

On-line Assessment of Comprehension Processes

Tomás Martínez, Eduardo Vidal-Abarca, Laura Gil, and Ramiro Gilabert
Universitat de València (Spain)

In this paper we describe a new version of a former paper-and-pencil standardized comprehension test called Test of Comprehension Processes (Vidal-Abarca, Gilabert, Martínez, & Sellés, 2007). The new version has been adapted to a computer-based environment based on the moving window technique. It can be used to assess comprehension strategies of students from fifth to tenth grades (11 to 16 years old). Comprehension strategies are registered on-line using reading times and visits to relevant sections of the text during the question-answering process. Data show that the computer-based version draws similar results to those provided by the paper-and-pencil version. In addition, we identify the particular strategies deployed during the question-answering process by high, medium and low comprehenders.

Keywords: comprehension test, assessment of reading processes, on-line reading measurement, moving window technique in comprehension

En el presente artículo presentamos una nueva versión de un test de comprensión estandarizado de lápiz y papel llamado Test de Procesos de Comprensión (Vidal-Abarca, Gilabert, Martínez, & Sellés, 2007), el cual ha sido adaptado a un entorno electrónico mediante la utilización una técnica de ventana móvil. Esta versión electrónica puede ser usada para diagnosticar estrategias de comprensión de escolares entre 5º de Primaria y 4º de Educación Secundaria Obligatoria (11 a 16 años). Las estrategias de comprensión se miden registrando tiempos de lectura y visitas a segmentos relevantes del texto de forma on-line durante el proceso de responder a preguntas del texto. Los resultados muestran en primer lugar que la versión electrónica es similar a la versión papel y lápiz. En segundo lugar, se muestran diferentes estrategias durante el proceso de respuesta a preguntas del texto que son características de estudiantes con estrategias de comprensión de nivel alto, medio y bajo.

Palabras clave: test de comprensión, evaluación de procesos de lectura, medida de variables on-line en lectura, técnicas de ventana móvil en comprensión

Correspondence concerning this article should be addressed to Tomás Martínez. Facultat de Psicologia. Departamento de Psicología Evolutiva y de la Educación, Universitat de València, Av. de Blasco Ibáñez, nº 21. 46101 Valencia (Spain). E-mail: Tomas.Martinez@uv.es

How to cite the authors of this article: Martínez, T., Vidal-Abarca, E., Gil, L., Gilabert, R.

Current theoretical models describe comprehension as a process in which readers construct a mental representation of the text information. This process involves three groups of variables: characteristics of the readers, the task, and the text, which are organized around a set of processing strategies (Entin & Klare, 1985; Goldman, 1997; Kintsch & Kintsch, 2005; McNamara & Kintsch, 1996; Trites & McGroarty, 2005). The aim of these strategies is to obtain a better or more efficient outcome from the reading task, which requires an adjustment of the processes to the task requirements (e.g., Alderson, 2000; Goldman, 1997).

Despite this process-oriented theoretical framework of comprehension, assessment tools are generally based on product variables, without considering the processes or strategies followed by readers to accomplish a task. These assessment tools have been useful to classify students in regard to their abilities, in a roughly reliable, valid and economic way (Magliano, Millis, Ozurur, & McNamara, 2007). However, in the last few years significant research efforts have been made to develop electronic procedures to assess comprehension processes using on-line variables (e.g., Magliano, Millis, The NIU R-SAT Development Team, Levinstein, & Boonthum, in press; McNamara, O'Reilly, Best, & Ozuru, 2006). As a result, tremendous advances have been made on the analysis of strategies and processes that play a role during text reading.

The procedures used to capture these processes on-line have been *think aloud* protocols and the study of reading times. Reading times are registered through two techniques: the analysis of eye-movements and the moving-window method. These procedures are becoming more accessible because of the development of computer-based systems that can easily handle the processing demands that they require at a relatively low cost.

The moving-window method requires students to read a text through a computer window that presents the text sequentially. Before being presented, the text appears as a set of dashes and blank-spaces that correspond to the words and spaces. When a reader presses a key, the current word is typed onto the screen, replacing the dashes corresponding to that word, and the previous word is replaced with dashes (e.g. Just, Carpenter, & Woolley, 1982). The reading time of a word is defined as the interval between successive key presses. This presentation format maintains the peripheral cues that are used during natural reading, based on the position and length of words and paragraphs. Furthermore, research conducted with this procedure has found similar results to those collected with other procedures of reading analysis (Aaronson & Ferrer, 1983; Just et al., 1982), except for reading times, which are somewhat longer when including a mechanical procedure to advance the reading. A theoretical assumption of these techniques is that students' reading speed depends on the cognitive processes involved in comprehension (Just & Carpenter, 1980). Thus, for each part of the text we can estimate the cognitive load associated with reading.

Nevertheless, during the last few years a debate has arisen about how reading times should be interpreted. The traditional view is based upon the *discrepancy reduction* model (Carver & Scheier, 1990; Koriat & Goldsmith, 1996; Thiede & Dunlosky, 1999). This model proposes that differences in reading times depend on the difference between readers' judgments of item difficulty and the performance level desired. More difficult items may require longer processing and task completion time. Contrasting with this theory, a new model has been proposed recently: the *region of proximal learning* hypothesis, which proposes that the allocation of study time for each item depends on the interaction between expertise of participant and item difficulty (Metcalf, 2002; Metcalf & Kornell, 2003). Readers may decide to devote longer study time to those items that represent a moderate difficulty for their level, because easier items do not require special attention, and more difficult items induce an early task drop-out due to a roof effect.

A variation of the moving window technique was used by Goldman & Saul (1990) to analyze the processes used by students while reading a text. The major innovation introduced by these authors was to show the text in bigger text units: paragraphs or groups of sentences. In addition, readers could select the 'window' on their own, using the computer mouse, and could regress in the text, which improved the external validity of the reading process. In this same line there is the computer tool *Read&Answer*, designed by our research team, whose main innovation is that it registers on-line variables related to the question-answering process. These variables allow analyzing the reading and self-regulation strategies used by students during a learning task.

