuously distributed over the underlying latent variable. The latent class analysis, on the other hand, is appropriate for those situations in which the examinees are viewed as belonging to only two, or at least very few, different groups." The latter might be the case, for example, on mastery tests. This distinction has been implicitly recognized by Lord and Stocking (1988), who stated that "the item response theory falls within the general class of latent trait models".

Item response theory refers to the set of models that relate one or more continuous latent traits θ to the probability of a certain response to an item

$$P(X_i = x|\theta)$$

maintaining the assumption of local independence and describing this relationship as a function, usually logistic, of one or more parameters. The item response model accounts for the observed covariation, which can be expressed by the fundamental principle of local independence (Steinberg and Thissen, 1996).

For a long period, from the work of Lazarsfeld (1950) to the beginning of the 1980s, what we now know as item response theory was called latent trait models or latent trait theory. These are more descriptive labels from a taxonomic point of view and were included in the classification by Hershberger (1994). The current label, item response theory, emerged with force after the publication of the book by Lord (1980) *"Applications of Item Response Theory to Practical Testing Problems"*. Now it is more common to keep the term latent structure models for the general case, and to distinguish latent class analysis when θ is discrete and use item response theory or latent trait analysis when θ is continuous.

In turn, within item response theory, the one-dimensional and multidimensional models can be distinguished according to whether they consider a single latent trait θ or multiple latent traits. Within each of these categories, it is still possible to distinguish models designed for dichotomous items and models designed for polytomous items. Additional distinctions can be made depending on the type of function described by the item characteristic curve and the number of parameters considered in the model.

Most of the traditional work in item response theory has been carried out with one-dimensional models for dichotomously scored items in a parametric frame and with logistic functions. However, in recent decades, item response theory has experienced a process of expansion and generalization towards other types of items and classes of

models, developing models for polytomous items, multidimensional models, and non-parametric models, which have received growing interest (Samejima, 1998).

Latent class analysis

Latent class models are closely linked to item response theory, with which they share not only the concept of explaining discrete variables with latent variables under the principle of local independence but also similarities in some estimation methods.

Latent class models were introduced by Lazarsfeld (1950) and Lazarsfeld and Henry (1968). In these models, the sample can be grouped according to a finite discrete number M of latent classes, under the basic assumption of the conditional independence of the variables observed, given the latent variable.

The analysis of latent class assumes the existence of a latent categorical variable that explains the relations between a set of categorical manifest variables (Langeheine, 1988). Once the level of this unknown latent variable is given, the manifest variables are independent. A set of conditional probabilities describes the relationship between the manifest variables and the latent class, expressing the probability of belonging to an observable class given the latent class. The probabilities of the latent class specify the probability that an observation will fall on each level of the latent variable (van der Heijden, Dessens and Bockenholt, 1996).

Latent class analysis is a general class of models in which different subclasses can be distinguished. For example, Clogg and Goodman (1984) developed the so-called simultaneous latent class analysis, which consists of the application of latent class analysis to a set of multidimensional contingency tables, defined by a grouping variable. The model admits constrictions of homogeneity between the different groups to look for simpler solutions. Dayton and Macready (1988) and van der Heijden, Dessens and Bockenholt (1996) extended this class of models, allowing the estimation of explanatory variables that are continuous, quantitative, and possibly different for each observation class. This extension of the simultaneous latent class analysis models that reincorporate continuous variables exemplifies how the development of the models frequently leads to mixed models that in some way do not fit the general introductory schemes.

Latent class analysis is usually solved by one of these two methods, either by the Newton-Raphson algorithm or by the EM (Expectation Maximization) algorithm (Mooijaart and van der Heijden, 1992). The former is used, for example, by Haberman in his LAT programs (Haberman, 1978) and NEWTON (Haberman, 1988), where latent class analysis is conceived as a log-linear model. The latter is applied in the so-called iterative proportional scaling of Goodman (1974; 1979), programmed in MLLSA (Eliason, 1988) or PANMARK (van de Pol, Langeheine and De Jong, 1989) and LEM (Vermunt, 1997).

A general outline of the position of item response theory

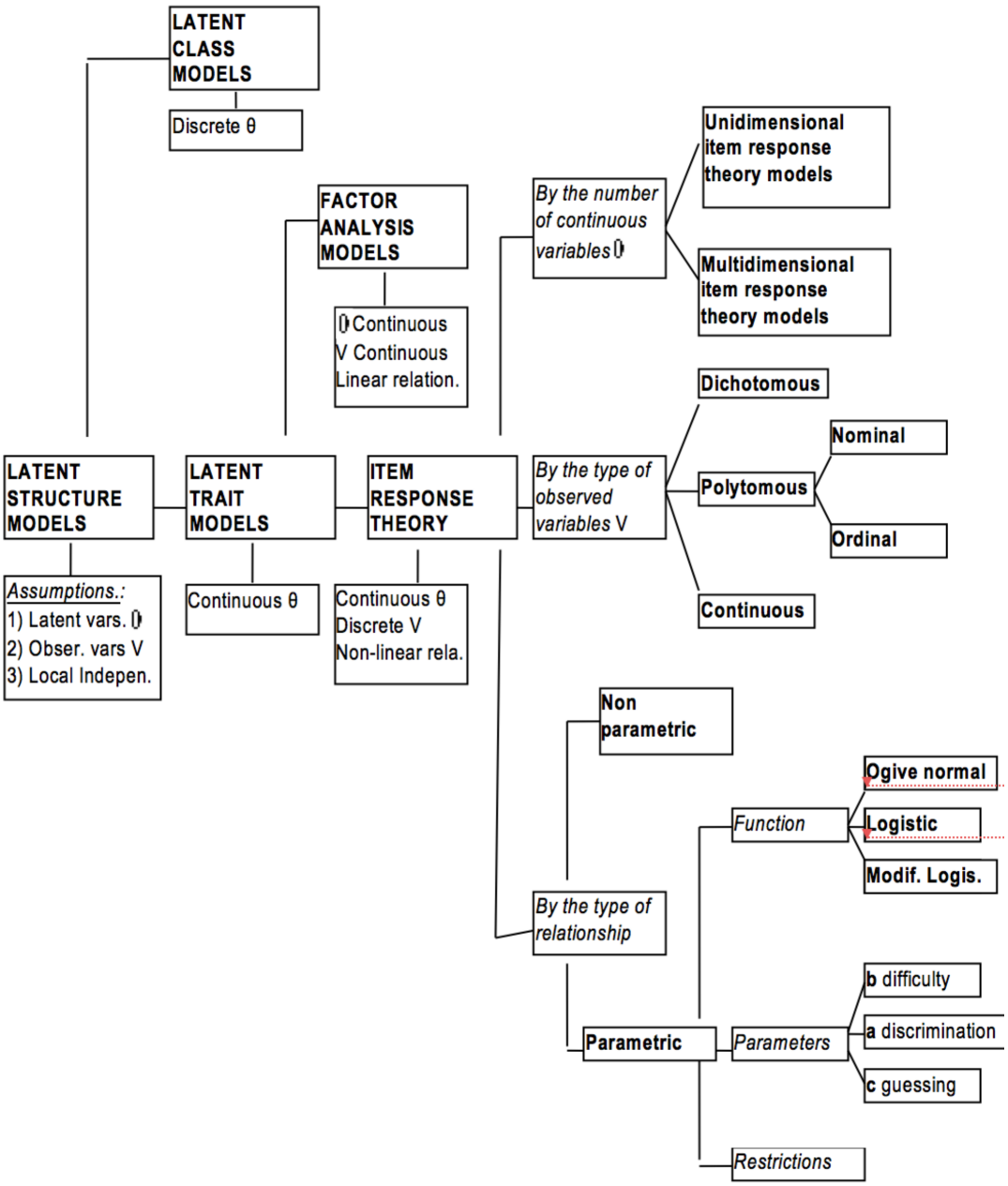
Considering various contributions, the following figure shows a scheme identifying the position of item response theory in the more general framework of models of latent structure analysis.

