PSICOLÓGICA

Validation of a reading comprehension efficiency test for Spanish university students

Ladislao Salmerón[1], **Lidia Altamura**[1], **M. Carmen Blanco**[2], **Sandra Montagud**[1], **and Cristina Vargas**[1]

[1] *ERI Lectura, Universitat de València.*
[2] *Universidad de Zaragoza.*

Abstract

Despite the importance of reading literacy in adulthood, research in this field in Spanish-speaking countries is limited due to the lack of a valid tool to assess adults' reading comprehension in their native language. Thus, we aimed to adapt the Deep Cloze test by Jensen and Elbro (2022) to Spanish and validate it. 453 undergraduate students from the universities of Valencia and Zaragoza completed the Spanish version of the Deep Cloze test. In addition, to test its criterion validity, subsets of the sample completed other text comprehension measures or provided their grades in introductory courses four months after completing the Deep Cloze test. Results indicated that while the Deep Cloze test has good internal consistency, correlations with other text comprehension measures and grades on a final exam are mostly small. The Deep Cloze fills an important gap by facilitating the assessment of reading comprehension efficiency in Spanish undergraduate students. This test presents important advantages, such as simplicity and administration time, and it can be used for research purposes. Finally, we discuss future research needed to test its validity to be used as a diagnostic tool to help deliver targeted support to struggling adult readers.

Keywords: reading comprehension, pedagogical assessment, undergraduate students.

Introduction

Reading literacy in adulthood has been related to better indicators of health and well-being (Mumper & Gerrig, 2017). Reading is much more than a hobby or obligation because it improves our way of understanding, thinking, and communicating. In the Spanish context, the focus on reading literacy has mainly involved the early stages of education, mostly ignoring undergraduate students. Accordingly, research on reading comprehension assessment in Spanish focuses on primary and secondary school students (ACL, Catalá et al., 2001; CompLEC, Llorens et al., 2011; ECOMPLEC, León et al., 2012; PROLEC-R, Cuetos et al., 2007; PROLEC-SE-R, Cuetos et al., 2016; TALE, Toro & Cervera, 1980; TEC-e, Martínez et al., 2009). Tests evaluating adults address reading comprehension in the field of learning Spanish as a foreign language (Marchante, 2021; Quezada & Westmacott, 2019), or they assess adult literacy at the country level, but without providing individual assessments of each participant, as in the Survey of Adults Skills (OECD, 2012). Thus, to date, there is no widely accepted test to measure reading

Psicológica (2022)

DOI
10.20350/digitalCSIC/14809

Corresponding author
Ladislao Salmerón
Av. Blasco Ibáñez, 21.
46010 Valencia, Spain.
ladislao.salmeron@valencia.edu

comprehension in the Spanish adult population, either for diagnosing and treating reading comprehension problems or simply for research purposes.

One of the most popular reading comprehension tests in the 1970s was the cloze-type test, which estimated students' reading comprehension level based on how well they chose the words that best fit into the syntactic structure (McBain & Mhunpiew, 2014). However, several years later, cloze tests are highly criticized, precisely because they measure grammatical and linguistic knowledge more than discourse level representations (Kintsch & Yarbrough, 1982). From that time on, and with the cognitive revolution, any comprehension measure that did not require integration across sentences was ruled out because cloze gaps could be inferred simply by using the closest context, the sentence itself, which would be 'local' comprehension (Carlisle & Rice, 2004). Furthermore, these cloze tests did not correlate with other measures of reading comprehension, such as question answering tests (Shanahan & Kamil, 1983).

Cloze tests are general indicators of proficiency (Hadley & Naaykens, 1999), and despite all the criticism received, they continue to be regarded as useful tools if the type and number of gaps in the text are considered (Gellert & Elbro, 2013; Kleijn et al., 2019). Traditional cloze tests systematically place gaps every 5, 6, or 7 words (regardless of the context), and they are only valid for measuring readability, but not reading comprehension, because they do not require the integration of information across sentences (Kintsch & Yarbrough, 1982).

Currently, a variant of the traditional cloze test has been designed, Deep Cloze tests, where gaps are chosen that go beyond the comprehension of the single sentence, requiring global comprehension processes. According to current reading comprehension models, readers construct a mental representation to understand texts, starting from the ideas provided in the text (text-based representation) (for a review of reading comprehension models see McNamara & Magliano, 2009). Proficient readers go beyond this representation and construct a situation model that integrates readers' background knowledge to fully interpret the situation described in the text. The situation model gives readers the ability to integrate information in order to maintain the coherence of the representation, while weaving this story into what is being read (Kintsch, 1998). At first glance, and depending on the text level, any of the Deep Cloze word choices can fill this gap, but once the reader is able to figure out the situation described in the text, only one word fits correctly (Gellert & Elbro, 2013; Kleijn et al., 2019). The HyTeC-cloze is a recently validated variant that combines a mechanical-rational gap deletion strategy and semantic scoring of answers (Kleijn et al., 2019). In the same line, Gellert and Elbro (2013) proposed establishing comprehension-demanding gaps, where the gaps have to be filled by attending to more information than what is included in only one sentence by making inferences across different sentences. The authors applied the Deep Cloze test in a Danish adult sample, and the results showed high correlations with a standard question-answering comprehension test. In a latter unpublished study (Elbro & Jensen, 2017), the authors improved the previous test by going beyond the word and sentence level, and applied it to a sample of students from adult courses in Danish language and literature, with a reading level corresponding to 9-10th grade. Specifically, they created short texts with gaps that require making a global inference by first selecting the setting (situation model) and then the word that fits the gap and the setting. The results showed that the Deep Cloze test scores explained unique variance to global reading comprehension even after controlling for lexical measures (Jensen & Elbro, 2022). In interpreting these results, it should be considered that Danish, as opposed to Spanish, is an opaque language with an inconsistent system of grapheme – phoneme correspondences. To the extent that language transparency has a strong influence on reading acquisition (Ardila & Cuetos, 2016), caution should be taken in generalizing those results into Spanish.