In this vein, the main objective of this work has been to develop a tool that allows testing the possibility of measuring and evaluating the on-line processes used by students while performing a comprehension test, a research topic that has not received much attention. A major contribution is the work by Farr, Pritchard & Smitten (1990), which initiated assessment of the different strategies used by students during the completion of a comprehension test. The authors divided such strategies in three groups: a) *overall strategies used*, which refers to the general approach the subject used to complete the task (if they read first the text or the questions, if they read the text continuously or if they move constantly between the text and the questions, etc.), b) *specific reading strategies*, which refer to those strategies intended to boost text comprehension, and c) *test-taking strategies*, which are related to the selection of the alternatives (Farr, Pritchard, & Smitten, 1990). A follow up of this study was performed by Cordón & Day (1996), whose objective was to test for differences between the processes related to the test completion and other reading tasks. The conclusion from this research was that both types of tasks were similar, although readers from the test

condition, unlike the other condition, did not read linearly but read as part of the information search process in which they got involved in order to answer the test questions (Cordón & Day, 1996). In the context of the present work, a restriction of both studies is that they were conducted with the think aloud method, which makes it difficult to compare the variables observed in those studies with the ones analyzed here.

In sum, the objective of the present work was to investigate the feasibility of developing a reading comprehension test using a moving window environment. Such a system would facilitate the assessment of on-line variables with the objective of detecting and analyzing the strategies used during test completion. This objective implies the use of a comprehension test that would be computerized and would include a technique for evaluating on-line processes. The adaptation of a paper-and-pencil test to a computer-based system has been extensively done before, and normally has been validated using correlational studies, which have shown moderate to high correlation indexes (v.g. Evans, Tannehill, & Martin, 1995; Maguire, Knobel, Knobel, & Sedlacek, 1991; Vispoel, Boo, & Bleiler, 2001). In addition, meta-analyses of prior studies have corroborated that generally both types of tests can be considered comparable (Bugbee, 1996; Mead & Drasgow, 1993; Wise & Plake, 1989). Even in the particular case of comprehension, research has revealed that the cognitive process involved in both types of tests might be comparable, both during text reading, and during the question-answering process (Kobrin & Young, 2003).

Method

Sample

The original version of the Test of Comprehension Processes (TPC, *Test de Procesos de Comprensión*) in paper-and-pencil was validated with 1595 students, including boys and girls from 5th to 10th grades. Our sample was controlled for sex, students' school location (rural or urban), and type of school (public or private), although the only variable that reached significance level was age or grade? of students. To be able to compare this sample with that from the computer-based version (TEC-e), in the current study we only used 804 students from 5th, 7th, and 9th grades. The validity assessment of the original test was performed using a representative sub-sample of 542 students.

To assess the validity of the Electronic Test of Comprehension Strategies (TEC-e, *Test de Estrategias de Comprensión-electrónico*) and the analysis of on-line strategies used by students we selected a sub-sample of 306 participants, distributed in the following way: 156 male and 150 females; 95 from 5th grade, 109 from 7th grade and 102 from 9th grade. We used alternate groups in order to maximize the sample size of each group in the profile

analysis, which allowed us to test for differences between each educational level. As was the case with the original test, the assessment of validity was performed using a sub-sample of 111 students, representative of the global sample.

The Test of Comprehension Processes

The comprehension test selected for the computer-based adaptation was the TPC (Vidal-Abarca, Gilabert, Martínez, & Sellés, 2007). The TPC is a reading comprehension test grounded in solid theoretical models: the Kintsch C-I model (Kintsch, 1988; Kintsch, 1998), studies on inference making by Graesser and colleagues (Graesser, Singer, & Trabasso, 1994), analysis of the text features (McNamara & Kintsch, 1996), and adaptive use of reading strategies based on the reading task and objectives (Goldman, 1997). In sum, the assessment done by the TPC is based on the analysis of basic comprehension processes identified in previous research (identification of main ideas, elaboration of inferences that connect textual elements, elaboration of inferences based on prior knowledge and the construction of macro-ideas).

The TPC is addressed to students of 10 to 16 years old, and consists of two expository texts, each followed by 10 multiple-choice questions, with four alternatives and only one correct response. The questions assess the different cognitive processes involved in comprehension. In addition the test contains a short training text with two questions that allows the evaluator to explain the test completion process. The texts were constructed controlling for vocabulary, grammatical structures, coherence and text organization, in order to assure its readability and comprehensibility. The psychometric properties of the test revealed high indexes of reliability (.798) and validity (.723) and a strong homogeneity across all test components (in all cases higher than .330). A more detailed description of this test can be found in Martínez, Vidal-Abarca, Sellés, and Gilabert (2008).

Computer-based adaptation of TPC. The TEC-e.

The Electronic Test of Comprehension Strategies (TEC-e) is a computer-based version of TPC, developed using the moving window technique, similar to that used in *Read&Answer*, described above. This technique allows the recording of on-line variables, both during text reading and during the question answering process. The TEC-e mimics TPC in terms of instructions, texts, questions, alternatives, and order. The only difference consists of the training session, which although it comprises the same text and questions as in TPC, it also includes instruction on the use of the tool: how to unmask, how to advance the reading and how to answer the questions.

The masked letters in TEC-e maintain the size, position and separation between letters and words; thus the masked text does not present structural change compared to the unmasked text. Thus, the masked text retains the peripheral

information while other window is read, or during the process of searching for information. In order to increase the pictorial cues and the identity between the masked and unmasked text only the alphanumeric characters, including tildes, are masked; whereas punctuation marks and other symbols are kept visible. (An example of a masked and unmasked text can be seen in Figure 1.) To unmask a text window, readers click on it using the computer mouse, which will automatically unmask the new text region selected while masking the previously visible text region.