Latent structure analysis is the general name for a set of psychometric models designed to deal with test scores under two interrelated assumptions: the existence of one or more latent variables and the statistical independence of the subjects' responses to the items conditioned by the latent variables.

Latent structure models can be classified into latent class models if the latent variable is considered discrete (nominal or ordinal), forming classes and latent trait models when the latent variable is assumed to be continuous (Langeheine and Rost, 1988).

In the latent trait models, the observable variables can be continuous or discrete but always related to latent variables, which are considered continuous, such that the latent variables explain the covariation between the observed variables. In other words, if the effect of the latent variables on the observable variables is removed, then the observed variables are no longer correlated. Thus, the observable variables, usually subjects' responses to items or sets of items, show local independence if the effect of the latent variables is partialized.

Figure 1. General classification of the latent structure models.



If the observable variables are continuous and the relationship between observable and latent variables is linear, we are dealing with a class of models traditionally known as factor analysis.

The models of item response theory are latent trait models where, in general, the observable responses are considered discrete and the relationship between observable and latent variables is non-linear.

Item response theory models are a class of latent trait models where the observed variables are discrete, the latent variable or variables are continuous, the relationship between the two is non-linear, and the principle of local independence rules this relationship. In fact, the principle of local independence is what allows the definition of these latent variables.

Given that local independence is a property inherent to this class of models, these are the three elements that define them: the class of observable responses, the number of continuous latent variables (traits, aptitudes, or abilities), and the kind of relationship between observable and latent variables.

Regarding the class of observable responses, traditional item response models have been developed for binary items, that is, items dichotomously scored. There are also models for polytomous items –discrete observable variables with more than two possible answers, either nominal or ordinal– and for continuous variables. In this third case, both the observable variables and the latent variable are considered continuous, but the relationship between the two and the estimation methods are those of item response theory.

Regarding the number of latent variables, the models can be classified into unidimensional models, a category in which all the traditional models can be included, and multidimensional models, where 2 or more latent variables are considered simultaneously.

The kind of relationship established between the latent and observable variables can be (1) parametric when its description can be exhausted by a set of parameters that determine a function, or (2) non-parametric when the relationship between the observed variables and the latent variables cannot be reduced to a set of parameters.

If a parametric function is determined, the class of function that relates the continuous latent variables and the observable variables, generally discrete, can be classified by adopting three criteria: (1) the mathematical form of the function, (2) the number and class of parameters, and (3) the restrictions that are imposed.

The main classes of the mathematical form of the function are: (1) the normal cumulative ogive function, (2) the logistic function, and (3) the modifications of the

logistic class that no longer belong mathematically to the logistic class, such as Birnbaum's popular function with three parameters.

With regard to the number and class of parameters, it should be noted that the dimensionality of the item response models always operates on the axis of the cases and not on the axis of the items. However, item response theory is interested in modeling some functional qualities of the items that, together with the latent dimensions of the cases, account for the observable responses. To do this, the models introduce one, two, or more parameters that describe the properties of the items into the functions that relate the observed and latent variables. The most important parameters are the "b" parameter, which expresses a similar concept to the classic concept of difficulty, the "a" parameter, which expresses an item property related to the classic concept of discrimination, and the "c" parameter, which refers to the concept of probability of success by random guessing. The restrictions of the functions are specific to various models, especially models for polytomous items and, due to their specificity, they will not be presented here.

Although this framework describes the main theoretical lines, some models and approximations are difficult to classify. For example, the so-called factor analysis of items has been described as referring to factorial analysis models where the observable variables are discrete variables rather than continuous variables, which are generally of a binary nature. Additionally, some types of factor analysis, known as nonlinear factor analysis, have been developed in close connection with the multidimensional models of item response theory.

Item response theory general assumptions

As well as their own particular ones, all item response theory models share some general assumptions.

These general assumptions of item response theory models include:

- The existence of a common latent trait, called θ, underlying the observed variables.

- The inclusion of the set of items or observed variables that make up a test in the estimation of the same latent trait is based on the mathematical dependence of these variables.

- The latent trait explains the mathematical dependence of the observed variables in such a way that if the latent trait is partialized –that is, if the effect of the latent trait is statistically removed from the observed variables, then the observed variables are mathematically independent. This is called the principle of local independence.

- If only one latent trait is sufficient to explain all the mathematical dependence among the items, the test is called unidimensional. If two or more latent traits need to be considered to exhaust the items' dependence, the model is multidimensional.

A comparison of the factor analysis model and the item response theory models

These assumptions lead to certain similarities and differences between the factor analysis model and classical test theory.

Like the factor analysis model, item response theory hinges on the idea of one or more latent traits. Unlike factor analysis, item response models do not provide a way to estimate the number of latent traits. However, any application of item response theory begins by assuming a certain dimensionality (generally, assuming a unidimensional model, that is, only one latent trait) although the estimation procedures may provide a way to evaluate the model fit.

Like the factor analysis model, item response theory models assume that the presence of a latent trait (or more than one for the multidimensional models) may be deduced from the interrelationship among the set of observed variables –(generally, the set of items from a test). Because the items are statistically related, there should be a latent variable that can explain all these interrelationships.

The factor analysis model generally assumes that the relationships among the items are of a linear nature and so can be summarized using covariances, Pearson correlations, or estimations of Pearson for dichotomized variables such as the tetrachoric coefficient. Unlike factor analysis, item response theory models are not based on linear relationships; for this reason, item response models hinge on the more general concept of mathematical dependence.

Moreover, the relationship between the observed variables and the latent factors is also linear for the mainstream factor analytical procedures. In contrast, the relationship between each item and the latent trait in the item response theory models is non-linear; specifically, such a relationship can be described by a logistic function.

This logistic function is particularly well suited to predicting a dichotomous output (for example, a pass/fail behavior) from a continuous variable (the latent trait). This is the main reason why item response models are immediately appropriate for the typical binary scored optimal performance test items, whereas the original factor analysis techniques are not.

Factor analysis uses only one parameter to relate each item to each factor. This one parameter expresses the relationship between the item and the latent trait and makes it possible to identify the relative importance of every item in defining the factor. In factor analysis, items may have strong or weak relationships with a given factor, and this is a

key feature in interpreting the psychological meaning of the factor. Conversely, item response models do not allow different degrees of relationships between each item and the latent trait but they establish a complex relationship based on one, two, or three parameters that characterize the item. Because the different items do not make different contributions to the definition of the latent trait, in item response theory models there is nothing like the psychological factor interpretation that we find in factor analysis.

In item response theory models, the psychological interpretation of the latent trait is a simple issue. As in a simple total score from classical test theory, there is no opportunity to figure out the differences in the contributions from the several items to the formation of the composite. An item response theory model simply assumes that the latent trait is made up of all the test items under analysis but differentiated contributions from different items cannot be identified.

Item dependence and local independence

In both classical test theory and factor analysis models, the linear relationships that characterize the item-item relationships and the item-test or item-latent trait relationships are a really restrictive class of relationships. Two variables may be demonstrated to be related through an infinite number of possible functions and through endless possibilities that cannot be easily described using functions.

The linear relationship is just one of these infinite possibilities, though it is particularly interesting for several reasons. First, it is one of the simplest ways; second, it is an effective way that has been found to be useful for describing many phenomena and relationships throughout the sciences; third, there are very well known, easy to use, and convenient statistics for estimating linear relationships. Not surprisingly, linear relationships are one of the most popular solutions for relating two variables in an atheoretical setting, that is, without a well-developed theory that explains the kind of law or relationship expected.

It is true that, most of the time in psychology and social sciences, linear relationships show only a moderate-to-low capability to explain the data but this is often the best option when lacking a well-developed theory and more precise measurements.

