2022 ■ Vol. 18 ■ no. 3

Evaluating different scoring methods for the speeded Cloze-elide test: The application of the Rasch Partial Credit Model

Farshad Effatpanah [a] and Purya Baghaei [b]

[a]English Department, Tabaran Institute of Higher Education, Mashhad, Iran
[b]English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran

Abstract ■ Cloze-elide tests are overall measures of both first (L1) and second language (L2) reading comprehension and communicative skills. Research has shown that a time constraint is an effective method to understand individual differences and increase the reliability and validity of tests. The purpose of this study is to investigate the psychometric quality of a speeded cloze-elide test using a ploytomous Rasch model, called partial credit model (PCM), by inspecting the fit of four different scoring techniques. To this end, responses of 150 English as a foreign language (EFL) students to a speeded cloze-elide test was analyzed. The comparison of different scoring techniques revealed that scoring based on wrong scores can better explain variability in the data. The results of PCM indicated that the assumptions of unidimensionality holds for the speeded cloze-elide test. However, the results of partial credit analysis of data structure revealed that a number of categories do not increase with category values. Finally, suggestions for further research, to better take advantage of the flexibilities of item response theory and Rasch models for explaining count data, will be presented.

Keywords ■ Reduced redundancy, speeded cloze-elide test, scoring methods, item response theory, Rash partial credit model.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

farshadefp@gmail.com

10.20982/tqmp.18.3.p241

Introduction

The concept of reduced redundancy (RR; Spolsky, 1969; Spolsky, Sigurd, Sato, Walker, & Arterburn, 1968) has long been considered as a theoretically-sound procedure for constructing language tests in the field of language testing to allow researchers to measure language ability of test takers in both first (L1) and second/foreign language (L2). The rationale behind the concept of RR is that all natural languages inherently include redundancy so that speakers of a particular language are able to restore missing linguistic items. In fact, understanding a language entails the competence to perceive a mutilated text or distorted message and provide some valid guesses about the removed elements (Klein-Braley, 1997; Spolsky, 1968). As Spolsky (1969) argued, redundancy is a property of the human verbal communication system which reduces the viability of errors, and permits communication in which there is some interference in the communication channel (e.g., noise).

In order for researchers or test developers to construct tests based on the principles of RR, noise should be added into utterances and written texts, as authentic materials, or some portions of a test should be masked to test a subject's ability to reformulate the omitted elements. The way in which test takers perform to restore mutilated texts is considered as a valuable method to provide evidence for language proficiency levels of test takers. Research has shown that tests developed based on RR are reasonably good ways of overall language ability and tap both productive and receptive processes in an integrative manner (Klein-Braley, 1997). Examples of tests that have been devised based on RR include the standard dictation (Oller, 1971), the (classical) cloze test (Oller, 1976, 1979; Oller & Conrad, 1971), the partial dictation test (Johansson, 1973, 1974), the multiple-choice cloze test (Jonz, 1976), rational deletion cloze tests (Bachman, 1981, 1985), the noise test (Gaies, 1987; Gaies, Gradman, & Spolsky, 1977; Gradman & Spolsky, 1975; Spolsky, 1971), C-tests (Klein-Braley & Raatz, 1984; Raatz & Klein-Braley, 1981), and cloze-elide tests (Manning, 1987). Although there are differences between different facets of these tests, they all incorporate the supposition of RR to test language abilities of test takers in repairing the mutilated texts.