Thus, given the importance of having a valid tool for research or initial diagnosis to evaluate reading comprehension efficiency in the Spanish adult population, the main aim of the present study was to adapt the Deep Cloze test by Jensen and Elbro (2022) to Spanish and validate it in Spanish university students across different studies. Table 1 provides an overview of the goals of each study, as well as the criterion used to test specific aspects of the test's criterion validity. First, we describe the procedure undertaken to adapt the test to Spanish and its psychometric properties (Study 1). Next, we describe a set of studies carried out to analyze the test's criterion validity, i.e. the extent to which the operationalization of the Deep Cloze relates or predicts a theoretical representation of the construct (AERA, APA & NCME, 2014). Criterion validity is usually divided into concurrent and predictive validity, depending on the time between measurements of the test and the criterion. Accordingly, studies 2a-2b analyze the Deep Cloze concurrent validity, by measuring during the same session the test together with other criteria: a standardized test for adolescents, as well as comprehension questions about a scientific text written for an undergraduate audience. Finally, Study 3 analyzes the Deep Cloze predictive validity, by measuring the criterion variable (in this case, grades on a final exam), 3-4 months after having administered the test.

Table 1. Overview of the goals of each study.

Study	Goal	Criteria used
1	Adaptation and Reliability	
2a	Criterion validity (concurrent)	PROLEC-SE-R (Cuetos et al., 2007)
2b	Criterion validity (concurrent)	Long scientific text (human learning)
3	Criterion validity (predictive)	Final exam grade

Study 1. Adaptation of the Deep Cloze test to Spanish

In Study 1, we describe the process of adapting the Deep Cloze test to Spanish and evaluate its reliability.

Method for Study 1

Participants

The final version was tested in a sample of 453 undergraduates from the Universities of Valencia and Zaragoza, Spain. Thirty-six students were excluded due to technical problems in registering their data (*n* = 9), learning disabilities (*n* = 15), and missing tasks (*n* = 12), yielding a final sample of 417 students. The sample consisted of 86.5% women, with a mean age of 21 (*SD* = 4.7). Parents' main spoken language at home was mostly Spanish (76.8% for mother's language). Respondents studied Psychology (28.7%), Speech and Language Therapy (10%), Teacher Training (42.7%), or postgraduate studies (15.4%). The undergraduate students were in their first (60.4%), second (30.5%), third (7.65%), or fourth (1.5%) year.

We applied the recommendations by Wilson Van Voorhis and Morgan's (2007) to determine the minimum sample size required for the analyses in the different studies: no less than 50 participants for a correlation, and $N > 104 + m$ (where m is the number of predictors) for testing individual predictors.

Materials

Deep Cloze test. We adapted the Deep Cloze test, developed by Jensen and Elbro (2022) in Danish, from an English version provided by the authors. The original test contains 34 short daily-life stories, ranging from 2-4 sentences, with 24-59 words each story. In the middle or end of the story, a word is missing, and four options are provided. All four alternatives fit grammatically and are thematically relevant in the sentence. The correct alternative is the most coherent one with the gist of the story. Thus, to respond correctly, readers must infer the gist of the story because they cannot merely rely on a textual representation of the story to answer properly. For example, in the story "They were seated on the terrace in shorts with bare upper bodies. They had fetched some fruit juice with ice. It became necessary to set up the ___ [tent, stage, parasol, loudspeakers]", all the alternatives were nouns and corresponded to objects that can be set up on a terrace. Thus, if readers only rely on the textual representation, they will conclude that all four are correct responses. Readers must infer that the protagonists of the story are having fun on a terrace on a sunny day ("in shorts with bare upper bodies"), in order to understand that they risk getting sunburned. Thus, it becomes evident that the correct word is "parasol", given that it is the only object of the four available that would be used on a terrace to prevent sunburn.

When translating the test into Spanish, we tried to come as close as possible to the original version. However, minor changes due to the grammatical characteristics of Spanish were needed. Specifically, when the gender and/or number of the four options differed, we included each specific determinant along with the alternative. Thus, in the example above, we used *[la carpa, el escenario, la sombrilla, los altavoces]*. In addition, we adapted one story to make it more familiar to the Spanish audience. Specifically, the story "During the break they were playing outside in the cold. It gave him a shock when he was hit. His back was quickly soaked. He wanted to take revenge and hurried to collect some [balls, mud, sand, snow]", we changed the initial sentence to "He decided to have a rest outside with some classmates while the others kept on skiing" because in Spain snow is typically associated with mountains and skiing.

Procedure

The initial version was piloted in a sample of 27 undergraduate students who completed the task in paper format, following the original instructions, which included an example with the solution. They had 10 minutes to complete the test. After completion, participants were debriefed to assess any potential difficulties. Overall, the initial version of the test was rather easy for this population. Eighteen of the participants completed 28 items, and five of them finished all 34 items. Five items were answered correctly by all 27 participants, and eight other items were answered correctly by at least 24 participants. Thus, we modified the original version in order to make the easiest stories harder, using the following strategies: a) delete parts of the story to make it harder to infer the gist. For example, in the story "The table was already laid. There was a slightly burnt smell although the timer had not gone off yet. He adjusted the knob and removed the [duvet, papers, radio, pan]", in the modified version we removed the reference to the knob, and so the third sentence was changed to "Next, he removed the"; b) when there was a clear association between the verb and the correct noun, we changed the critical verb and used different nouns to make the four options equally salient. For example, in the story "She struggled with her son's [homework, zip, duvet, painting], but it was difficult because he would not stand still", we changed the original verb "was" to the action verb "subir" (to go up, to increase) and the nouns to "stairs, zip, prices, volume". In addition, we restricted the completion time to eight minutes to ensure that none of the participants could finish the test. The final version of the adapted test consisted of 34 short daily-life stories ranging from 2-4 sentences, with 19-58 words each story (see the final version of the text in Annex I).