The information from each text was divided in two pages, accessible through buttons on a constantly visible menu, so that scrolling was not needed. This allows maintaining a rigid structure of the information units, similar to what is apparent on a piece of paper. Thus, the interface assists the reader in the information search processes. Most of the criticisms of computerized texts are related to how the information is visualized. Current programs do not allow showing a big amount of text in each screen, which requires the use of the scrolling in long texts. However, scrolling has also been criticized because it can hide contextual clues during the information search process (Choi, Kim, & Boo, 2003).

The texts in TEC-e are divided in several text regions that correspond to the moving windows. The criteria used to divide the text were that each window should contain at least one complete idea and that those parts necessary to respond to a question were isolated. The first text consisted of 550 words and was divided in two pages and 11 regions, of a variable length in between 17 to 78 words. The second text consisted of 473 words and was divided in two pages and 9 regions, which ranged from 29 to 74 words. The training text included 237 words and was divided into 5 segments shown on only one page. The 20 test questions were also shown in a masked way, with two segments each one. In one region the question was presented and in second region the four choices appeared.

While working with TEC-e, the student starts in a screen with a text, which is shown totally masked (see Figure 1). In order to read the text the student will have to unmask the different regions with the computer mouse. Once the student has read the entire text, the button 'Questions' is activated, allowing the student to answer the questions. The fact that students can only access the questions once the text has been read completely is because the application instructions of the TPC require the student to read all the text before proceeding to the questions. The process is similar for the two texts and the corresponding questions.

The *question* screen is divided in two regions that are shown in a masked way; the upper part of the screen presents the question and the central part the response alternatives (see Figure 2). In this screen the student has first to read the question, clicking on it and keeping the mouse button pressed. Then if students move the mouse to a different region or if they release the mouse button, the question is masked again. Once the question has been read the response alternatives are activated and can then be unmasked after the procedure described above. Once the question and the response alternatives have been read, the program activates the *response options* and the button *Turn back to the text*. In other words, once students have read the questions and the different response alternatives, they can decide between answering or coming back to the text to look for information before answering. After students respond to the question they can proceed to the next question, although they are allowed to go back to a previous question and to change a previous response.

The main advantage of this tool over the original paper-and-pencil test is that it allows us to measure for how long and how many times students read each region, question or response alternatives. In addition, the program performs an automatic correction of the test.

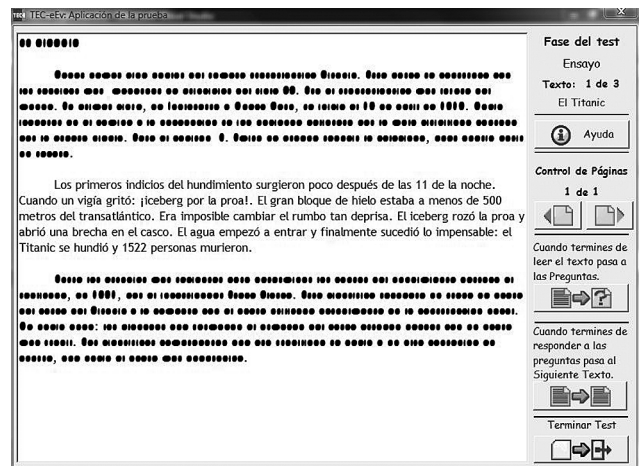
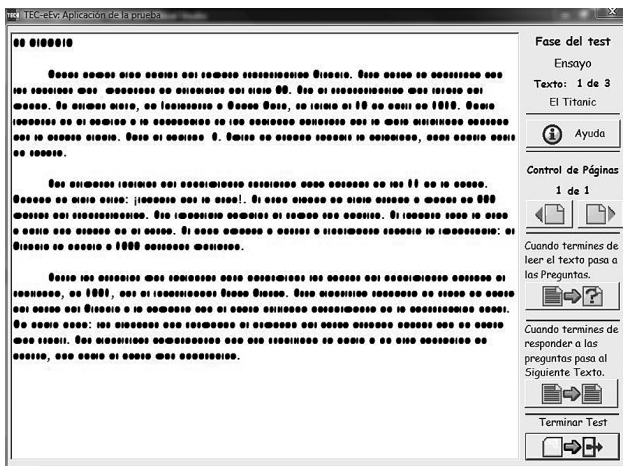


Figure 1. Screen capture of masked and unmasked text by TEC-e.

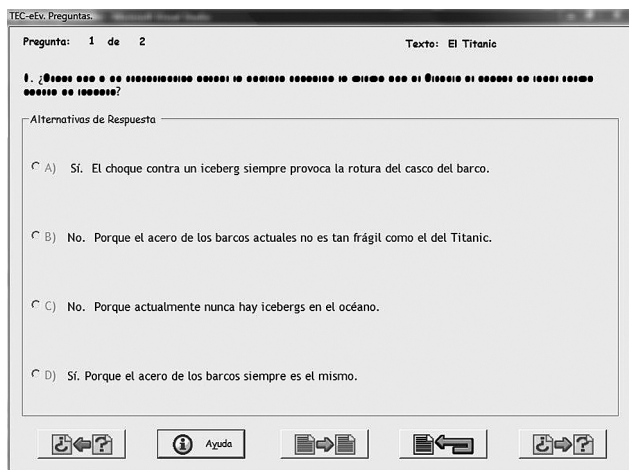
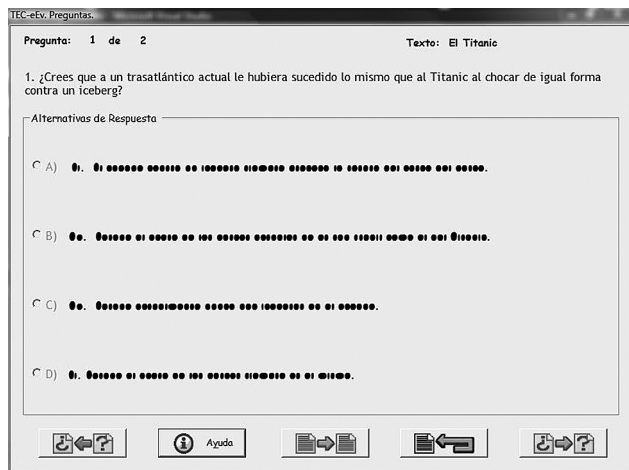
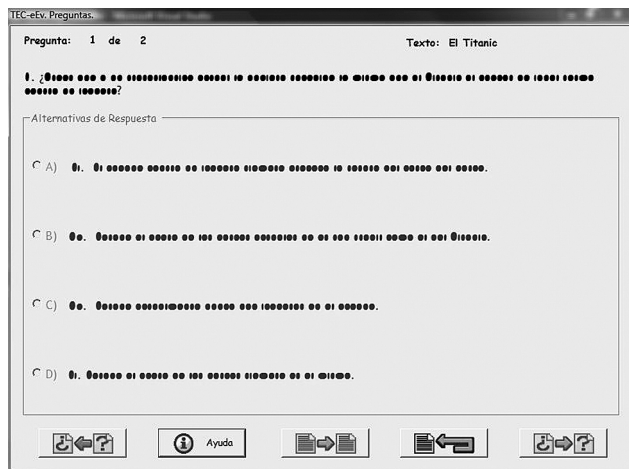