The most notable operationalization of RR is cloze testing procedure as a reliable and efficient measure of language abilities in both L1 and L2 (Oller, 1973, 1979; Oller & Conrad, 1971). Deriving from the concept of closure in the Gestalt school of psychology (Stansfield & Hansen, 1983), in which individuals are inclined to reformulate the missing parts by using their prior experiences or background knowledge, cloze procedure was originally developed in the early 1950s as a psychological tool for assessing readability of written materials (Taylor, 1953). In cloze tests, after a short intact lead-in for introducing context of the text, every nth word is completely deleted from a piece of text and replaced by blanks, and test takers must fill out the missing words (or group of words). Unlike 'discrete-point' methods (e.g., multiple-choice, true/false, and fill in the blank) in which language is divided into different discrete components or "points" and these components are assessed at a time, the cloze procedure is viewed as an 'integrative' method of assessment developed from the unitary trait hypothesis (Oller, 1979), which states that language is indivisible. More specifically, cloze test intends to assess the total communicative abilities and language use of test takers, and they should read and comprehend a considerable amount of discourse (Carroll, 1961; Farhady, 1979; Oller, 1979). Research has shown that cloze test can be used as a measure of overall English proficiency of native speakers (Anderson, 1976; Bormuth, 1967; Oller, 1973, 1979; O'Reilly & Streeter, 1977; Ruddell, 1964) and non-native speakers (Anderson, 1976; Friedman, 1964; Oller, 1979).

Some researchers, on the other hand, have maintained that irrespective of the strengths of cloze test, it has drawn criticism, and a number of studies have strived to modify the format of the measure to tackle the proposed deficiencies (Alderson, 1980, 1983; Jonz, 1976; Klein-Braley, 1981, 1997; Ozete, 1977; Porter, 1976). For instance, Klein-Braley (1997) pinpointed that the deletion rates are too high, and nth word deletion is dissimilar to random deletion. As a test of reading comprehension, Porter (1976) argued that cloze tests focus on productive language aspects rather than comprehension, that is, in addition to the comprehension of the cloze test, test takers should write their responses in blanks which divert their attention away from the reading task (Ozete, 1977). For this reason, Ozete (1977) suggested a multiple-choice version as an effective solution to overcome the perceived restrictions of cloze test. Jonz (1976) proposed a multiple-choice cloze test (M-C cloze) as a way of enhancing the cloze test by limiting range of choices for blanks and reducing the number of items in the test. In the M-C cloze, test takers are provided with three to five choices for each blank and they should choose only one option. This practice is easier to score, increases reliability and placement accuracy, and improves objectivity and ease of administration (Chapelle & Abraham, 1990; Jonz, 1976).

Along the same lines, another important shortcoming of cloze tests is that they tend to have low reliability and validity coefficients for homogeneous samples, and any particular change in deletion frequency can have a great impact on the relationship of the cloze test to measures of language proficiency (Alderson, 1979, 1980). C-tests were introduced to improve the limitations of the cloze tests (Klein-Braley, 1997; Raatz & Klein-Braley, 1981). Compared to cloze tests which include a single long passage and a word is completely removed in every nth word, in C-tests every second half of every word is omitted, but the first and last sentences remain intact. C-tests have already been developed and validated in various languages (Norris, 2018), and are viewed as valid instruments for measuring overall language ability in both L1 and L2 (Raatz & Klein-Braley, 1981). A large number of researchers have used various quantitative and qualitative research methods to show the reliability and validity of C-tests (Babaii & Ansary, 2001; Baghaei & Grotjahn, 2014; Eckes, 2010; Eckes & Grotjahn, 2006a, 2006b; Forthmann, Grotjahn, Doebler, & Baghaei, 2020; Raatz, 1985).

Another modified format of the cloze test, which is the main focus of this study, is cloze-elide (Manning, 1987), also known as the *intrusive word technique* (Davies, 1975, 1989). In this type, a certain number of words are randomly interspersed throughout the text and test takers are required to read the text, identify, and cross out or elide extraneous words. The cloze-elide is also labelled "text retrieval", "text interruption", "doctored text", "mutilated text", and "negative cloze" (Alderson, 2000, p. 225). Manning (1987, pp. 9-10) maintains that the cognitive processes underlying the cloze-elide interact in the following ways:

> . . . the readers collect information as they read along, developing evidence about what the text might mean. These data come from basic elements in the text ("bottom-up" or "data-driven" processes) and from hypotheses in the reader's mind ("top-down" or "hypotheses-driven" processes). These hypotheses are tested against the accumulating information

> base and occur at several levels (lexical, syntactic, semantic) more or less concurrently. They also more or less compete, as alternative hypotheses, for acceptance. As a particular hypothesis gains power and acceptance, it tends to facilitate or inhibit other hypotheses by more or less channeling the simultaneous spread of activation and the focusing of attentional processes. Thus comprehension comes to depend upon the reader's ability to use her or his own knowledge and the textual information both within and between levels of analysis.