In all studies reported here, participants provided an informed consent form before starting the study and completed the tasks using an online tool. The tasks in all studies were carried out online, using the moodle platform of the University of Valencia. Participants were given access to the platform and the duration of the experiment was indicated in advance so they could plan to carry it out in a single session. The Deep Cloze test was automatically programmed to close the task after the 8 minutes limit.

Results for Study 1

To test psychometric properties, we evaluated evidence of the test's reliability, using the ordinal Omega coefficient, which is the most appropriate one for use with tests like ours (Viladrich et al., 2017), using the 'psych' package (Revelle, 2022) from R 3.6.3 (R Core Team, 2020). Scores on the Deep Cloze test were normally distributed (see Table 2).

Omega coefficient was .77. Furthermore, based on Gellert & Elbro (2013), we ran an additional analysis of reliability because the Deep Cloze test combined accuracy and speed, and the value could be overestimated. We found a reliability value of .73 with the first 19 items attempted by all participants. Therefore, the evidence suggested an acceptable level of reliability.

Discussion for Study 1

The Deep Cloze test was found to have appropriate internal consistency. A major difference with the original test is that we used eight minutes as a time limit instead of 10 because many participants finished the test with a high threshold in the pilot study. The need to adjust the time limit may stem from potential literacy differences between the samples used. The original Danish sample used by Elbro and Jensen (2017) included adults from a vocational training program, who, on average, may read slower than the sample of undergraduate students in our study.

Table 2. Descriptive statistics (raw scores) for all measured variables in Study 1 (n = 453), Study 2a (n = 99), 2b (n = 118), and 3 (n = 113).

Variable	M	SD	Skewness	Kurtosis
Study 1				
1. Deep Cloze	18	4.42	0.04	-0.09
Study 2a				
1. Deep Cloze	18.72	4.21	-0.19	-0.63
2. PROLEC-SE-R	15.40	2.12	-0.05	-0.81
Study 2b				
1. Deep Cloze	18.42	4.14	0.03	-0.17
2. Prior knowledge	5.29	1.30	-0.42	1.08
3. Text comprehension	7.83	2.91	0.19	-0.64
Study 3				
1. Deep Cloze	16.81	4.20	0.18	0.12
2. K-BIT (non-verbal)	38.43	4.26	-0.58	0.38
3. Exam grade (0-10)	6.58	1.70	-0.01	-0.87

Study 2. Concurrent validity

Because there are no standardized measures of reading comprehension ability in Spanish for undergraduate students, there is no clear-cut solution for testing the criterion validity (either concurrent or predictive) of the Deep Cloze test. Nevertheless, a major reference in assessing reading comprehension in Spanish is the reading subtest of the PROLEC-SE-R test (Cuetos et al., 2016), designed for adolescents from 12 to 18 years old. Our sample of interest, freshmen undergraduate students, would be in the upper extreme of the validation sample, and so they may find the test too easy. Thus, in Study 2a, we expected that scores on the Deep Cloze test would positively correlate with scores on the PROLEC-SE-R test, although the size of the relationship might be small due to ceiling effects. In addition, we aimed to further assess the criterion validity (specifically, concurrent validity) by comparing scores on the Deep Cloze test to participants' understanding of complex expository texts, measured by reading comprehension questions. Specifically, in Study 2b, we used a long authentic text published in the science-dissemination magazine *Investigación y Ciencia*. We expected scores on the Deep Cloze test will show small and positive correlations with scores on the reading comprehension questionnaires, after controlling for participants' prior knowledge about the topic.

Method for Study 2

Participants

Participants were undergraduate students who were studying Psychology, Teacher training, or Speech therapy degrees at the Universities of Valencia and Zaragoza and volunteered for class credit. In Study 2a, 102 undergraduate students participated. We excluded three participants with incomplete data (*n* = 3), resulting in a final sample of *n* = 99, *M* = 20.82 years, 85.9% women, 72.8% first year of studies. In Study 2b, 125 undergraduate students participated (26 of whom also participated in Study 2a). We excluded seven participants with incomplete data, resulting in a final sample of *n* = 118, *M* = 21.58 years, 82.2% women.

Materials

Deep Cloze test. See description in Study 1.

PROLEC-SE-R. In Study 2a, we used the reading comprehension subtest of the standardized test PROLEC-SE-R (Cuetos et al., 2016), which consists of two expository texts and 10 open-ended inferential questions for each text. The questions were corrected based on the test criteria.

Human learning text. In Study 2b, we used an expository text on human learning and artificial intelligence (see Table 3 for details). We used 21 multiple-choice comprehension questions with four options each, created in previous research (Delgado & Salmerón, 2021). Questions targeted different comprehension processes: text-based (i.e., a single idea explicitly stated in a sentence), local inference (i.e., a bridging inference linking two adjacent sentences), and global inference (i.e., a bridging inference linking information located more than two sentences apart). In the current study, after removing five items with negative loads, the test showed acceptable reliability (ω = .76). Participants rated their knowledge about subtopics related to human learning and artificial intelligence (e.g., computer programming), using an eight-item questionnaire with a scale from 1 (I know nothing) to 10 (I am an expert). Cronbach's alpha was good (α = .87).

Procedure

Participants completed the tasks using an online tool at a time and place of their choice (Studies 2a and 2b). First, participants completed a demographic information questionnaire and then the Deep Cloze test. Then, the procedure changed for each study. In Study 2a, participants completed the PROLEC-SE-R, with no time limit. In Study 2b, participants filled out a prior knowledge questionnaire, read the text or texts in a limited time, and answered the corresponding comprehension questions. In all the studies, participants were not able to refer back to the text to respond to the comprehension questions.