Figure 2. Screen captures from the question presentation screen by TEC-e with different regions masked.

Variables and measurement instruments

The variables used in order to compare both tests were a set of descriptive variables (age, course, gender, rural or urban area, type of school and reading habits), the TPC test scores obtained by the initial population of 1595 students, and TEC-e test scores obtained by the new sample of 206 students. The test scores are based in the sum of correct responses without considering errors or missed cases, thus they range from 0 to 20.

The analysis validity of the two tests was conducted using the comprehension subtest from the PROLEC-SE test (Ramos & Cuetos, 1999), a test aimed at assessing the cognitive processes involved in reading (lexical, syntactic and semantic) of students from the third cycle of primary schooling (10-12 year olds) and compulsory secondary education (12-14 year olds). This subtest also includes two texts and a set of open questions that students have to respond to after reading the texts. In our research not all students completed PROLEC-SE, but only a representative sub-sample from the two original samples, consisting of 542 students from the TPC sample, and 111 from the TEC-e sample.

The second part of our study, which focused on the assessment of the strategies used during test completion, was based on the measurement of a set of on-line variables. First, we measured *Reading speed*, both for the text and for the questions and the response alternatives, because this is the main variable used in most moving window techniques. The reading speed was measured in milliseconds per words, based on the unmasking time. We also included other variables that could allow us to replicate the results about strategy use found in previous studies, although we should recall that those were performed using the think aloud technique. Some of the variables used in those studies could be assessed without the need of student verbalizations, such as: a) *Rereading behaviour*, both for the text and for the questions, coded in our study as *Number of visits* and *Percentage of First Reading* (percentage of time devoted in the first reading over the total reading time) and b) *Information Search Processes*, which is coded as *Information Search* (a dichotomy variable that indicated whether students went back to the text after reading a question).

Finally, we included a control variable, *Lexical Access*, measured also using PROLEC-SE. The subtests used in this case were: *Word reading*, *Pseudo-word reading*, and *Lexical decisions*, from which we measured both number of errors and reading speed.

Performance profiles

To be able to analyze the results of the on-line processes it was necessary to divide the tasks and students into categories, based on question difficulty and test scores respectively. Question difficulty was used as dependent

variable in the analysis of reading times because the time invested in each question depends on task demands, as asserted by both the Discrepancy model and the Region of Proximal Learning hypothesis described above. In order to categorize questions we used as a criterion the Index of Question Difficulty (i. e., Number of Correct Responses / Number of Participants), from which we obtained three ranks with six questions each: *Easy* (indexes from .12 to .28), *Moderate* (indexes from .40 to .55) and *Difficult* (indexes from .61 to .84). The two remaining questions, situated between these ranks, were eliminated to maximize the differences between categories and to keep the number of questions in each rank balanced.

In addition, student categories were established using their own test performance, with the aim of providing only a descriptive analysis. Assessing comprehension processes does not necessarily imply evaluating final comprehension. Indeed, to interpret comprehension processes it is necessary to know the effects that these processes have on the product (Myers, 1991). In this sense, students were classified in three levels based on the percentile obtained in TEC-e: High (with a percentile above 75), Medium (between 35 and 65) and Low (with a percentile below 25). It is important to stress that students' classification was not affected by the use of the particular normative samples of TPC and TEC-e, because both are totally comparable. We excluded from the study students from percentiles between the three levels used, in order to improve intra-group homogeneity and to maximize between-group differences.

Results

Correspondence between TPC and TEC-e

The first and main objective of this study was a computer-based adaptation of the TPC based on the moving window technique. To evaluate the similarity between the original and the new test we compared their descriptive indexes, for both tasks and items, including reliability, validity and distribution by course.

We first compared the mean scores (TPC: 11.11; TEC: 10.69) and standard deviations (TPC: 4.18; TEC: 3.83) of each of the two versions, although we did not use a mean comparison because we had grouped scores from different courses, and the distributions of each course were not similar for both test samples. To verify if the means were similar we used an ANOVA (inter 2 x 3), using as fixed factors the test version (computer-based or paper-and-pencil), and course (5th, 7th and 9th grade), and test scores obtained on each version of the test as dependent variable. The statistic used to perform the ANOVA was the *Sum of Squares Type III*, due to the differences on sample size on each of the test versions. The results obtained in this test allow us to conclude that both test versions are statistically comparable.