Manning (1987) also highlights that the cloze-elide test is not solely a measure of reading comprehension ability, but it can be a measure of communications skills, such as listening, speaking, and writing, because to resolve linguistic problems given in a cloze-elide test, test takers should employ several underlying cognitive operations at deeper levels.

The processing of the text in cloze-elide is somewhat contrary to the standard cloze test (Alderson, 2000; Farhady, 1996). The cloze test requires test takers to read the text and insert some words while, in the cloze-elide, test takers should read the text and remove redundant words. Because cloze-elide tests differ from the other types of cloze test, some researchers (Alderson, 2000; Baker, 2011; Bowen, 1978; Farhady, 1996; Hudson, 2007; Lee, 2008; Manning, 1987) have suggested certain precautions in constructing and using cloze-elide tests including: (1) selecting a suitable passage with acceptable length and difficulty; (2) specifying the exact locations where the redundant words should be inserted based on an appropriate method, such as "pseudo-random", "rational", and "random insertion" procedures; and (3) selecting the words to be inserted in the text, which can be chosen randomly or based on the similarity to the adjacent words.

Davies (1975) argued that the best use of cloze-elide may not be only as a measure of reading comprehension, but it can be used as a measure of processing speed or "speeded reading test" because test takers should cross out superfluous words from a text within a limited period of time. The number of redundant words correctly identified minus the number of items incorrectly identified can be taken as a measure of reading speed (Alderson, 2005). Cloze-elide, in this case, is turned into an error recognition task (Manning, 1987). In language proficiency testing, Alderson (2005) maintains that speeded tests are informative about test takers' implicit knowledge of the target language and that non-speeded tests provide information about test takers' explicit knowledge. Therefore, processing speed under time pressure is a central component of language proficiency (Alderson, 2005). Bachman (1990) asserts that test scores obtained from speeded tests reveal different ability levels of test takers. Time constraints have been shown to be useful for identifying individual differences (Oller, 1973) and increasing test reliability and validity (Grotjahn, 2010). High proficient readers are usually fast readers and they can comprehend a text more than low proficient readers.

A large number of researchers have shown the value of cloze-elide as a measure of language proficiency and reading comprehension (Bowen, 1978; Manning, 1987). In fact, the ability of test takers to recognize and elide the redundant words in a text are regarded as their level of language proficiency; more proficient test takers are expected to be more successful in detecting inserted words in the text. Elder and von Randow (2008) also showed that there is a significant moderate-to-high correlation between cloze-elide scores and a wide variety of language proficiency tests, including listening, reading, speaking, and writing. More importantly, Elder and von Randow (2008) indicated that cloze-elide scores are predictive of diagnostic abilities of examinees. The construct validity and reliability of the cloze-elide have been examined with various quantitative data analysis methods, including regression models (Manning, 1987) and exploratory factor analysis (Baker, 2011; Klein-Braley, 1997; Manning, 1987; Zare & Boori, 2018). Despite the importance of cloze-elide in language proficiency tests, there is limited research on investigating the validity of the test, and language researchers are not well aware of its psychometric quality.

The Current Study

The purpose of the present study is to examine the fit of the cloze-elide test to partial credit model (PCM; Masters, 1982). In this study, time constraints are imposed to analyze the performance of examinees under time pressure. This study builds on and extends previous research on the quantitative data methods used for exploring psychometric quality of cloze-elide tests. To date, too little attention has been paid to the application of item response theory (IRT) models in analyzing the psychometric properties of cloze-elide tests. This can be due to incompatibility of data structure in processing speed tests with the requirements of IRT models (Doebler & Holling, 2016).