Results for Study 2

First, we analyzed data distributions in search of outliers (±2SD from the sample mean). Outliers in this and in the following studies were determined at the sample level. Outlier values (between 1.1-5.3% of data) were replaced by the next highest or lowest score that was not an outlier (i.e., winsorization; Field, 2013). Once outliers were removed, all measured variables were normally distributed (see Table 2, middle rows). Next, to test the criterion validity of the Deep Cloze test, we used: a) Pearson correlations between scores on the Deep Cloze test and scores on PROLEC-SE-R (Study 2a); b) hierarchical multiple regression model to predict scores on the comprehension questions on the human learning text, using prior knowledge and Deep Cloze scores as factors (Study 2b). To facilitate the interpretation of the results, we standardized all the measures prior to the analyses.

First, Pearson correlations between Deep Cloze and PROLEC-SE-R scores (Study 2a) were in the expected direction, but the pattern was not significant: *r* = .12, *p* = .23; *n* = 99. Second, a two stage hierarchical multiple regression model was conducted with the human learning text as the dependent variable (Study 2b). An examination of correlations revealed that predictor variables were not highly correlated (*r* = -.086, *p* = .357). Prior knowledge was entered at stage one of the regression as a control variable, and the Deep Cloze variable at stage two. There was no association between prior knowledge and human learning text at stage one. Adding Deep Cloze to the regression model explained an additional 5.5% of the variation variance in the human learning text and this change in R2 was significant (*p* = .011). There was a significant association between Deep Cloze and human learning text (*p* = .011). Finally, the VIF values for all predictors were close to 1 (Table 4).

Discussion for Study 2

Results from Study 2a showed that the correlation between Deep Cloze scores and PROLEC-SE-R, a widely used reading comprehension test for high school students, was rather weak (Study 2a). Nevertheless, the texts and questions on PROLEC-SE-R were designed to be sensitive to students from 13 to 18 years old. Study 2b used a text and questions explicitly designed to measure undergraduates' reading comprehension. As expected, scores on the Deep Cloze test positively predicted students' comprehension, even after controlling for the level of prior knowledge about the text's topic. Finally, although observed relations on the study using texts appropriate for undergraduate levels (i.e. studies 2b) were rather low, they are in the actual range observed in the scientific literature (Ozuru et al., 2009). In sum, results show that the Deep Cloze test presents adequate concurrent validity when tested against comprehension tests designed to test text comprehension of undergraduate students. However, it should be kept in mind that there is no gold standard to assess reading comprehension in native speakers of Spanish, and accordingly any claim regarding concurrent validity is only tentative.

Table 3. Description of the texts employed in studies 2a, and 2b.

Study	Text employed	Topic	Number of words	Legibility (INFLESZ scale) *
Study 2a	Short expository text from PROLEC-SE-R test (1)	Eskimo lifestyle	338	57.9
	Short expository text from PROLEC-SE-R test (2)	Papuan lifestyle	373	63.2
Study 2b	Long expository text	Learning process and AI systems	3010	46.7

Note: *Values of the INFLESZ scale between 55-65 indicate a normal legibility, values between 40-55 indicate a somewhat difficult legibility

Table 4. Summary of hierarchical regression analysis for variables predicting text comprehension (Study 2b; n = 118).

	Model 1			Model 2		
Variable	B	SE (B)	β	B	SE (B)	β
Intercept	-0.01	0.09		-0.01	0.09	
Prior knowledge (human learning)	0.02	0.10	.02	0.04	0.10	.04
Deep Cloze				0.24	0.09	.22*
R^2	0.00			0.06		
F for change in R^2	0.05			6.71*		

Note: *Values of the INFLESZ scale between 55-65 indicate a normal legibility, values between 40-55 indicate a somewhat difficult legibility

Study 3. Predictive validity

As a second component of criterion validity, we aimed at studying the predictive validity of the Deep Cloze test, i.e. its ability to predict other criteria measured at a different time. Reading comprehension ability supports academic learning because students must decode and integrate written information from textbooks to gain and use subject knowledge. A recent metaanalysis has established a small (r = .29) positive relationship between undergraduates' reading comprehension ability and learning achievement measured by exam grades (Clinton et al., 2022). Accordingly, we expect that scores on the

Deep Cloze test measured at the beginning of university studies will show a small and positive relation with scores on final exams in introductory courses four months later.

The magnitude of these effect sizes could be considered small according to benchmarks proposed by Cohen (1969) in the context of small-scale experiments in social psychology. However, Cohen emphasized that his benchmarks were "recommended for use only when no better basis for estimating the index is available" (Cohen, 1988, p. 25). One solution to this general strategy that would provide improved interpretation is to identify benchmarks for a specific area of research (Kraft, 2020). Therefore, we follow this recommendation.

Method for Study 3

Participants

We invited three professors and all their students in three first-year introductory classes to participate. All students were enrolled in similar Developmental Psychology courses, differing in the degree program and the location of the university: a) Psychology degree, University of Zaragoza; b) Speech Therapy degree, University of Valencia; c) Teacher training degree, University of Valencia. In all, the sample consisted of 125 participants (*M* = 19.2 years, *SD* = 4.16, 93.4% women). We excluded participants who reported a diagnosis of previous learning difficulties (*n* = 4) or had missing data (*n* = 8), resulting in a final sample of 113 students.

Materials

Deep Cloze test. See method section of Study 1.

Kaufman Brief Intelligence Test (K-BIT). We used K-BIT (Kaufman & Kaufman, 1990/1996), a standardized screening tool for verbal and non-verbal intelligence. We used the non-verbal intelligence Matrices subtest. Each item corresponds to a matrix of images, where one is missing. Participants must select from a series of options the one that fits the matrix. Participants were presented with items 15 to 48 because items 1-14 are meant to be completed only by young students, according to the test manual.