First, the homogeneity of the error variance, between versions, showed no significant differences between test version, either if analyzed with the Levene test ($F=1.89$; $df(1)=5$; $df(2)=1122$; n.s.), or with the F test ($F=1.72$; $df=1$; n.s.). As expected, we obtained significant differences between course levels ($F=83.84$; $df=2$; $p<.000$), revealing that both test versions correctly discriminated between courses. In addition, the interaction between these two variables was not significant ($F=2.012$; $df=2$; n.s.), thus showing that the adaptation was equally valid for all courses. Finally, we introduced in the analyses the rest of descriptive variables (gender, rural or urban area, etc.) but, as was the case with the TPC validation, none of these variables resulted as being significant.

Regarding item analysis, globally we obtained satisfactory results. Considering the *Difficulty Indexes* of each item, both test versions showed similar indexes with differences inferior to 0.09. Only items 2 (TPC: .52; TEC-e: .64), 4 (TPC: .43; TEC-e: .30) and 5 (TPC: .53; TEC: .66) of the Penguin text and only item 2 of the Sioux text (TPC: .65; TEC: .54) showed indexes slightly different between both versions of the test. Similarly, the *homogeneity* of all elements revealed similar indexes; the items that showed a higher difference were: Penguin 1 (TPC: .327; TEC-e: .194), 5 (TPC: .338; TEC-e: .210), 6 (TPC: .302; TEC-e: .155) and 9 (TPC: .405; TEC-e: .265) and item 9 of Sioux text (TPC: .465; TEC-e: .305). These slightly lower indexes were responsible for the lower reliability of the computer-based version of the test (TPC: .79; TEC: .72).

Regarding the Validity analysis, (TPC: .723; TEC: .637), we performed a Pearson correlation between *Total Scores* obtained on each of the versions, and scores on the comprehension subtest of PROLEC-SE. Results showed a similar trend with different indexes, although slightly lower on the computer-based version. In sum, both tests revealed similar characteristics, although the computer-based version introduced an additional source of variance that diminished slightly its reliability and validity.

Profile analysis

We used two types of ANOVAs in order to perform the profile analysis: one for the dependent variable related to text reading, and another for those related to question answering. In the first case we performed a simple between-subjects ANOVA, with Levels of Test Performance (High, Medium and Low) as an independent variable (there was no grouping by difficulty based on reading). For variables related to question answering processes we used mixed ANOVAs (3 x 3), adding the factor Question Difficulty (Easy, Moderate and Difficult) to the independent variable Levels of Test Performance.

Analysis of initial text reading. During initial text reading we considered two dependent variables: Number of Visits,

Table 1
Relation between lexical access variables from PROLEC-SE and reading speed and TEC-e scores

| | Reading Time | | | Reading Errors | | |
|---------------|--------------|---------|---------|----------------|--------|---------|
| | Words | PsW | LD | Words | PsW | LD |
| Reading Speed | -.565** | -.401** | -.301** | -.289** | -.124* | -.239** |
| TEC-e scores | -.259** | -.264** | -.186** | -.254** | -.114* | -.150** |

Note: Reading speed measured as words by second. Words = word reading subtest. PsW = pseudoword reading subtest. LD = lexical decision subtest. Statistical significance: * $p = 0.05$ and ** $p = 0.001$.

that is, the number of times that a text region was unmasked, and Reading Speed. Results showed that Number of Visits to each region was not related to Level of Test Performance ($F = 1.244$; $df = 2$; n.s.), probably because most readers read initially the text linearly, that is, with only a few regressions, which implied only one reading of each region. On the contrary, Reading Speed was indeed related to Performance Level ($F = 7.084$; $df = 2$; $p < .001$). Follow up analyses showed significant differences between all groups: students reading the text faster obtained higher test scores. These results indicate that a higher investment of time during reading did not mean a higher level of processing, but the existence of a difficulty for the student. Indeed, Reading Speed correlated significantly with the Lexical Access variables evaluated (see Table 1).

To better understand these results we also analyzed the relation between Lexical Access and students' comprehension scores. Significant correlations were obtained in all cases, showing a slightly relation between comprehension and lexical access. Indeed, after controlling for Lexical Access through an ANCOVA, the relation between Reading Speed and Performance Level disappeared. In this sense, the variable most related to performance during initial reading was the speed of lexical access, which actually determines Reading Speed.

In a second group of analyses, focused on Question Answering, we obtained several significant effects. First,

the analysis of Visits to the Question resulted significant both for Question Difficulty—within participants analysis— ($F = 66.717$; $df = 1$; $p < .000$), and for the interaction with Levels of Performance ($F = 17.277$; $df = 2$; $p < .000$). The main effect of Levels of Performance was not significant. In Figure 3 the reader can see how students with medium and high levels increased the number of visits together with question difficulty, whereas students with a lower performance level did not show this increase; in fact they kept constant their number of visits independent of question difficulty. These results were similar to those obtained through the analysis of Percentage of First Reading, because lower visits were linked to a higher percentage of first reading; therefore, they did not provide additional information.

Results of the analyses for *Questions' Reading Speed* also revealed significant differences for the within participants variable, Question Difficulty, ($F = 65.370$; $df = 1$; $p < .000$), the between-participants variable, Levels of Performance, ($F = 12.749$; $df = 1$; $p < .000$), and their interaction ($F = 4.819$; $df = 2$; $p < .009$). These pattern of results indicated that students distributed their resources based on task difficulty. Similar to what we found with the first reading data, the between-participants effect could be explained as a function of differences in Lexical Access, because previous research has reported a robust relation between Reading Speed, Lexical Access and Performance Level. Figure 4

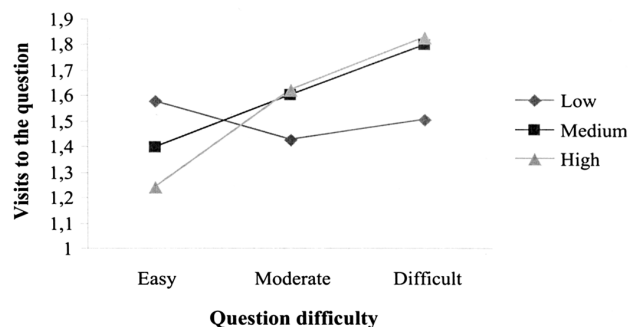


Figure 3. Number of visits to the question as a function of performance level and question difficulty.