Classical test theory and the mainstream factor analysis models rely on linear relationships, for which we do have very well-known statistics. For this reason, classical test theory and factor analysis models work with covariances, all kinds of correlations, and linear regressions. For example, if we look at classical test theory, the reliability index is the correlation between the true scores and the observed scores. Its square, the coefficient of reliability, is also defined as the Pearson correlation between two parallel measurements. The alpha coefficient is based on the items' correlations or covariances. The item coefficient of homogeneity is the correlation between the item and the total

test score, and the test coefficient of validity is defined as the Pearson correlation between the total test scores and externally measured criteria. If we look at factor analysis, most of the procedures work on the variances-covariances matrix, the unrotated factor matrix is based on linear relationships between the observed variables and the latent traits, and the factor loadings after an orthogonal rotation can be interpreted as correlations between the observed variables and the underlying factors. In the realm of validity, linear regressions (simple or multiple) are the main tools for identifying relationships between several tests and between tests and other factors. Even many procedures classified as construct validity, such as the multitrait-multimethod matrix, are fully based on linear relationships.

Classical test theory and the factor analysis model are fully consistent in this regard: they assume and expect linear relationships and subsequently analyze the relationships among the items and between the items and the latent variables using statistics for linear relationships.

The case for item response theory is somewhat more complex. Everything starts when the analysis of the relationship between the continuous variable that represents the position on the construct (that is, the total test score, the total true score, or the latent trait measured by a factor coming from factor analysis) and the binary variable that represents the success in solving an optimal performance test discovers that linear regression is not a suitable statistical function to represent this relationship. This is clear: the probability of a correct answer ranges from 0 to 1, but any attempt to predict this probability from a continuous variable using linear regression will produce values outside the [0 1] interval. This is a bad result, especially after taking into account that any formal definition of probability always restricts it to a number within the interval [0 1].

After this discovery, authors such as Thurstone, Lazarsfeld, and Lord identified the cumulative normal ogive –the popular function that provides the even more popular tables of left-hand cumulative probabilities for a Gauss distribution– as a function that solves the problem. Based on a continuous metric, the z-score probabilities never cross the limits. Moreover, the s-shaped form of this cumulative function resembles and somehow fits the empirical item response curves –that is, the values of the index of difficulty p plotted along the several values of the total test score or the latent trait score. The discovery of this use of the cumulative ogive is not surprising given the familiarity of any well-educated psychologist with the Gaussian distribution and its various uses in statistics and psychological scaling.

The cumulative normal distribution is theoretically plausible as a model of the trait-binary item response relationship, but it comes from a definite integral of a rather complex function. This means that some problems in handling the model and estimating parameters may arise. Therefore, is there not a simpler function that can represent this s-shape relationship? The solution comes when several authors introduce the logistic

function as a natural solution for this problem. The logistic function is well known in several areas, is the basis of logistic regression, is easier to handle than cumulative logistics, does not involve any integral, makes it possible to integrate several parameters describing the item properties, and produces the same type of s-shaped curves as the cumulative logistic. In fact, the kingdom of logistic functions to describe the item-test or the item-latent trait relationships seems to have all the wind behind it. It began the reign of item response theory and, through successive and impressive advances in diversity, methods of estimation and feasibility for applications, has ruled the psychometrics arena since the final few decades of the last century. It is clear that the logistic function is much better qualified to describe and estimate the item-latent trait relationship for any optimal test measurement binary item. However, this advance involves the need to estimate the latent trait by assuming these logistic relationships and not the linear ones used by factor analysis. Also, once we have a latent trait, we need to base it on the relationships among the items. What other justification could there be for extracting a factor from a set of items? If the items were not related to each other, why would a common factor be assumed for all of them?

The classic ideas of factor analysis come to mind: we extract a factor because there is common variance to explain, because the items are related. If the items were not related, they would not share a common factor. Therefore, what we extract is based on these relationships; hence, after the factor has been extracted –partialized– from a pair of items, its relationship should vanish. If item response theory is to be an alternative test theory, it has to be able to extract the latent traits from the relationships between items. And after partializing the factor influence, the items should remain unrelated.

Thus far, all this seems okay. However, what would the class of relationships be among items that support logistic relationships with the latent trait? Well, because items are conceived as binary entities, how do we establish and why do we defend a logistic relationship between two binary entities? There is no need to do so. Assuming a logistic relationship between the latent trait and each of the binary items does not mean that we should expect the same kind of relationship between the observed variables. The relationship between one binary entity and another binary entity defines a four-point space that in a certain sense may well be defined by a linear relationship but these data do not seem to be easily or naturally defined by a logistic relationship. Item test theory becomes rather unspecific on this point. Clearly, items should be related –or there is no way of supporting the idea of a common factor– but the kind of relationship among the items is unspecific and open to any mathematical form. This is exactly the idea of mathematical or statistical dependence[2].

---

[2] Wermuth and Cox (2005) provide a more thorough description of several kinds of dependence: "If two variables are statistically independent, then the distribution of one of them is the same no matter at which fixed levels the other variable is considered and

Following Crocker and Algina (2008), for two dichotomously scored items, these concepts can be defined as follows. Let

$$P_k(k=1) = P_k(1)$$

denote the probability of answering the *k*th item correctly, and

$$P_k(k=0) = P_k(0)$$

represent the probability of answering the *k*th item incorrectly. Let

$$P_j(1) \text{ and } P_j(0)$$

denote the corresponding probabilities for the *j*th item.

These four probabilities are obtained independently for any item –that is, without taking into account the responses to the other item– and, because they are frequently obtained from the marginal distributions of a cross-tabulation table, they are also called marginal probabilities.

Further, let

$$P(1,1),\ P(0,0),\ P(1,0) \text{ and } P(0,0)$$

indicate the probabilities of the response patterns defined in parentheses.

For example, P(1,1) denotes the probability of answering the *k*th and *j*th items correctly.

---

observations for such variables will lead correspondingly to nearly equal frequency distributions. If there is deterministic dependence, then the levels of one of the variables vary in an exactly determined way with changing levels of the other. In other words, under independence, knowledge about one feature remains unaffected by information provided about the other, while under deterministic dependence it follows with certainty which level of one variable occurs as soon as the level of the other variable is known" (pages 4260-4261).

Taking these definitions into account, the scores on two items are statistically independent if:

$$P(1,1) = P_k(1)*P_j(1)$$

$$P(1,0) = P_k(1)*P_j(0)$$

$$P(0,1) = P_k(0)*P_j(1)$$

$$P(0,0) = P_k(0)*P_j(0)$$

and are statistically dependent if *any* of the four equalities is not met.

For example, if $P_k(1)=.7$, $P_k(0)=.3$, $P_j(1)=.8$ and $P_j(0)=.2$, then the scores on the two items are independent if and only if:

$$P(1,1)=.7*.8=.56,$$

$$P(1,0)=.7*.2=.32,$$

$$P(0,1)=.3*.8=.24,$$

$$P(0,0)=.3*.2=.06.$$

These latter probabilities are called the expected probabilities for the four cells of a cross tabulation.

How can we find out whether a pair of items empirically satisfies this definition? First, we should obtain the cross tabulation of both items, thus enabling the marginal distributions to be calculated.

The marginal frequencies divided by the total number of cases provide the probabilities of answering each item correctly or incorrectly without taking into account the results of the other item. These are called the marginal probabilities. There is one marginal probability for every value of each variable involved. In the case of the cross tabulation of two binary items, there are two possible results {0 1} for each item, so we have four marginal probabilities.

The product of these marginal probabilities then provides the expected proportions or expected probabilities for the four cells.

The frequency for each cell divided by the total number of cases then provides the observed proportion or observed probability for each cell.

Finally, for each cell, we compare the observed probability to the expected probability. If these two numbers are equal for the four cells, we can say that the two items are

mathematically independent –that is, the distribution for each level of each item is not affected by the other item. If, for any of the four cells, the expected probability is not equal to the observed proportion, there is some kind of dependence.