Final exams. For each of the three classes assessed, the corresponding professor developed a final exam to evaluate students' subject knowledge. The exam for the Psychology degree class contained 40 multiple-choice questions with 3 options; the exam for Speech Therapy had two parts, the first with 25 multiple-choice questions with 3 options, and the second with 2 open-answer questions (out of 3 options); finally, the exam for the Teacher training degree also had two parts, the first composed of 30 multiple-choice questions with 3 options, and the second with 2 open-answer questions. Scores ranged from 0 to 10 (pass = 5 or above), the standard grading system in Spanish universities.

Procedure

Participants completed the demographic information, K-BIT, and Deep Cloze test via an online tool at a time and place of their choice. All participants provided informed consent to use their responses and their final exam grade, and they were debriefed after completing the study. Data collection took place during the first three months of the autumn semester of 2020/2021. Each class's final exam took place physically in regular classrooms during the exam period (January, 2021).

Results for Study 3

We first inspected data distributions for Study 3. Outliers for each variable were identified (between 1.1-5.3% of data) and replaced by the next highest or lowest score that was not an outlier. Scores for the K-BIT subtest were within the normal range (percentile 68). Once outliers had been replaced, all the measured variables were normally distributed (see Table 2, lower row).

Next, to test the predictive validity of the Deep Cloze measure, we performed a two stage hierarchical regression model (see Table 5). In a first step, we entered two dummy variables to control the effects of class (Dummy 1: psychology degree = 0, teacher training degree = 1; speech therapy degree = 0; Dummy 2: psychology degree = 0, teacher training degree = 0; speech therapy degree = 1), as well as the K-BIT scores. We included the scores on the Deep Cloze in a second step, to assess its effect on participants' exam grade. To facilitate the interpretation of the results, we standardized all measures prior to the analysis. In addition, before running the hierarchical regression analysis, the association between each predictor variable was examined. We found that K-BIT scores, Dummy 1 and Dummy 2 were not significantly correlated with Deep Cloze (*r* = .02, r_{pb} = .03 and r_{pb} = -.18, respectively), as did between K-BIT test and Dummy 1 and Dummy 2 (r_{pb} = -.01 and r_{pb} = -.10, respectively). The hierarchical multiple regression revealed that, in the first step, only the Dummy 1 variable was a significant predictor of the Model 1. When Deep Cloze was added in the second step, this variable also contributed significantly to the regression model and accounted for 4.7% of the variation in participants' exam grade. All VIF values for continuous predictors were close to 1.

Discussion for Study 3

Results indicate that Deep Cloze scores obtained at the beginning of the semester positively predicted students' grades in introductory classes, assessed as their final exam grades four months later, even after controlling for the effects of non-verbal intelligence. The fact that this relationship was obtained in three classes in different disciplines provides evidence of the generalizability of the pattern. Although the relation between the Deep Cloze scores and final grades was small, it should be noted that it is within the range of correlations obtained in previous studies with undergraduate students (Jackson, 2005; Williams et al., 2007).

Table 5. Summary of hierarchical regression analysis for variables predicting exam grade (Study 3, n = 113).

	Model 1			**Model 2**		
Variable	**B**	**SE (B)**	**β**	**B**	**SE (B)**	**β**
Intercept	-0.27	0.17		-0.32	0.17	
Dummy 1 (psychology vs. teacher training)	0.56*	0.22		0.61**	0.22	
Dummy 2 (psychology vs. speech therapy)	0.13	0.25		0.26	0.25	
K-BIT (non-verbal)	-0.05	0.11	-.05	-0.05	0.10	-.05
Deep Cloze				0.23*	0.10	.22
R^2	0.07			0.12		
F for change in R^2	2.66			5.71*		

Note: *Values of the INFLESZ scale between 55-65 indicate a normal legibility, values between 40-55 indicate a somewhat difficult legibility

General discussion

We presented an adaptation and evidence of criterion validity (both concurrent and predictive) of the Deep Cloze test, designed to assess reading comprehension efficiency of undergraduate students in Spanish. The test is unique because it aims to assess a population, university students, for which there is no standardized test in Spanish. Results from three studies provide initial evidence of the validity and reliability of the Deep Cloze test for assessing undergraduates' ability to comprehend texts. Nevertheless, given that there is no gold standard test to analyze text comprehension for adult Spanish native speakers, those conclusions should be taken with caution.

Usefulness of the Deep Cloze test

For decades there has been a widespread belief that cloze tests only measure comprehension at the sentence level (Shanahan et al., 1982). Contrary to this belief, results from the Deep Cloze test in Spanish indicate that it can capture students' comprehension of the situation being described in the text (i.e. situation model), even after controlling for their background knowledge about the topic. Specifically, this was shown by the small to medium positive correlations between Deep Cloze test performance and reading comprehension scores in Study 2b. The test was also useful in predicting students' performance on a final exam, even after controlling for non-verbal intelligence (Study 3). This is a noteworthy result considering that participants completed the Deep Cloze test four months before the final exam took place.

The Deep Cloze test also has many advantages in terms of simplicity and administration time. An important advantage is that it only takes eight minutes to complete, whereas other traditional reading comprehension tests require at least 30 minutes. Second, it does not depend on students' writing skills because they answer by selecting one of the available options. Finally, it can be administered collectively and corrected automatically, saving time and personnel resources.

Uses of the Deep Cloze test

The Deep Cloze test was designed to be used as an instrument for research with Spanish undergraduate students. Studies from different areas of Psychology, Education, Behavioral Economics, or Cognitive Neuroscience that focus on linguistic input, such as problem-solving studies, may benefit from having a quick test to control participants' reading comprehension, in order to avoid further experimental noise in the data. The Deep Cloze test may be particularly useful for instructional studies, given that the effectiveness of some interventions may depend on students' prior reading comprehension levels, as in aptitude x treatment interactions (Preacher & Sterba, 2019).