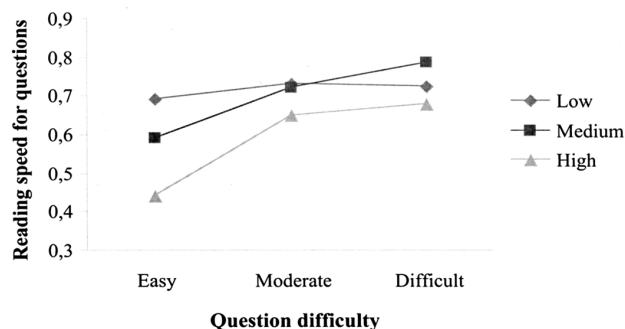


Figure 4. Reading speed for questions as a function of performance level and question difficulty.

Note: reading speed measured as words by second.

shows how participants of a higher performance levels varied their speed as a function of difficulty, while participants with lower performance levels kept their speed constant. This may indicate that either they did not perceive the difference or they simply did not possess more optimal strategies or cognitive resources.

The analysis of on-line variables related to the *Response alternatives* showed similar effects to those found for questions. For *Visits to Response alternatives*, results were significant for the within-participants variable ($F= 38.051$; $df=1$; $p<.000$) and for the interaction with Levels of Performance ($F= 21.041$; $df=2$; $p<.000$). The main effect of the between-participants variable was not significant, as was the case with the effects on question reading. As can be seen in Figure 5, there was a clear distinction in the patterns followed by students as a function of question difficulty. Students with lower performance did not increase the number of re-reading as a function of difficulty, but indeed they decreased their re-readings, contrary to what was found for the other two groups. This result may indicate a sort of 'surrounding' behaviour when low performers are confronted with more difficult tasks, which is consistent with the predicted effect by the Region of Proximal Learning model described above.

Regarding the variable *Response alternatives Reading Speed*, results also confirmed significant effects for the within-participants variable ($F= 22.772$; $df=1$; $p<.000$), the between-participants variable ($F= 11.607$; $df=1$; $p<.020$) and for their interaction ($F= 24.456$; $df=2$; $p<.000$). Nevertheless, the pattern of behaviours was completely different to that found during question reading. In figure 6 the reader can see how lower performers reduced the resources devoted to reading the response alternatives as the question difficulty increased, which could support the previously stated 'surrounding' hypothesis. Students performing at a medium level kept constant the time devoted to reading the response alternatives, whereas high performers were quicker in reading the response alternatives, especially the easiest, probably because they

quickly rejected false alternatives. In this case, the relation between Lexical Access and levels of performance could not account for the effect found. After controlling for Lexical Access through an ANCOVA, the relation between Levels of Performance and Reading Speed still remained significant.

Of all variables assessed, probably the one that showed a clearer strategic behaviour was *Information Search*. This variable measured the extent to which students decided to check the text in order to answer a question. By using the same analyses as in the previous section, we again found significant differences for the within-participants variable ($F= 9.28$; $df=1$; $p<.000$), the between-participants variable ($F= 5.009$; $df=1$; $p<.007$) and for their interaction ($F= 10.39$; $df=2$; $p<.000$). In Figure 7 we can observe how participants' behaviour is clearly different as a function of Performance Level. Lower performers checked the text less, especially when answering difficult questions; probably because they realized that they would not be able to answer them, which was consistent with results found previously. Participants performing at a medium level checked the text more times, even for easy questions. They may have 'verified' their responses more often. It is important to recall that these students devoted longer times processing the response alternatives. On the contrary, high performers showed an 'intermediate pattern': they barely checked the text on the easy questions, but more often on the moderate and difficult questions (there were no significant differences between these two last groups).

A major problem with the analysis of the variable *Information Search* was the procedure followed in TEC-e, which forced students to read the text completely before attempting to answer the questions. This particular feature of the test made students often answer the questions without checking the text, which impedes the comparison of these results with those obtained in other studies discussed in the Introduction. Nevertheless, we tried to gain more knowledge about the meaning of this variable through analysis of the percentage of correct responses obtained only on questions

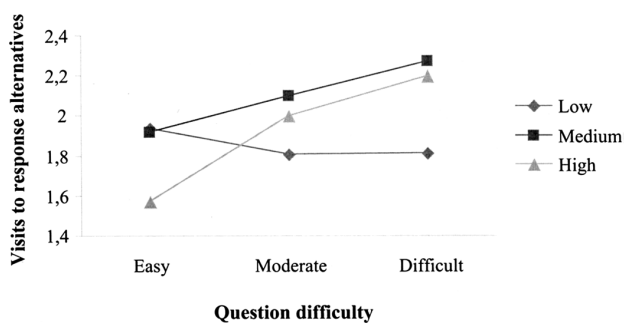


Figure 5. Number of visits to response alternatives as a function of performance level and question difficulty.

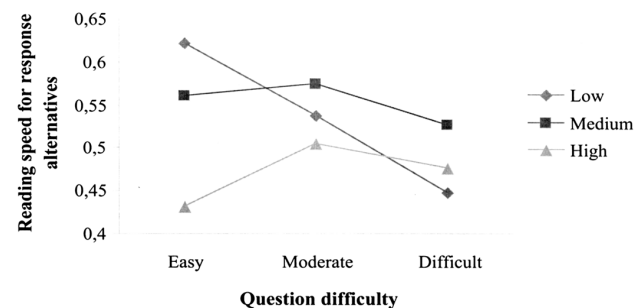


Figure 6. Reading speed for response alternatives as a function of performance level and question difficulty.

Note: reading speed measured as words by second.