Each of the four cells of a cross tabulation for a pair of binary items defines a response pattern: (11) (10) (01) and (00). Item scores are independent if the probability of each response pattern to both items, that is, the probability of the observed answers, can be calculated by knowing only the probabilities of the correct and incorrect responses to each item –that is, only the marginal probabilities.

Any situation in which one or more of the conditions for mathematical independence is not met produces statistical dependence. Mathematical dependence involves some kind of relationship between the variables because one or more of the conditional distributions of at least one variable shows one or more changes associated with the level of the other variable.

Correlation and linear regression involve a specific kind of relationship –specifically, a linear relationship– but two variables may be related in many other ways. Thus, two variables can be dependent but not correlated. If two variables show dependency but not a linear correlation, this means that there is some kind of non-linear relationship between them. However, if two variables are correlated, this necessarily involves mathematical dependence. Like correlation, statistical dependence does not necessarily involve a causal relationship. Two variables may present dependency because one affects the other or because another –a third variable not included in the analysis– affects both. There are many ways a set of variables may affect or be affected by others. These include some paradoxical effects.

Although classical test theory and the factor analysis model are based on linear relationships, which involve the use of covariances, correlations, and lineal regression, item response theory models involve non-linear logistic relationships between the observed items and the latent traits. For this reason, item response theory requires the analysis of statistical dependence rather than the use of correlations.

However, if item dependence is the basis for estimating a common factor, item independence –called local independence– after the effect of the factor is partialized is the basis of the definition of unidimensionality.

If a particular item response model fits the data, then the items should show dependence before the consideration of the latent trait. However, once the level of the latent trait is taken into account, the items should be locally independent, which means independent at each level of the latent trait.

This idea is similar to the foundations of factor analysis. However, to understand why this independence after the effect of the latent dimension is removed is called 'local

independence', it is worth understanding how the independence is tested after the latent trait is extracted.

Let us suppose that the latent trait can be divided into 10 homogeneous levels. These 10 latent trait levels allow us to classify the sample into 10 groups. All the cases within one level show the same aptitude, so they are assumed not to present any differences due to the latent trait. If we cross tabulate any pair of items, including the full set of cases (that is, the entire sample), with people from any of the 10 latent trait levels, then any pair of items should show dependency because this dependence is due to the latent trait. However, if we select any of the 10 latent trait levels and we cross tabulate the same two items for the subsample of this specific trait level again, then the two items should yield local independence –that is, independence at the local level of the trait we selected.

If the model fits the data and all the item dependence can be explained by only one latent trait, then all relationships or dependence between the items will disappear when we drop the latent trait variability by choosing homogeneous samples on the latent trait. If, for example, all the people come from level one and there is therefore no variability in the latent trait, then the two cross-tabulated items cannot show any dependence because, if the model is unidimensional and fits the data, all the dependency between items is explained by the latent trait differences.

Practical and theoretical drawbacks when checking the local independence assumption

To test the assumption of local independence –which involves checking the dimensionality of the model (i.e., that only one latent trait explains all the dependency for the unidimensional case)– all possible pairs of items should yield independent relationships when cross tabulated within any level of the latent trait.

This formal way of checking local independence is quite demanding because it involves a large number of checks by cross tabulating all the pairs of items.

For example, if we have an optimal performance test made up of 20 binary items that may be either correct or false and the continuous scale of the latent trait is simplified to just 10 levels of aptitude, we have to perform n(n-1)/2 cross tabulations at each of the 10 levels of the latent trait and in each cross tabulation we have to check 4 equalities (one per cell). That is, we have to perform

$$[20(20-1)/2]\bullet 10\bullet 4 = 7{,}600.$$

checks.

This is really an enormous number of checks for a common test of binary scored items, especially if we bear in mind that classifying the sample into only 10 levels may be the result of a balanced decision.

For a large sample, 10 levels may seem to be a suitable partition if we consider the number of cases in each level and especially the number of cases in the levels at the extremes of the scale. Because many traits (whether for psychologically based reasons or methodological artifacts –as discussed earlier– or for both) tend to have a more or less Gaussian distribution, the intervals close to the center of the distribution often have the highest frequencies, while the intervals at the extremes of the distribution have low frequencies.

The problem is that we need a reasonable sample size in each interval, not only in those around the mean of the latent trait.

This requirement of a suitable sample size becomes more challenging if we consider that the local independence must be tested for any pair of items. This means that, for example, a pair of difficult items also has to show their local independence at the low levels of the latent trait. However, few cases if any would have correctly solved a difficult item at a low level of the latent trait. This means that, given the meaning of the latent trait, we expect some or many zero or close to zero observed frequencies at the extremes of the scale. In turn, this is a serious drawback for any test of dependency.

This kind of problem suggests that wider intervals are required in the latent trait in order to classify the sample in aptitude groups with a large enough sample size. However, the latent trait is, by definition, a continuous variable that is theoretically defined in the interval $(-\infty, +\infty)$, though in practical terms this interval might be [-3, +3] on a standardized metric around a 0 mean. The problem is that the wider the intervals, the greater the latent trait differences within each interval. If there are differences in the latent trait score within the interval, checking local independence become useless because this check is based on the idea that all cases within an interval share the same level of the latent trait. Only if all the cases that enter in a cross tabulation between two items have the same level of latent trait can we claim that the latent trait effect has been removed and thus expect local independence.

Theoretically speaking, the fact that, as a distinctive feature of the item response theory, the latent trait has been defined as a continuous variable suggests that the latent trait values estimated for a sample of cases would rarely, if ever, produce exactly the same latent trait score for two examinees.

If there are no examinees with exactly the same latent trait score, the idea of checking local independence (and therefore the unidimensionality or, in general, the dimensionality of the model) by identifying subsamples of cases with the same latent trait score, though theoretically appealing, is also theoretically contradictory. Theoretically, it can in no case be expected that a sample, even a huge one, will produce

enough examinees with the same latent trait score at each level of latent trait score to allow a cross tabulation check for any pair of items.

Empirically, this brilliant idea of checking local independence –and therefore unidimensionality or, in general, the model dimensionality– requires the formation of groups of homogeneous latent trait scores, i.e., subsamples in which the differences in the latent trait score within the group would be negligible.

The idea of negligible latent trait differences is appealing but contradictory. There is no theoretical support for such a concept, which involves differences in the latent trait –ultimately, the important result of any measurement– that are not real differences or at least can be treated as non-real differences. It would be fun to try to define an epsilon value ε, such that

$$\theta_j-\theta_k<\varepsilon$$

is a no measure value, i.e., it identifies the threshold at which a measure difference loses its measurement value to become meaningless.

Of course, in some way, this idea is implicit in the use of any measure –there is a point of precision below which there is no particular practical advantage. However, somehow it is intrinsically contradictory with the same idea of the latent trait as a continuous meaningful measure. Unless an external criterion of usefulness is declared, searching for this embarrassing ε would be entirely contradictory because it involves:

$$0<\left|\theta_j-\theta_k\right|<\varepsilon$$

but also:

$$\left|\theta_j-\theta_k\right|=0$$

If ε is 0, then such a point exists, which means that no difference can be defined between two latent trait values that, although greater than 0, is at the same time meaningless. However, if ε appeals to a difference greater than 0, it has no meaning simply because it tries to define a difference that works like a 0 difference.

The solution comes from the external criterion. In real life, such epsilon values are implicit for many measures. If we intend to measure the length of a wall, we might be interested in meters and centimeters, and perhaps millimeters for some particular tasks, but we are probably not interested in microns. Of course, though millimeters and even

kilometers are made up of microns, for practical reasons, microns are useless in most everyday situations.

Given this example, someone may suggest a practical rule of thumb for negligible latent trait differences when forming groups to check local independence: two latent trait scores are reasonably equal (and so can be mixed in the same interval) if, when items within this interval are cross tabulated, they show local independence. This would be a nice suggestion with practical effects. In fact, this is more or less the idea when the number of levels of the latent trait is determined for checking local independence. It may run and (though laborious) may be enough to solve the practical problem. From a theoretical point of view, however, this is a good example of a tautological definition: the interval size is defined in such a way that the condition of local independence –which is precisely what we are trying to verify– is satisfied.