Limitations and future research

Our study does not come without limitations. First, conclusions from our study are limited by the samples we used. Specifically, because we sampled undergraduate students without any diagnosis of reading problems, it is still an open issue the extent to which the Deep Cloze test could also be used as a screening tool to identify students who are struggling with reading comprehension. Future studies could use different methods to identify struggling readers, such as teachers' referrals, think aloud protocols during reading, or self-reported learning strategies. In addition, they should assess other relevant factors that may help to explain the difficulties of struggling readers, such as decoding ability or oral comprehension (Talwar et al., 2021). Having a valid screening tool would be really useful in applied contexts, as universities could provide further support for freshman students identified as struggling readers. In addition, our sample came from just two universities and included mostly female students, a limitation that should be solved in future studies. Second, results from our studies should be interpreted in light of some limitations of the tasks and procedures used. Given that our studies didn't include low-level linguistic measures such as word decoding and vocabulary knowledge, we can't rule out the possibility that those factors influence undergraduate's reading comprehension as measured by the Deep Cloze test. In addition, because the studies were conducted on-line, some participants could have not followed properly the instructions. Future research should address these issues in controlled laboratory studies. Another limitation of our study was the use of the PROLEC-SE-R test which was developed to test 13–18-year-old students to assess the concurrent validity of the Deep Cloze test. As we already discussed, there are currently no standardized tests to assess reading comprehension in adult native speakers of Spanish to compare the Deep Cloze test with.

Conclusion

Reading comprehension is a critical skill during university and young adult learning. Its good psychometric properties, along with the ease and adaptability of its administration, make the Deep Cloze test a useful instrument to evaluate reading comprehension efficiency in Spanish undergraduate students, opening up new research pathways in this field.

Acknowledgments

We would like to thank Laura Gil, Pablo Delgado, Laura Royo for their assistance during the project, and to Carsten Elbro for sharing with us the original Deep Cloze test and for his continuous support. Participation of L. Altamura was funded by Grant PRE2021-099420 from Spanish Ministry of Science & Innovation.

Conflict of interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Disclaimer

Formatting, following the templates provided by Psicológica, spelling, grammar-checking and correct referencing are the sole responsibility of the authors.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.

Ardila, A., & Cuetos, F. (2016). Applicability of dual-route reading models to Spanish. *Psicothema, 28(1),* 71–75. https://doi.org/10.7334/psicothema2015.103

Carlisle, J. F., & Rice, M.S. (2004). Assessment of reading comprehension. In C.A. Stone, E.R. Silliman, B.J. Ehren & K. Apel (Eds.), *Handbook of language and literacy* (pp. 521-540). The Guilford Press.

Catalá, M., Catalá, G., Monclús, R., & Molina, E. (2001). *Pruebas ACL para la evaluación de la comprensión lectora [ACL tests for reading comprehension assessment].* Graó.

Clinton-Lisell, V., Taylor, T., Carlson, S. E., Davison, M. L., & Seipel, B. (2022). Performance on reading comprehension assessments and college achievement: A meta-analysis. *Journal of College Reading and Learning*. https://doi.org/10.1080/10790195.2022.2062626

Cohen, J. (1969). Statistical power analysis for the behavioral sciences (1st ed.). New York, NY: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cuetos, F., Arribas, D., & Ramos, J. L. (2016). *Batería para la evaluación de los procesos lectores en secundaria y bachillerato-revisada (PROLEC-SE-R*) *[Assessment of reading processes for secondary and senior high-school students-Revised].* TEA Ediciones.

Cuetos, F., Rodríguez, B., Ruano, E., & Arribas, D. (2007). *Batería para la evaluación de los procesos lectores revisado (PROLEC-R) [Assessment of reading processes-Revised].* TEA Ediciones.

Delgado, P., & Salmerón, L. (2021). The inattentive on-screen reading: Reading medium affects attention and reading comprehension under time pressure. *Learning & Instruction, 71,* 101396. https://doi.org/10.1016/j.learninstruc.2020.101396

Elbro, C., & Jensen, K. L. (2017, July 31). *How deep is your cloze? The construct validity of a Deep Cloze test.* [Poster presentation]. 27th Annual Meeting of the Society for Text and Discourse, Philadelphia, PA, United States of America.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock "N" Roll* (4th ed). Sage.

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*, 16–28. https://doi.org/10.1177/0734282912451971

Hadley, G., & Naaykens, J. (1999). Testing the test: Comparing SEMAC and exact word scoring on the selective deletion cloze. *The Korea TESOL journal, 2*(1), 63. https://koreatesol.org/sites/default/files/pdf_publications/KTJ2-1999web.pdf

Jensen, K. L., & Elbro, C. (2022). Clozing in on reading comprehension: a deep cloze test of global inference making. *Reading and Writing, 35*, 1221–1237. https://doi.org/10.1007/s11145-021-10230-w

Kaufman, A.S., & Kaufman, N.L. (1996). *Kaufman Brief Intelligence Test* (K-BIT) (Cordero, A., & Calonge, I, Trans.). Pearson. (Original work published 1990).