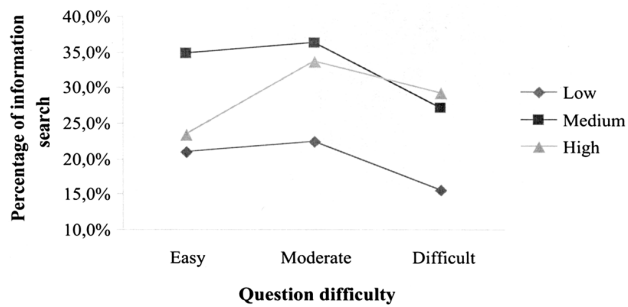


Figure 7. Percentage of times that students search information in the text as a function of performance level and question difficulty.

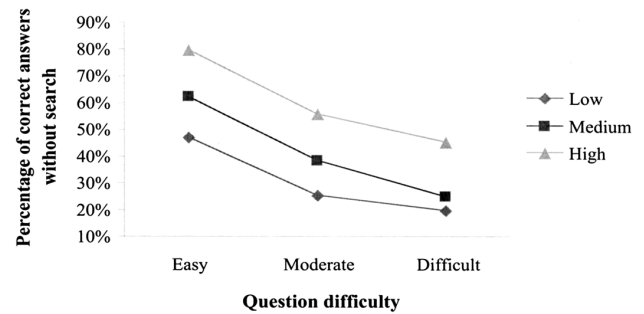


Figure 8. Percentage of times that students answer correctly without having searched for information in the text as a function of performance level and question difficulty.

that were answered without checking the text. The percentage of correct responses diminished as a function of task difficulty ($F= 127.53$; $df=1$; $p<.000$), and high performers obtained a higher percentage of correct responses without checking the text ($F= 43.91$; $df=1$; $p<.000$), which could be considered obvious provided the way we constructed the variable. Even so, the interaction between these two variables did not reveal significant effects, as can be seen in Figure 8. This relation indicated that higher levels of global performance were associated with optimal decisions on when to re-read the text in order to answer a question (and when not), which constituted an optimal use of self-regulation processes.

Discussion

The main goal of this study is the development of a computer-based comprehension test called TEC-e, which is an adaptation of a previous paper-and-pencil reading comprehension test called TPC. TEC-e provides not only an automated correction of responses but also capturing variables related to the reading process. These new process variables can draw a more complete picture of student assessment. They provide additional information to that already supplied by test scores, which can be informative on how comprehension evolves during reading. The results from the study have been satisfactory, revealing that the new version of TPC is highly similar to the original version, although the new source of variance introduced by the 'moving window' technique reduces slightly its indexes of reliability and validity. Nevertheless, this decrease does not alter significantly the measures or the distributions by course, as results on the scores obtained from the original sample and those from the computer-based version are similar.

Regarding the analyses of on-line processes it is important to note that there are significant differences in all variables assessed. These effects point to the fact that on-

line techniques can be considered as an adequate methodology to the study of comprehension processes and their assessment, and can also be easily used in natural environments. An essential limitation of the current study is that results can not be contrasted with previous research, because these were performed using the think aloud methodology, which captures variables that are substantially different from those analyzed here. In addition, the application procedure of our test is very strict: it forces students to read the texts before seeing the questions, which prevents readers from using global strategies, for example, skipping initial reading and focusing on the information search. Nevertheless, global strategies are revealed to be of great significance in previous studies of the on-line processes involved in the completion of comprehension tests (Farr et al., 1990; Rupp, Ferne, & Choi, 2006).

In any case, the results obtained from our on-line analyses seem to show that the processes followed by high and low test performers are clearly different, as the significant differences between most of the variables analyzed demonstrate. The different set of variables studied show that during initial text reading students' behaviour is really homogeneous, as determined by test instructions, which in some way prevents the involvement of individual differences. In this way we obtained a linear reading and with a unique visit per text region.

Reading times, as expected, are mainly determined by students' level of lexical access. Thus, students with a more fluent lexical access show higher reading speed, and get the best test results. In regard to processing times results reveal that lower performers devote most of their resources to lexical access, which interferes with the rest of comprehension processes. Students with better comprehension levels have a fairly automatic lexical access, which allows a more fluent reading (Perfetti, 1985; Perfetti, Beck, Bell, & Hughes, 1988).

Regarding the strategies used while reading the questions, low performers behave in a less adaptive way: they deploy similar behaviours independent of question difficulty. In

addition, processing times from these students show that, again, they read the texts more slowly; which is not an adaptive behaviour. Nevertheless, medium or high performers reveal a more adequate distribution of their resources: they devote greater amounts of time and visit to more difficult questions.

With respect to students' behaviours during reading the response alternatives, results reveal that they slightly change in each of the three groups. Low performers are again less adaptive and visit the alternatives the same number of times independently of task difficulty, whereas medium and high performers increase their number of visits together with task complexity. Clearly, the most intriguing effect of this set of analysis is the results of processing times. Low performers keep reading slowly only the easy questions, whereas with questions of medium or high difficulty, notably, they increase their reading speed. In these cases, students may not be completely reading the response alternatives, which may indicate either a surrounding behaviour after a low comprehension of the task, or the use of low levels strategies, such as lexical matching. Students with a medium level of performance keep their reading speed constant, similar to when they read the text, but in this case they are not so adaptive because they are not influenced by task difficulty. Only high performers, as expected, show a more strategic behaviour: again they read easy questions faster, probably because they quickly reject incorrect response alternatives, whereas they slightly decrease their speed on medium and difficult questions. These differences in reading times for the response alternatives are similar to those found during the information search task, where high performers access text segments relevant to the answer, open fewer text regions and for less time. Low performers, on the contrary, show an indirect search pattern, which makes them slower (Cataldo & Oakhill, 2000).