Other drawbacks stem from the fact that any sample is exposed to sampling error, and any act of measurement is exposed to measurement error. These two factors imply that even if all the items in a population have local independence, any sampling of them, subject to measurement error, may show some deviance from the expected patterns of local independence. The mathematical definition of independence does not allow the presence of deviance due to sampling error due to or measurement error.

Given a dataset for a certain test obtained in a certain sample and subject to some error of measurement, if the cross tabulation of two items for a given level of the latent trait does not satisfy the rule of independence, the items are dependent for whatever reason. The rule of dependence comes from the realm of mathematical probabilities and does not take into account sources of error for statistical reasons such as sampling or for psychological measurement reasons such as measurement error. It may sound somewhat paradoxical but if the independence rule were a scaling model, we would call it a deterministic model instead of a probabilistic one.

That is, there is no model such that:

$$P(1,1) = P_k(1)*P_j(1)+ \Lambda$$

$$P(1,0) = P_k(1)*P_j(0)+ \Lambda$$

$$P(0,1) = P_k(0)*P_j(1)+ \Lambda$$

$$P(0,0) = P_k(0)*P_j(0)+ \Lambda$$

where capital lambda Λ means a mixture of sampling error S and error of measurement E:

$$\Lambda = S + E$$

where both random variations have either positive or negative signs.

For any cross tabulation, this would imply developing a metric defined within the probability scale, which in turn would involve developing a probabilistic definition of the amount of sampling error and a probabilistic definition of the amount of measurement error for any bivariate binary distribution.

As this Λ parameter has not been proposed, the definition of local independence relies on a deterministic approach, which means that, from a mathematical point of view, any deviance from the basic rule of products of marginal proportions involves some dependence.

However, such a harsh rule would not be helpful for any real sample. It is hard to believe that even one of the

$$(n(n-1)/2)\bullet(\theta \text{ levels})$$

cross tabulation tests of dependence, each of which is based on

$$(v_j \bullet v_k)$$

equalities (where v_j is the number of possible scores for the item j, and v_k is the number of possible scores for the item k, –2•2 for binary scored items) would yield a result of independence. Somehow, the Λ terms, even though they are not formally defined, have to be taken into account for real cases.

A standard solution for these drawbacks is to apply standard statistical tests of independence, e.g., a chi-square test. Of course, chi-square is a test with some assumptions that are difficult to fulfill in some situations for the local independence case we are discussing. Fortunately, however, many other statistical tests have been developed to cover some or many of these situations in which the assumptions of chi-square are hard to fulfill. However, all these statistical tests follow the logic of a statistical test of hypothesis based on the probability of the given result when the null hypothesis happens to be true. I will discuss this topic now and, for this discussion, will use the well-known chi-square test as a model.

Let us suppose that all the assumptions for a chi-square test are fully satisfied for a given test of local dependence. We test the local independence for a couple of binary scored items at a certain latent trait level. Of course, the result does not tell us that the four equalities are exactly true –if it did, we would have perfect mathematical independence, and no hypothesis testing would be needed. We then apply the chi-square test, whose

assumptions have been fully justified. Imagine that the result is that the chi-square test has a p=0.2. If we apply the standard interpretation for a statistical test of independence, because this p is greater than the conventional level of 0.05, we cannot reject the null hypothesis and therefore do not have enough evidence to claim that there is dependence in the population. That is, we cannot say that the dependence that in fact both items show in the sample is big enough to claim that there is a dependency relationship in the population. But is this proof of statistical independence? Note that if we increase our alpha level to 0.01 to reject a null hypothesis, it would be even harder to identify dependent items and more pairs of items could then be declared statistically independent. If we increase our alpha level to the rather compulsive (in psychology and the social sciences but not in physics or other natural sciences) threshold of 0.001, then nearly all our items will be declared statistically independent and the item response assumption of local independence (and therefore the unidimensionality property) will be warranted. Of course, something is wrong with this way of reasoning.

Statistical tests of independence such as chi-square were designed to seek statistical dependence. That is, these tests try to ensure that when we say that two variables are related, this is true in the population and not an accidental result that stems from sampling error. They are not intended to test the equality of the expected frequencies and the observed frequencies for each cell as they are not designed to prove mathematical dependence. Of course, not rejecting the hypothesis of independence when applying such statistical tests of independence is a basic prerequisite for accepting mathematical independence. But this is not enough. This does not prove mathematical independence.

If we were to use a test such as chi-square to test the statistical independence except for sampling error, the region of acceptance of the null hypothesis –what we are trying to prove in this case– should be reduced to 5% of the distribution around the null hypothesis value, so a bidirectional alpha level would have to be 0.95. This would accept only small variations around the equalities that define the mathematical independence that may be attributed to sampling error. The chi-square test and many other tests of independence were not designed for this task and it is hard to think of them as tests for proving statistical independence –except for sampling error– rather than tests for showing dependence when it is improbable that this would only occur due to sampling error.

Logistic models to represent the probability of a correct answer

Item test theory is a general family of mathematical models that represent the relationship between the item answers and a latent trait, called θ, using a set of mathematical functions (mainly logistic functions).

In general, a logistic function adopts the form

$$P_j(x_{ij} = x \mid \theta_i) = \frac{e^{\theta_i}}{1 + e^{\theta_i}} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)}$$

and can also be expressed as

$$P_j(x_{ij} = x \mid \theta_i) = \frac{1}{1 + e^{-\theta_i}} = \frac{1}{1 + \exp(-\theta_i)}$$

or

$$P_j(x_{ij} = x \mid \theta_i) = \frac{1}{1 + e^{-\theta_i}} = [1 + e^{-\theta_i}]^{-1} = [1 + \exp(-\theta_i)]^{-1}$$

All these equivalent ways of expressing the logistic function say that the probability P of obtaining a certain answer x to the item j by the respondent i is a function of the location of this respondent θ_i on the trait θ. The e in the equations is the number e, i.e., the mathematical constant e=2.71828, which is the base of the natural logarithm.

The first part of any of these equalities, i.e., the expression

$$P_j(x_{ij} = x \mid \theta_i)$$

can be read as the probability P that the answer to item j by the case i would be equal to a certain value x, given his/her level on the trait θ. Generally, for most models the answer the model tries to explain is the response of interest, which is usually the correct answer to an item on an optimal performance test. The expression "given the θ level" is really important and shows one of the changes in perspective introduced by item response theory compared to classical test theory.

The item response theory approach versus classical test theory

In classical test theory, the relationship between the item score and the position in the trait, represented by the total test score (or the true total test score under the observed total test score) is analyzed by the coefficients of homogeneity. Since the coefficient of homogeneity is a correlation, the implicit assumption is a linear model. However, when the purpose is to explain a binary item response model (such as a pass/fail item), the linear regression is not a good solution because its output –that is, the predicted answer for the item– is not going to be in the interval [0 1] for all abscissae. This is a main advantage of the logistic function for this purpose: the logistic function is able to show the relationship between an independent continuous variable (the latent trait) and a

binary output (the presence or absence of a correct answer) producing an output in the interval [0 1].

In classical test theory, the probability of a correct answer was originally analyzed for each item as the proportion of respondents who pass the item. This approach gives a global value p_j for the item j. For example, p=.5 means that half of the cases have passed the item j. If the sample is good enough, this p may also be interpreted as a probability: if we randomly choose a new case from the same population, there would be a .5 probability that this new case is able to answer the item correctly. However, this probability clearly depends heavily on the position of the respondent on the trait. If, for example, this trait is an aptitude and the respondent belongs to the best performers, the general p probability, as a prediction of the respondent's performance, may be clearly improved. That is, for a high-level performer, it is natural to expect a high probability of giving the correct answer, whereas for a respondent on the lowest levels of the aptitude scale, a low probability of answering the item correctly may be expected. The probability p depends on the examinee's ability.