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology, 74*(6), 828–834. https://doi.org/10.1037/0022-0663.74.6.828

Kleijn, S., Pander Maat, H., & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing, 36*(4), 553–572. https://doi.org/10.1177/0265532219840382

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*, 241-253. https://scholar.harvard.edu/files/mkraft/files/kraft_2019_effect_sizes.pdf

León, J. A., Escudero, I., & Olmos, R. (2012). *Evaluación de la comprensión lectora (ECOMPLEC) [Reading comprehension assessment]*. TEA Ediciones. https://doi.org/10.5093/ed2013a9

Marchante, B. M. (2021). Relationship between grammatical and pragmatic competence in EFL of Spanish learners through a computer adaptive test. *Research in Education and Learning Innovation Archives*, 18-34. https://doi.org/10.7203/realia.26.17694

Martínez, T., Vidal-Abarca, E., Gil, L., & Gilabert, R. (2009). On-line assessment of comprehension processes. *The Spanish Journal of Psychology, 12*, 308-319. https://doi.org/10.1017/S1138741600001700

McBain, R. A., & Mhunpiew, N. (2014). The development of a vocabulary instruction model for content and language integrated learning for English language learners in Bangkok. *Latin American Journal of Content & Language Integrated Learning*, *7*(1), 82-97. https://doi.org/10.5294/laclil.2014.7.1.5

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation, 51*, 297-384. https://doi.org/10.1016/S0079-7421(09)51009-2

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 109–120. https://doi.org/10.1037/aca0000089

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adult skills.* OECD Publishing. http://dx.doi.org/10.1787/9789264128859-en

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*, 228-242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children, 85*, 248–264. https://doi.org/10.1177/0014402918802803

Quezada, C., & Westmacott, A. (2019). Reflections of L1 reading comprehension skills in university academic grades for an undergraduate translation programme. *The Interpreter and Translator Trainer*, *13*, 426-441. https://doi.org/10.1080/1750399X.2019.1603135

R Core Team. (2020). *R: A language and environment for statistical computing* (Version R-3.6.3) [Computer software]. R Foundation. https://www.R-project.org/

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.2.3. https://CRAN.R-project.org/package=psych

Shanahan, T., & Kamil, M. (1983, November 29 – December 3). The relationship of concurrent and construct validities of cloze [Paper presentation]. 33rd Annual Meeting of the National Reading Conference, Austin, TX, United States of America. https://files.eric.ed.gov/fulltext/ED239219.pdf

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229–255. https://doi.org/10.2307/747485

Talwar, A., Greenberg, D., Tighe, E. L., & Li, H. (2021). Examining the reading-related competencies of struggling adult readers: nuances across reading comprehension assessments and performance levels. *Reading and Writing, 34,* 1569-1592. https://doi.org/10.1007/s11145-021-10128-7

Toro, J., & Cervera, M. (1980). *Test de análisis de lectoescritura [Reading and spelling analysis test].* TEA Ediciones.

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología, 33*(3), 755–782. http://dx.doi.org/10.6018/analesps.33.3.268401

Wilson Van Voorhis, C. R., & Morgan, B. L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology, 3*, 43-50. https://doi.org/10.20982/tqmp.03.2.p043

Annex I

Test Deep Cloze de comprensión lectora

Este test se compone de 34 historias cortas, que son independientes las unas de las otras. A cada texto le falta una palabra para que la historia tenga sentido. En cada historia se proponen cuatro palabras –entre corchetes- para completar la historia. De entre las cuatro opciones, debes subrayar la palabra que mejor complete la historia.

Por ejemplo, en la historia “Ellos salieron volando con la esperanza de ser los primeros en llegar al objetivo, pero otra persona [cojeó, calculó, corrió, comió] más rápido”, deberías subrayar la opción correcta que es “corrió”.

Se presentan 34 historias. Sin embargo, solo dispones de 8 minutos para completar el mayor número de historias que puedas. Si lo ves necesario, puedes saltarte una pregunta, y continuar con el resto. Cuando se te indique, pasa la página para empezar con el test.

1. Ellos acordaron reunirse justo después del trabajo. La mujer del traje sacó una bandeja tras otra, mientras ellos se cogían de la mano y miraban intensamente. Les parecía difícil tomar una decisión. No estaban seguros de que fuera [anual, **oro**, grande, a cuadros].

2. La mesa ya estaba puesta. Olía ligeramente a quemado, aunque el temporizador no había sonado todavía. A continuación, sacó [el edredón, los papeles, la radio, **la bandeja**].

3. Intentó subir [la escalera, **la cremallera**, los precios, el volumen], pero fue difícil porque no se estaba quieto. No iban a salir a tiempo.

4. Había cinco delante de él y todos/todas sus [**cestas**, bolsos, manos, coches] estaban llenos/llenas. Con todos esos artículos iba a tardar mucho tiempo.

5. Estaban sentados en la terraza en pantalones cortos con el torso desnudo. Habían pedido un zumo de frutas con hielo. Fue necesario montar [la carpa, el escenario, **la sombrilla**, los altavoces].

6. Se podía escuchar el sonido de [los perros, **las ambulancias**, los manifestantes, los truenos]. Ya se habían aglomerado muchas personas alrededor de los dos coches. Algunas de ellas estaban llorando.

7. Tenía que estar lista en dos horas, así que tenía un poco de prisa. La bolsa estaba ya en el coche y el billete, las llaves y la cartera estaban en su bolsillo. Su marido corrió tras ella con su [almuerzo, **pasaporte**, lista de la compra, llave USB]. Tuvo suerte, de lo contrario no habría llegado muy lejos.

8. Eran tantos que no había suficiente espacio para que todos pudieran estar de pie resguardándose de la lluvia. Cuando lo vieron en la distancia, todos los [visitantes, oficiales, estudiantes, **pasajeros**] dieron un paso adelante.

9. Sara buscó a tientas su [bolígrafo, **entrada**, cartera, calendario] cuando entró en la sala. No le gustaba el olor a palomitas de maíz. Después le molestó el hombre que estaba sentado justo detrás de ella y leía los subtítulos a su hijo en voz alta.

10. Se sentaron en silencio muy juntos en el sofá y lamentaron lo que habían empezado. Indicaba que debías ser mayor de 15 años. Pasado un tiempo, el más joven se decidió a coger [su brazo, las gafas, el mango, **el mando**].