IRT models are a family of mathematical models which explain the relationship between the performance of an examinee on an item and location of the examinee on the latent trait continuum (Effatpanah, 2019; Effatpanah & Baghaei, 2021; Hambleton & Swaminathan, 1985). The probability of getting an item right or endorsing a response category is assumed to be a function of both an examinee's ability and a set of item characteristics. Test takers with greater ability parameters have higher probabilities to give a correct response to an item. The relationship between individuals' ability or θ and the probability of correct response is graphically illustrated in a set of graphs called item characteristic curves (ICCs). The use of IRT models allows researchers and practitioners to optimally design tests, analyze items in more detail, provide standard errors of measurement for various ability levels and item difficulties, develop computerized adaptive testing, analyze differential item functioning (DIF), and devise test equating (Doebler & Holling, 2016).

Partial Credit Model (PCM)

The partial credit model (PCM; Masters, 1982), also referred to as the adjacent category logit model, is a latent trait model for analyzing polytomous responses to a set of test items. It is considered as an extension of the Rasch model (RM) for dichotomous responses (Rasch, 1960). PCM is appropriate for modeling instruments in which polytomous items contain several ordered categories, such as achievement or aptitude test items and attitude questionnaires. The model assumes that (1) the number of response options vary across items, that is, some items can be on a 5-point scale and some on a 4-point scale, or some even can be dichotomous; (2) steps within items should be completed in sequence, although the steps do not need to be equally difficult or ordered in terms of difficulty (Baghaei & Effatpanah, 2022; Desjardins & Bulut, 2018); and (3) the model assumes that all items discriminate equally among examinees. Furthermore, the PCM emphasizes adjacent categories when estimating the thresholds (i.e., difficulty) between the ordered response categories (Desjardins & Bulut, 2018; Masters, 1982). Because an item has K ordered option responses, PCM estimates $K-1$ thresholds for the item. The model estimates a unique set of thresholds for each item. As noted by Desjardins and Bulut (2018), "PCM does not require the thresholds to follow the same order as the response categories. Because PCM considers adjacent categories in each step, the adjacent response categories are treated as a series of dichotomous items, but without order constraints beyond adjacent categories" (p. 145). When an item only contains two categories, then the PCM reduces to the RM. Although the PCM is a generalization of the RM, the model enjoys distinctive properties of the RM such as sufficiency of raw scores, independent item and person parameter estimates, and specific objectivity.

Under the standard RM, the probability of getting an item i correct by person ν with regard to his/her ability θ_ν and the item difficulty β_i is defined as:

$$P(X_{vi}=1 \mid \theta_\nu, \beta_i) = \frac{\exp(\theta_\nu - \beta_i)}{1+\exp(\theta_\nu - \beta_i)} \tag{1}$$

For the PCM, the examinee-item interaction is modeled as:

$$P(X_i \mid \theta_v, \beta_{ih}) = \frac{\exp\left(\sum_{j=0}^{x}(\theta_v - \beta_{ih})\right)}{\sum_{k=0}^{m_i}\exp\left(\sum_{j=0}^{k}(\theta_v - \beta_{ih})\right)} \tag{2}$$

where P_{xi} is the probability of obtaining X_i points ($X_i = 0, 1, \ldots, m_i$) on item i; θ_ν is the latent trait; β_{ih} is the step difficulty (also known as step parameter) associated with category score i_h of item i with m_i categories. A 'category score' is the number of successfully completed steps, and a 'step' refers to a stage required to complete an item.