The traditional focus of classical test theory has been to provide this p value for each item and summarize the proportion of cases that answer the item correctly, instead of providing a different p for the different levels of aptitude of the examinees.

However, classical test theory has also developed some analyses that provide p values based on the stratification of the respondents by their level on the total test score. The index of discrimination based on p is a good example.

This index of discrimination based on p, sometimes called the D_j index, is a discrimination index; i.e., it reports the ability of item j to discriminate the total test score value. To calculate D_j, the sample is ordered by the total test score. The total sample is then divided into groups determined by the total test scores. The classical way of applying the D_j index takes into account two extreme groups. Thus, 27% of the cases with the highest total test scores form the group of highest performers, whereas 27% of the cases with the lowest total test scores form the group of lowest performers. The proportion of cases that pass item j is calculated separately for these two groups. That is, p_h represents the proportion of cases that pass the item for the high performers, whereas p_l represents the proportion of cases that pass the item for the low performers. D_j is just the difference:

$$D_j = p_h - p_l$$

If this difference D_j is large enough, the item is considered discriminant because the item answers reflect the position of the examinees on the total test score.

This idea of a separate and different probability of answering an item correctly for examinees belonging to the extreme opposite levels on the measured variable can be generalized to a more detailed definition of levels. This is the objective of the empirical item response curve.

The empirical item response curve or p profile

Although the empirical item response curve is not an item response theory concept, it is key to understanding item response theory concepts. It may be seen as a generalization of the separate p values for the extreme groups used in the D_j index.

With a large enough sample, the cases may be classified by their position on the total test score. For example, if we have a test whose total score ranges from 1 to 10, the cases can be classified into 10 groups. These ten groups represent the ten levels of aptitude that this test is able to distinguish.

Now, for a certain item j, a separate p value can be calculated for each aptitude level. If p_v represents the p value for the group of examinees that score v on the total test score, then p_1, p_2, p_3 ... p_v ... p_9, p_{10} represent the p values for the ten groups of aptitude. The ordered vector $\{p_1, p_2, p_3 \ldots p_v \ldots p_9, p_{10}\}$ is the profile of difficulty for item j.

If the item works as expected, the proportion of cases giving a correct answer to the item should increase as the level of aptitude increases. That is

$$\{p_1 < p_2 < p_{3 <} \ldots p_v \ldots < p_9 < p_{10}\}$$

If the sample is large enough and representative of a certain population, these proportions may well be interpreted as probabilities. Examinees with a high level of aptitude are expected to show a higher probability of answering the item correctly than examinees with a low level of aptitude.

In some ways, this p profile for each item, also known as the empirical item response curve, is what item response theory models try to explain. The output of the item response theory functions is just the probability of a certain response –generally, the correct one– given the level of aptitude.

There are three main differences between, on the one hand, the empirical item response curves or p profiles throughout the total test score levels and, on the other hand, the item response theory functions –or their graphical representations, which are called item response curves or item characteristic curves:

First, the empirical p profiles classify the observed total test scores into discrete levels, whereas the item response functions work on the continuous latent trait θ, which is also estimated from the item answers.

Second, the empirical p profiles do not provide any function to estimate the probability of success from the aptitude levels, whereas the item response theory functions do provide such a function, usually a logistic function.

Third, whereas the empirical p profiles do not parameterize the item characteristics that influence the probability of a correct answer, the item response theory functions include the item parameters that allow us to understand which item properties –such as item difficulty, item discrimination, or response guessing– influence the form and location of the item characteristic curve on the scale of aptitude provided by the latent trait θ. These item parameters describe the items, provide the model with flexibility, and improve the estimation of θ, thus allowing a better fit between the model predictions and the real answers.

The classic item response logistic models for binary items and their item parameters

The classic item response theory functions express the probability of a correct answer to an optimal measurement test as a logistic function of the position of the respondent on the latent trait θ.

All the logistic models give the probability of an observed answer as a function of the θ value, i.e., as a function of the location of the respondent on the latent trait. The logistic function may be more or less complex depending on the number of parameters it includes.

There are three classic item-response models: the one-parameter, two-parameter, and three-parameter logistic models.

The one-parameter logistic model

The one-parameter logistic model is the simplest. In this model, the probability of a correct answer to the item depends on the respondent's level on the latent trait and the item difficulty. So, in this model, all items are assumed to be equal on any item property except difficulty. Item difficulty is usually represented by the so-called b parameter.

The one-parameter logistic function adopts the form:

$$P_j(x_{ij}=1\,|\,\theta_i)=\frac{e^{Da(\theta_i-b_j)}}{1+e^{Da(\theta_i-b_j)}}$$

or, alternatively, the equivalent form:

$$P_j(x_{ij}=1\,|\,\theta_i)=\frac{1}{1+e^{-Da(\theta_i-b_j)}}$$

The left-hand side can be read as "the probability of a correct answer to item j, given that the examinee i has level θ_i on the latent trait".

In the formula, e is the number e, i.e., the mathematical constant that is the base of the natural logarithm, which is approximately equal to 2.71828.

The parameter "a" expresses the item discrimination. For the one-parameter logistic model, this parameter is constant for all items belonging to the same test; i.e., all the items under analysis have the same parameter "a". For this reason, "a" appears in the formula without the subscript j. When the one-parameter logistic model is applied to a set of items, some item response theory software estimates a unique parameter "a" value for all the items, whereas others may yield results for the parameters θ_i and b_j that somehow integrate "a" in such a way that the results do not produce a separate parameter value for the test.

In the function, θ_i is the latent trait value of examinee i. For a certain item, once the item parameters have been estimated, all the item parameters in the formula are fixed values, whereas θ_i varies. Hence, θ_i plays the role of the independent variable in this function: the probability of a correct answer –the dependent or response variable– will change as the value of the individuals on the latent trait θ_i changes.

The latent trait θ_i is usually measured on a scale from $-\infty$ to $+\infty$, with a mean of 0. This metric resembles the standardized z-scores on the horizontal axes of a standardized normal distribution. Although the θ variable goes from $-\infty$ to $+\infty$, when the metric is adjusted to the classical normal distribution, most cases are assumed to be between –3 to +3.

D is a scaling constant usually set at 1.7. If D=1.7, the values of P_j for the logistic models and the values of P_j for the cumulative ogive model differ absolutely by less than .01 for all values of θ. D remains the same for all items and does not affect the nature of the model. As a scaling constant, D is not a substantive part of the logistic model and can be

eliminated from the function. Given that D multiplies the term $a(\theta_i-b_j)$, eliminating the D constant from the model is the same as fixing its value to 1, i.e., the same as assuming D=1.

In the one-parameter logistic model, parameter b_j is the only item parameter that is allowed to vary from item to item. Parameter b_j expresses the item difficulty. It is the only real item parameter for the one-parameter logistic model and in some ways the most important item parameter for any classic logistic model.

For the one-parameter logistic model, the item difficulty b_j is what differentiates one item from another. It allows us to identify the different levels of item difficulty from very easy to very difficult just by looking at the b value associated with each item. When the θ for a sample are transformed to a normal metric, so that their mean is 0 and the standard deviation is 1, the values of b_j typically vary from -2.0 to +2.0. Values of b_j close to -2.0 correspond to very easy items, and values of b_j close to 2.0 correspond to very difficult items.

An interesting and valuable property of the b parameter is that it is expressed on the same scale as the latent trait θ. This allows us to compare items and examinees on the same scale. If an item j shows a difficulty b parameter greater than the θ value for an examinee i, we can write

$$b_j>\theta_i$$

which means that it is probable that examinee i will fail item j. This statement allows us to examine the stochastic –that is, probabilistic– nature of the item response theory models. If $b_j>\theta_i$, this does not mean that examinee i has a probability equal to 1 of failing item j but that this probability is greater than the probability of passing the item. In fact, for any item, parameter b means the point where the probability of answering the item correctly is 0.5, Therefore, if an examinee i has a trait level θ_i equal to the parameter b value of an item j, the interpretation is that this examinee has a 0.5 probability of answering this item correctly. Therefore, if an examinee has $\theta_i > b_j$, this means that his/her probability of passing the item is greater than 0.5 but not equal to 1.