11. Aunque lo había intentado con diferentes términos, no había logrado encontrar lo que buscaba. Ahora tenía siete [habitaciones, armarios, instituciones, **páginas**] abiertos/abiertas al mismo tiempo. Todo se había vuelto muy confuso.

12. Ella estaba de pie en la acera. Sin pensarlo, insertó el pin. Unos segundos más tarde tenía [la conexión, **el dinero**, el correo, los nombres] en las manos.

13. Sacó los utensilios del armario y los preparó. Entonces se puso a trabajar. Al terminar, volvió a colocar todo en su lugar y dejó pasar un rato para que se secara. Todo estaba [completo, vendido, descuidado, **brillante**].

14. Los padres hicieron [la presentación, **la tarta**, la recepción, la reunión]. La mesa estaba puesta y bien presentada. Cuando los invitados estaban a punto de llegar, su hijo corrió hacia la puerta de entrada con su nuevo tractor en la mano.

15. Se puso de pie junto a la larga mesa y se aclaró la garganta. En su bolsillo tenía [el móvil, **el discurso**, el contrato, los códigos de seguridad]. Después, se pudo escuchar el tintineo de los vasos.

16. No podían colaborar y se frustraron. Algunos fueron sustituidos y los asistentes empezaron a abuchear. Al final, no hubo [veredicto, dinero, clientes, **goles**].

17. Cuando la pareja regresó después de pasar el fin de semana fuera, todo el sótano estaba destrozado. Rápidamente se pusieron en contacto con [la policía, la compañía de mudanzas, los testigos, **la compañía de seguros**]. Escucharon en las noticias que muchas áreas habían sido afectadas.

18. Los policías sostuvieron [al detenido, la cómoda, **el ataúd**, la mesa]. Pasaron por delante de varias coronas de flores mientras caminaban lentamente por el pasillo. Después, fueron invitados a tomar café en la casa de un pariente cercano.

19. [El automóvil, **La fecha límite**, El agobio, El mal tiempo] había pasado sin que se hubiera dado cuenta. No era típico en él. Debía haber colocado el sobre en un lugar incorrecto, tal vez en el montón de papeles para reciclar. Ahora le costaría aún más.

20. La puerta se abrió y la niña entró. Nadie aparecía. Fue directamente a por una botella y luego esperó de pie. Pasó medio minuto antes de que pudiera [chatear, caminar, comer, **pagar**].

21. Decidió quedarse a descansar fuera con unos compañeros de clase, mientras el resto seguía esquiando. Se asustó cuando fue golpeado. Su espalda se empapó rápidamente. Quería vengarse y se apresuró en recoger [pelotas, barro, arena, **nieve**].

22. Había cartones y recortes de folletos y periódicos esparcidos por todas partes. Llevaría todo el descanso limpiar la habitación, pensó el profesor. Fue adecuado que todos usaran [harina, ordenadores, **delantales**, cepillos de dientes], de lo contrario, algunos padres se habrían quejado.

23. Ella sostuvo firmemente sus manos húmedas sobre el volante tal y como le habían enseñado. De repente, sintió con toda claridad cómo pisaba el freno en vez del embrague. Ahora estaba segura de que [ganaría, pasaría, gritaría, **suspendería**].

24. Los coches se detuvieron y la multitud comenzó a moverse. Un hombre corrió para coger su autobús. La mayoría de la gente se apresuró a cruzar mientras [el payaso, el camión, **el hombre**, el meteorito] estaba en verde.

25. La cuerda colgaba sobre el borde de la taza. Ahora ella solo tenía que [**esperar**, sorber, respirar, apretar] durante unos minutos. Sacó un par de galletas.

26. Habían pasado semanas planificando y habían elegido cuidadosamente el lugar. Su compañero esperaba en el coche a la vuelta de la esquina. Su ritmo cardíaco aumentó mientras se introducía por el hueco que habían hecho en la verja. Rezaron para que no hubiera [quejas, turistas, supervivientes, **perros**].

27. Había mucho que hacer. Ella comenzó con [las cajas, **los vasos**, las ventanas, los correos electrónicos]. Cuando terminó, sus manos estaban completamente arrugadas y olían a detergente. Tenían que convertirlo en una rutina diaria.

28. Todos los días, cuando llegaba a casa, él se acurrucaba alrededor de sus piernas y dejaba pequeños pelos en sus pantalones. Estaba muy contento porque sabía que [una sorpresa, una charla, un balanceo, **una merienda**] le estaba esperando.

29. María pensaba que siempre solían tener buenas conversaciones, pero esta vez los temas resultaron no ser tan interesantes, y el [gerente, profesor, **invitado**, cuidador] tenía una voz molesta. María acabó apagándolo.

30. Un grupo de niños vino corriendo hacia ellos. La anciana trató de agarrarse fuerte. Pudo sentir cómo [el aire, el abrigo, la mano, **la correa**] se extendió por completo. Ella siempre se sentía tensa cuando se encontraban con los niños durante su paseo diario.

31. Primero hizo el lado izquierdo. Giró la silla ligeramente para estar en mejor posición. Luego hizo el lado derecho. Después de una breve conversación con el cliente, ella cogió [el pincel, la imagen, **las tijeras**, el documento] nuevamente para terminar los últimos retoques.

32. Ella estaba prácticamente tumbada y la luz de la lámpara situada justo encima le molestaba mucho en los ojos. Ella estaba un poco nerviosa. Él apartó el pequeño espejo. Entonces ella giró la cabeza ligeramente y vio [**el taladro**, la bicicleta, la carta, el martillo] que tenía en la mano.

33. Los dedos del alumno se movieron rápida y adecuadamente. Él se sentó y golpeaba el suelo con su pie. Ella observaba y se aseguraba de que él siguiera [la receta, las líneas, las reglas, **las notas**].

34. Iba a acostarse. Apretó el tubo, se miró en el espejo y comenzó a [soñar, calcular, **cepillar**, escribir].