Method

Participants

The data analyzed in this study consists of the performance of 150 Iranian English as a foreign language (EFL) students to a speeded cloze-elide test. This dataset was previously examined by Zare and Boori (2018) for examining the psychometric quality of the speeded cloze-elide test and its relationship with multiple-choice cloze test, C-test, and reading comprehension. The dataset is available on the website of the journal. There were 92 (61.3%) females and 58 (38.7%) males. The ages of these participants ranged from 19 to 39 (M = 23.46, SD = 3.51). The participants were 2nd (48%), 3rd (34%), and 4th (18%) year college students and recruited from the English department at the Islamic Azad University of Mashhad, Iran. The participants were non-native speakers of English. They provided their written consent form to voluntarily participate in the study on conditions of anonymity and confidentiality. The students were given no award, but a personalized test performance report was generated for each student. As the research involves human participants, the Ethics Committee of the Islamic Azad University of Mashhad reviewed and approved of the study (institutional review board decision no. d/8651).

Instrument

All participants were given a cloze-elide test and asked to complete the test within 2 minutes. The test contained five passages, each containing 20 redundant words. To select appropriate passages for the study, *Select Readings* (Second Edition), as a four-level American English reading course book series, by Lee and Gunderson (2011) was used. The passages were of various length (287-318 words) and based on different text genres and topics. As to the reading passages, the first one was an elementary passage of 307 words on different reasons why bikes are so popular in Denmark. The second passage was a pre-intermediate text of 296 words on the presence of colors in many English expressions. The third passage was a pre-intermediate text of 288 words and discussed the effects of using mobile phones. The fourth passage contained an intermediate text with 287 words in lengths that argue how babies develop language skills. And finally, the fifth passage was an upper-intermediate text with 318 words in length on differences between the life of a man who does no reading and that of the man who does.

To randomly insert a number of superfluous words into the passages to construct the cloze-elide test, several words were selected from Frequency Dictionary of Contemporary American English (Davis & Gardner, 2010). The dictionary provides a list of 5000 most commonly used words in English and depends on data from a 385 million word corpus, including spoken English, fiction, magazines, newspapers, and journals. The dictionary classifies all words into five parts: (a) from 1 to 1000, (b) from 1001 to 2000, (c) from 2001 to 3000, (d) from 3001 to 4000, and (e) from 4001 to 5000. For this study, 20 words were randomly selected from each part. To produce random numbers, the following site was used http://www.random.org. The randomly-selected words included both content (nouns, adjectives, adverbs, verbs, numbers, and interjections) and function words (articles, conjunctions, determiners, pronouns, prepositions, existential, negations, the infinitive marker (to), and genitive). In the next step, the selected words were interspersed at random locations in the text, and the first two sentences of each passage remained intact. The positions of the words were specified using software producing random numbers. If the suggested place was before or after the proper noun, the location of the inserted word was changed (e.g., moving one word further or back). Using http://www.random.org, the random numbers were determined, ranging from seven to fourteen for the location of words in the text. The passages were meticulously examined and revised for several times to assure that the embedded words do not conform to the sentence structure.

Finally, three university instructors, as experts in this field, were invited to review the constructed cloze-elide test and provide comments on content, appropriateness, wording, clarity of the test, and more importantly, the plausibility and authenticity of the inserted words. All of the instructors (two men and one woman) were non-native English speakers, and held PhDs in Applied Linguistics and Second Language Education with at least 10 years of experience in teaching reading comprehension. Based on the length of the passages, and the number of words can be read in a minute, testing time for each test was set at 5 minutes including instructions. This amount of time was halved to make the passage appropriate for being a speeded test. The finalized version of the cloze-elide test was administered in the current study and students were required to cross out the redundant words within 2 minutes.