The item characteristic curves of all the items on a test analyzed under the one-parameter logistic model show the same form but different locations on the horizontal axis, because b changes the location of the curve over the θ axis. Thus, when the set of item characteristic curves corresponding to a set of items of the same test analyzed using the one-parameter logistic model is represented on the same graph over the same latent trait axis, the result is a set of parallel logistic s-shaped curves, all of which have the same slope (determined by the parameter a) but different locations on the latent trait θ (determined by the parameter b, which is different from item to item).

The two-parameter logistic model

The second parameter introduced in the two-parameter logistic model is the parameter a_j. The two-parameter logistic model removes the restriction of equal item discrimination for all the items. In the two-parameter logistic model, there is one a_j value for each item. For this reason, we now write parameter "a" with the subscript j because it changes for each item j.

This model adopts the function:

$$P_j(x_{ij} = x \mid \theta_i) = \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}}$$

If the numerator and the denominator are multiplied by

$$e^{-Da(\theta_i - b_j)}$$

this function can also be expressed as:

$$P_j(x_{ij} = x \mid \theta_i) = \frac{1}{1 + e^{-Da_j(\theta_i - b_j)}}$$

The only change from the previous one-parameter logistic model is the subscript j for the "a" parameter. The rest of the terms have exactly the same meaning.

Parameter a_j expresses the capability of the item characteristic curve to discriminate around point b_j. For this reason, a_j is called the discrimination parameter and is closely related to the slope of the curve at point b_j. In a two-parameter logistic model, the values of parameters b_j and a_j will vary across the items on a test.

Hambleton and Swaminathan (1985) pointed out that the item discrimination parameter a_j is defined, theoretically, on the scale $(-\infty, +\infty)$. However, negative values of a_j have no meaning because they imply that the probability of a correct answer will decrease as the latent trait score increases. Negatively discriminating items are therefore discarded from ability tests. Although possible, it is unusual to identify items with a_j values larger than 2. Hence, the usual range for item discrimination parameters is approximately from 0.1 to 2.

When a_j is high, there is a strong and fast change in the probabilities of a correct answer between the zone of θ values below the item b_j value and the zone of θ values above the b_j value. Conversely, when a_j is low, there is little change in the probabilities of a correct answer between the zone of θ values below the item b_j value and the zone of θ values above the b_j value.

High values of a_j result in item characteristic curves that are very 'steep' around the b_j value, whereas low values of a_j lead to item characteristic curves that increase the probability of a correct answer gradually as a function of θ around the b_j value.

It should be noted that parameter a_j represents the slope of the logistic curve around the b_j value but in some way represents the slope or discriminant power on the rest of the curve. Because the logistic curves are s-shaped along the θ axis, those items with high a_j parameters show steep slopes around b_j, i.e., in the central zone of the s-curve, while the tails of the curve are left relatively horizontal. In other words, items with a high a_j value are especially discriminant around b_j and especially non-discriminant on the parts of the curve (usually at the beginning and the end) that are far from b_j. In contrast, items with a low a_j value are not especially discriminant around b_j, but may show a certain slope in the parts of the curve (usually at the beginning and the end) that are far from b_j.

The three-parameter logistic model

The one-parameter logistic model allows the items to differ in difficulty but not in discriminative power. The two-parameter logistic model allows the items to differ in difficulty as well as in discriminative power. The three-parameter logistic model allows the items to differ in difficulty, discriminative power, and the probability of a correct answer due to guessing.

The form of the three-parameter logistic model is:

$$P_j(x_{ij} = x \mid \theta_i) = c_j + (1 - c_j)\frac{1}{1 + e^{-Da_j(\theta_i - b_j)}}$$

where c_j is called the pseudo-guessing parameter and represents the probability of a correct answer for an examinee completely lacking in ability, i.e., an examinee with $\theta_i = -\infty$.

The other terms are the same as those presented for the two-parameter logistic model, the only difference being that the difficulty parameter b_j no longer represents the θ level where the probability of a correct answer is 0.5 but rather the θ level where the probability of a correct answer is $(c_j + 1)/2$.

Strictly speaking, the introduction of the terms $c_j + (1 - c_j)$ breaks the logistic nature of the function, so this is a logistic-based function rather than a proper logistic function.

The c_j parameter represents the probability of a correct answer when the respondents do not know the correct answer and guess.

The importance of this parameter lies in the presence of circumstances that allow and encourage respondents to guess answers. The circumstances that *allow* random guessing are mainly related to the use of multiple-choice formats for educational or aptitude tests. All multiple-choice tests allow the respondents to choose the correct answer for items they do not master. The respondents are *encouraged* to use random guessing on those items that they are unable to solve in cases where they are eager to get a good score, such as exams, personnel selection processes, and other kinds of evaluative assessments.

Selecting an item response model

How should we select an item response model for a test? On this point, we will assume that our test is a multiple-choice optimal performance measure that scores 1 for the correct answer and 0 otherwise.

A general scientific principle known as Ockham's razor or the law of parsimony fits here. This principle says that we should choose the simplest model with the same explanatory power. Translated to this case, this means that if two models fit the data similarly, the best choice is the simplest one, i.e., the model with the fewest parameters.

Some psychometricians have defended the simplicity of the one-parameter logistic model, mainly on account of its exceptional properties from a methodological and theoretical point of view. The one-parameter logistic model can be shown to be a way of expressing the Rasch model –a measurement model with convenient mathematical properties introduced by George Rasch (1960) and based only on item difficulty and the respondent's aptitude. Following this principle, some Rasch psychometricians think a test should follow this theoretical simplicity, so they look for items that fit the Rasch model rather than looking for the model that best fits the items.

It should be accepted that a model rarely, if ever, perfectly fits the characteristics of a test and the data obtained in real circumstances. More often than not, one or more item

response assumptions are not fully satisfied by the data, which means that assumptions have to be relaxed in some way in order to apply the models. The robustness of the model to the violation of the item response theory assumptions is therefore an additional consideration to take into account when selecting a model. A model is said to be robust if it yields reasonably accurate results, even if one or more of its assumptions is violated.

For most psychometricians, the most important considerations when selecting an item response model stem from the model's ability to fit the circumstances of the measurement and the properties of the dataset of interest. This involves making it clear whether the assumptions of the model are realistic for the test analyzed. Because different models with a different set of parameters involve different assumptions, these differences provide some guidance for choosing an item response theory model.

If the test is multiple-choice and the examinees are motivated to answer the items by random guessing –as in many educational or organizational situations– then our dataset probably requires a three-parameter logistic model.

If the type of test items does not allow guessing (for example, when the test is made up of open-answer items) or the examinees are not motivated to answer the test by guessing (as in some clinical situations), then we probably do not need a three-parameter logistic model and can use either of the other two solutions, i.e., a two-parameter model or a one-parameter model.

Sometimes the items are exposed to possible guessing but the actual impact of guessing on the scores may be negligible, which allows the psychometrician to avoid the three-parameter logistic model. If, after applying a three-parameter model all items seem to have similar and really low c_j values (close to 0), we may try a two-parameter logistic model. If both models show a similar fit, then the simpler of the two should be the general case choice.

In fact, the items on a test may or may not show different discriminations. If, after applying a two-parameter logistic model, all items show a similar a_j parameter, i.e., similar discrimination, this result suggests that a one-parameter logistic model may fit the data reasonably well.

If a pair of items depicts a similar a_j parameter, then these two items should show a similar slope when the item characteristic curve is drawn. A set of items with similar discrimination a_j will show a set of parallel item characteristic curves when these curves are drawn on the same graph over the same latent trait θ dimension.