Finally, with regard to searching behaviours, again we find three different patterns: low performers show fewer searching behaviours, and they do not adapt them based on task difficulty. Students with medium performance levels search for more information, mostly in questions of easy and moderate difficulty, in which they verify their responses. Nevertheless, this behaviour decreases dramatically in difficult questions, which logically are the ones in which students committed more errors. Finally, high performers engage in a more strategic search pattern: search fewer times while answering easy questions, and increase their searching pattern equally for medium and difficult questions. Results of high and medium performers are in line with those obtained with other experimental procedures based on the judgments students make on the probability of correctly answering a question, which is one of the key factors influencing search behaviour. Studies that have focused on this issue show those students are confident in difficult

questions, while they may hesitate in easy questions (Schraw & Roedel, 1994), a similar effect to that found in our searching patterns.

In sum, results seem to reveal that the ability to adapt to task difficulty by using a strategic processing is responsible for the performance level obtained in the comprehension test. These analyses help us to understand the differences between good and poor readers, although we cannot draw a causal relation. We should stress that we have been able to develop a valid analysis tool that can complement the results from the comprehension test, which increases its diagnostic power.

Finally, we would like to address the debate between the Region of proximal learning hypothesis and the Discrepancy reduction model. Our results clearly show a better fit with the Region of proximal learning hypothesis, because students do not devote a higher quantity of cognitive resources (i.e., reading times) to the more difficult questions, but they seem to adjust their abilities to task difficulty, as predicted by the theory of Metcalfe and colleagues (Metcalfe, 2002; Metcalfe & Kornell, 2003). In our study the time devoted to each task depends primarily on students' comprehension ability: they invest more time in those questions that have a moderate complexity for their level, and they reduce the resources used for really easy or really difficult tasks. This pattern contradicts the discrepancy reduction model, because for this model students should have spent more time on more difficult tasks.

Referentes

- Aaronson, D. & Ferres, S. (1983). Lexical categories and reading tasks. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5), 675-699.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28(3), 282-299.
- Carver, C. S. & Scheier, M. F. (1990). Origins and functions of positive and negative effect. A control-process view. *Psychological Review*, 97(1), 19-35.
- Cataldo, M. G. & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text representation and spatial memory on the ability to locate information in text. *Journal of educational psychology*, 92(4), 791-799.
- Choi, I., Kim, K. S. & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320.
- Cordón, L. A. & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of educational psychology*, 88(2), 288-295.

- Entin, E. B. & Klare, G. R. (1985). Relationships of measures of interest, prior knowledge, and readability to comprehension of expository passages. En B. Hutson (Ed.) *Advances in Reading/Language Research* (pp. 9-38). Greenwich: JAI Press.
- Evans, L. D., Tannehill, R. & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. *Behavior Research Methods, Instruments & Computers*, 27(2), 162-165.
- Farr, R., Pritchard, R. & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27(3), 209-226.
- Goldman, S. R. (1997). Learning from text: Reflections on the past and suggestions for the future. *Discourse Processes*, 23(3), 357-398.
- Goldman, S. R. & Saul, E. U. (1990). Applications for tracking reading behavior on the macintosh. *Behavior Research Methods, Instruments & Computers*, 22(6), 526-532.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative test comprehension. *Psychological Review*, 101, 371-395.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4), 329-354.
- Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2), 228-238.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction integration model. *Psychological Review*, 95(2), 163-182.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Kintsch, W. & Kintsch, E. (2005). Comprehension. En S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment*. (pp. 71-92). Mahwah: Lawrence Erlbaum Associates Publishers.
- Kobrin, J. L. & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education*, 16(2), 115-40.
- Koriat, A. & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103(3), 490-517.
- Magliano, J. P., Millis, K. K., Ozurur, Y. & McNamara, D. S. (2007). A multidimensional framework to evaluate assessment tools. En D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107-136). Mahwah, NJ: Erlbaum.
- Magliano, J. P., Millis, K. K., The NIU R-SAT Development Team, Levinstein, I. B. & Boonthum, C. (En prensa). Comprehension strategies and performance with the reading strategies assessment tool (R-SAT). *Journal of Educational Psychology*.
- Maguire, K. B., Knobel, M. M., Knobel, B. L. & Sedlacek, L. G. (1991). Computer-adapted PPVT-R: A comparison between standard and modified versions within an elementary school population. *Psychology in the Schools*, 28(3), 199-205.
- Martínez, T., Vidal-Abarca, E., Sellés, M. P. & Gilabert, R. (2008). Evaluación de las estrategias y los procesos de comprensión: El test de procesos de comprensión (TPC). [Evaluation of comprehension strategies and processes: The test of Comprehension Processes]. *Infancia y Aprendizaje*, 31(3), 319-332.
- McNamara, D. S. & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3), 247-288.
- McNamara, D. S., O'Reilly, T. P., Best, R. M. & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, 34(2), 147-171.
- Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.
- Metcalf, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, 131(3), 349-363.
- Metcalf, J. & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, 132(4), 530-542.
- Myers, S. S. (1991). Performance in reading comprehension -product or process? *Educational Review*, 43(3), 257-272.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.
- Perfetti, C. A., Beck, I., Bell, L. C. & Hughes, C. (1988). *Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children*. Detroit, MI: Wayne State University Press.
- Ramos, J. L. & Cuetos, F. (1999). *PROLEC-SE: evaluación de los procesos lectores en alumnos de tercer ciclo de educación primaria y secundaria*. [PROLEC-SE: evaluation of reading processes for students of third cycle of primary and secondary school]. Madrid: TEA.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441-474.
- Schraw, G. & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & cognition*, 22(1), 63-69.
- Thiede, K. W. & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(4), 1024-1037.
- Trites, L. & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing*, 22(2), 174-210

- Vidal-Abarca, E., Gilabert, R., Martínez, T. & Sellés, M. P. (2007). *Test de estrategias de comprensión*. [Test of Comprehension Strategies]. Madrid: Instituto Calasanz de Ciencias de la Educación.
- Vispoel, W. P., Boo, J. & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the rosenberg self-esteem scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement*, 61(3), 461-474.
- Wise, S. L. & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5-10.

Received November 14, 2007

Revision received July 22, 2008

Accepted September 4, 2008