Results

Examining the Fit of Different Soring Techniques

In scoring cloze testing procedures and C-tests, the common method is that each gap is considered as an item. This way of scoring, however, has two problems (Forthmann et al., 2020): (1) it substantially increases the number of item parameters for estimation; and (2) because the blanks within a passage are connected (directly or indirectly) to each other in terms of content and different linguistic features (Harsch & Hartig, 2010), the assumption of local, or condition, item independence as an important assumption in most IRT and Rasch models is violated. Local independence assumption implies that item responses of an examinee to test items are independent or uncorrelated based on the latent trait (Hambleton & Swaminathan, 1985). The violation of local independence may lead to overestimation of reliability coefficients, biased item difficulty and item discrimination parameter estimates, and overestimation of the accuracy of individuals' ability estimates (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Yen & Fitzpatrick, 2006). Furthermore, Heckman, Tiffin, and Snow (1967) argue that when these tests are administered under some pre-established time limits, all items may not be attempted by all test takers; therefore, in this case, addressing gaps as individual items is troublesome (Forthmann et al., 2020). Such issues are totally true for cloze-elide tests as a variant of cloze testing procedure.

To resolve these problems, researchers typically aggregate all the scores on the gaps for each passage and then, total scores are used for applying Rasch and IRT models. This modeling strategy is known as item bundle approach (Rosenbaum, 1988). As each passage is viewed as polytomous with various ordered categories, ordinal Rasch models such as rating scale model (RSM; Andrich, 1978), the partial credit model (PCM; Masters, 1982), and the continuous rating scale model (CRSM; Manning, 1987) can be employed.

In this study, the five passages were considered as polytomous items and four different scoring techniques were taken into account, including (1) the total number of redundant words correctly identified (Correct(C)), (2) the total number of redundant words incorrectly identified (Wrong (W)), (3) the total number of redundant words missed or unidentified (Miss(M)), and (4) the number of redundant words correctly identified minus the number of redundant words incorrectly identified (Correct-Wrong (C-W)).

Table 1 ■ Overall Model Fit across the Four Scoring Techniques

Scoring Techniques	N	$-2LL$	AIC	BIC	$CAIC$
Correct (C)	72	1904.31	2048.31	2265.07	2337.07
Wrong (W)	65	1093.66	1223.66	1419.35	1484.35
Miss(M)	90	2134.92	2314.92	2585.88	2675.88
Correct-Wrong (C-W)	71	1783.93	1925.93	2139.68	2210.68

Note. Note: N = number of parameters, AIC = Aikaike's Information Criterion, BIC = Bayesian Information Criterion, and $CAIC$ = Bozdogan's Consistent AIC.

The WINSTEPS computer package Version 3.73 (Linacre, 2009) was used to examine the fit of the speeded cloze-elide test to PCM. The overall model fit of the four scoring techniques was firstly evaluated to identify which scoring technique has a better fit to the model. To explore the fit of the four techniques or models, -2log-likelihood$(-2LL)$ and a set of information criteria were compared. The most widely used information criteria are:

$$\begin{aligned}\text{Akaike's Information Criterion } &= -2LL + 2P,\\ \text{Bayesian Information Criterion } &= -2LL + P\ln(N),\\ \text{Bozdogan's Consistent AIC } &= -2\ln L + p\,(\ln(N)+1)\end{aligned}$$

where LL is the log-likelihood of the estimated model; P is the number estimated parameters; N is sample size; and $\ln(N)$ is the natural log of sample size. As a rule of thumb, the model with the lowest log-likelihood and fit values, indicating parsimony, is more preferred (Janssen & De Boeck, 1999). As Table 1 shows, the scoring technique based on wrong (W) has a better fit compared to the other scoring techniques, suggesting that W scoring technique can better account for variability in the data. Therefore, the scores based on wrong (W) scoring techniques were used for further analyses.