If the parameters a_j of a set of items are similar enough, perhaps they can be summarized by estimating a common parameter a, thus simplifying the model to a one-

parameter logistic model. If the items on a test suggest the possible fit of the one-parameter logistic model, we might estimate it and compare the model fit for both models.

As a general rule, a model tends to fit the data better as it introduces more and more parameters. Therefore, the psychologist has to balance the attractiveness of simplicity with the improvement in fit that a more complex model produces in order to make a better decision. In general, improvements of scarce magnitude for a more complex model lead to choosing the simplest model, whereas substantial improvements in model fit strongly suggest allowing more parameters.

Sample size requirements

Another factor to consider when choosing an item response theory model is sample size. As a general rule for any statistical model, estimating more parameters requires more cases. This means that small samples may suggest simpler models (such as the one-parameter logistic model), whereas large samples make it possible to choose the model while taking into account other considerations that are only related to the test characteristics and model fit.

As a rule of thumb, a simple model such as the one-parameter logistic model requires at least roughly 100 cases. The more complex three-parameter logistic model requires at least roughly 500 cases. Other more complex models, apart from the three basic ones discussed here, would probably require even more cases.

Therefore, even if you are analyzing a multiple-choice test in which there is an obvious tendency to guess the unknown answers, you cannot apply a three-parameter logistic model if your sample size is rather small, e.g., around 100 cases. You can fairly argue that a more complex model such as the three-parameter logistic model suits the tests and sample characteristics. In this case, however, the solution would be to increase the sample size.

After all, item response theory models not only involve more parameters and stronger assumptions than classical test theory but they also impose harder sample size requirements. These sample size requirements are based not only on the simultaneous presence of several item parameters but also on the demands of the computational methods required to estimate the parameters.

References:

American Psychological Association (2018). Publication Manual of the American Psychological Association, Sixth Edition. APA

Andersen, E.B. (1980). Discrete statistical models with social science applications. Amsterdam: North-Holland.

Barlow, D.H., Bullis, J. R., Comer, J. S., and Ametaj, A. A. (2013). Evidence-Based Psychological Treatments: An Update and a Way Forward. *Annual Review of Clinical Psychology, 9*, pp. 1–27

Barlow, J., Johnston, I., Kendrick D., Polnay, L., and Stewart-Brown, S. (2006). Individual and group-based parenting programmes for the treatment of physical child abuse and neglect. Cochrane Database of Systematic Reviews 2006, Issue 3. Art. No.: CD005463. DOI: 10.1002/14651858.CD005463.pub2.

Becoña, E. and Lorenzo, M.C. (2001). Tratamientos psicológicos eficaces para el trastono bipolar. *Psicothema, 13*(3), pp. 511-522.

Beutler, L. E. (2006). The dodo bird is extinct. *Clinical Psychology: Science and Practice*. https://doi.org/10.1093/clipsy.9.1.30

Box, G. E. P. (1979). Robustness in the strategy of scientific model building in Launer, R. L.; Wilkinson, G. N., *Robustness in Statistics,* Academic Press*, pp. 201–236*.

Carroll, L. (1865). *Alice's Adventures in Wonderland* by Lewis Carroll, first published in 1865 – A Book Virtual Digital Edition, 2000.

Clogg, C.C. and Goodman, L.A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79,* 762-771.

Crocker, L. and Algina, J. (2008). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1990). *Essentials of psychological testing. (5th ed).* New York: Harper and Row.

Dayton, C.M. and Macready, G.B. (1988). A latent class covariate model with applications to criterion-referenced testing. In Langeheine, R. and Rost, J. (Eds.), *Latent trait and latent class models* (pp. 129-143). New York: Plenum Press.

Eliason, S.R. (1988). *The categorical data analysis system. Version 3.00 A user´s manual.* University Park, PA: Pennsylvania State University, Department of Sociology.

Eger, S. (2013). Restricted Weighted Integer Compositions and Extended Binomial Coefficients. *Journal of Integer Sequences,* Vol. 16, Article 13.1.3.

Grinnell, R. (2018). Operational Definition. *Psych Central*. Retrieved on November 30, 2018, from https://psychcentral.com/encyclopedia/operational-definition/

Grinstead, C. M. and Snell, J. L. (1997). *Introduction to Probability*. American Mathematical Society.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215-231.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association, 74,* 537-552.

Haberman, S.J. (1978). *Analysis of qualitative data (2 vols.)* New York: Academic Press.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for loglinear models for frequency tables derived by indirect observation. In Clogg, C.C. (Ed.),

Sociological methodology (pp. 193-211). Washington, DC: American Sociological Association.

Hambleton, R.K. and Swaminathan, H. (1985). *Item response theory. Principles and applications.* Boston: Kluwer Academic Publishers.

Hershberger, S. L. (1994). The specification of equivalent models before the collection of data. In von Eye, A. and Clogg, C.C. (Eds.) *Latent variables analysis: Applications for developmental research* (pp. 68-108). Thousand Oaks, CA: Sage.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington: Joint Committee on Testing Practices.

Langeheine, R. and Rost, J. (Eds.) (1988). *Latent Trait and Latent Class Models.* New York: Plenum Press.

Lawson, D. and Marion G. (2008). An introduction to mathematical modeling. https://people.maths.bris.ac.uk/~madjl/course_text.pdf

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In Stouffer, S.A, Guttman, L, Suchman, E.A, Lazarsfeld, P.F, Star, S.A. and Clausen, J.A. (Eds.), *Measurement and prediction* (pp. 362-412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent structure analysis.* New York: Houghton Mifflin

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22,* 1-55.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science.*

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale: Lawrence Erlbaum Associates.

Lord, F.M. and Stocking, M.L. (1988). Item response theory. In Keeves, J.P. (Ed.), *Educational Research, methodology, and measurement. An international Handbook* (pp. 269-272). Oxford: Pergamon Press.

McDonald, R.P. (1989). Future directions for item response theory. In Hambleton, R.K. (Ed.), *Applications of item response theory*. International Journal of Educational Research. Special Issue 13. (pp. 205-220).

Meliá, J. L. (1990). Introducción a la Medición y Análisis de Datos. Valencia: CSV.

Mooijaart, A. and Van Der Heijden, P.G.M. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika, 57(2),* 261-269.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment test. Copenhagen: Danmarks Pedagogiske Institute.

Rose, S.C., Bisson, J., Churchill, R., and Wessely, S. (2002). Psychological debriefing for preventing post traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews* 2002, Issue 2. Art. No.: CD000560. DOI: 10.1002/14651858.CD000560.

Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. *Psychometrika, 63(1),* 111-130.

Sijtsma, K. and Hemker, B.T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63(2),* 183-200.

Sijtsma, K. and Junker, B.W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49(1),* 79-105.

Steinberg, L. and Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1(1),* 81-97.

Traub, R.E. and Lam, Y.R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36,* 19-48.

van de Pol, F.J.R., Langeheine, R. and De Jong, W. (1989). *PANMARK user manual. Panel analysis using Markov chains, version 1.5.* Voorburg: Netherlands Central Bureau of Statistics.

van der Heijden, P.G.M., Dessens, J. and Bockenholt, U. (1996). Estimating the concomitant variable latent class model with the EM algorithm. *Journal of Educational and Behavioral Statistics, 21(3),* 215-229.

Vermunt, J.K. (1997). LEM: A general program for the analysis of categorical data. Tilburg University Press.

Wermuth, N. and Cox, D. R. (2005) Statistical dependence and independence. In Armitage, P. and Colton, T. (Eds.) *Encyclopedia of Biostatistics.* New York: Wiley. pp. 4260-4264.

Xarxa Vives d'Universitats (2017). Interuniversity style guide for writing institutional texts in English. Manual d'estil interuniversitari per a la redacció de textos institucionals en anglés. Tercera Edició. Castelló de la Plana: Xarxa Vives d'Universitats.