Item Difficulty Parameters and Fit Statistics

Table 2 demonstrates descriptive indices of the data, including mean and standard deviation, item difficulty parameters of the five items (or texts), standard error of measurement, infit and outfit mean squares, and point-measure correlations. The item difficulty parameters present the position of items on the latent trait continuum and are explained in logits (or log odd-units). The error of measurement shows to what extent the item difficulty parameters were accurately estimated. As Linacre (2002) argued, infit mean square (INFIT MNSQ) indicates information-weighted fit or inlier-sensitive. This fit statistic is sensitive to unexpected response patterns of examinees to items that are relatively targeted on examinees, and vice versa; however, outfit mean square (OUTFIT MNSQ), as an outlier-sensitive fit statistic, is sensitive to unexpected response patterns of examinees to items that are fairly very easy or very difficult for them, and vice versa. The acceptable boundary for fit values are 0.60 and 1.40 for measuring rating scales (Bond & Fox, 2015; Linacre, 1999; Wright & Linacre, 1994). The point-measure correlations for all items were also estimated to measure the degree to which the observed scores are in agreement with the expected latent trait.

As can be seen in Table 2, Item 1 has the lowest mean score and Item 5 has the highest, suggesting that a large number of test takers answered correctly to this item. The results of item difficulty parameters also indicated that item 3 is the easiest and item 1 is the most difficult. Except for item 2 regarding INFIT MNSQ, the fit values are within the ideal range of 0.60 and 1.40. The point-measure correlations show that all correlations are positive and medium-to-high. Point-biserial (or point-measure) correlations indicate to what extent the responses to each item within a measure are correlated with the overall measure. Taken together, these values indicate that the patterns of item difficulties in the data agree with the RM (Linacre, 2009).

In addition, person and item reliability coefficients as well as separation values were evaluated. The reliability of item and person indices show the accuracy of the test in measuring item difficulty and person performance (Linacre, 2009). Separation reliability is defined as the ratio of person or item true standard deviation to error standard deviation (e.g., root mean square error (RMSE)), and indicates to what extent the person and item parameters are separated on the latent trait. In other words, person separation is used to classify examinees. Low values of person separation (< 2, person reliability < 0.8) indicate that the scale may not be sensitive enough to discriminate between low- and high-level examinees (Linacre, 2009). Item separation, on the other hand, is used to confirm the hierarchy of items. Low values of separation (< 3 = high, medium, low item difficulties, item reliability < 0.9) suggest that the sample is not large enough to verify the hierarchy of item difficulties of the scale (Linacre, 2009). The value of separation reliability varies from zero to infinity. A greater value of separation for persons/items denotes that there is a high probability that persons/items with high ability/difficulty estimates have higher ability/difficulty estimates compared with items/persons with low estimates (Linacre, 2009). For this study, item reliability coefficients and separation values were 0.72 and 1.61, respectively. The values of person separation and reliability were 1.46 and 0.68, respectively, indicating the restricted range of test takers' abilities. As Linacre (2009) put, this is not considered as a serious problem because it reflects the natural characteristics of the examinees taking the test.

Checking Unidimensionality

As agued by Smith and Plackner (2009), outfit and infit indices are not sensitive to systematic threats to unidiemnsionality. For that reason, unidimensionality of the speeded cloze-elide test was examined. Unidimensionality assumption is an important requirement in IRT and Rasch models. It states that all items of a scale should measure a single latent trait at a time. When the data fit the model (e.g., the assumptions of the model are satisfied), it is assumed that there is only one latent trait that explains variability in the data, that is, all items only measure a single construct. To investigate the unidimensionality of the speeded cloze-elide test, the principal component analysis of linearized Rasch residuals (PCAR) was investigated. Residuals refer to the differences between the Rasch model expectations and the observed data. The model fits the data better when the values of residuals are low. Because residuals are unexpected part of the data, which are not in accordance with the RM, they are expected to be uncorrelated and randomly distributed (Linacre, 2009). Using PCAR is an effective method to check that the residuals are independent or uncorrelated, and no extra component (secondary dimension) can be extracted from them. It must be noted that when PCAR is implemented based on standardized residuals, the latent trait is removed from the analysis; thus, any dimension drawn from the residuals is considered as an unexpected dimension (Linacre, 2009). In principal, when PCAR reveals noticeable factors from residuals, it indicates that an additional dimension is at work and the test is multidimensional. To specify whether the extracted factors are negligible or substantial, the magnitude of eigenvalue is analyzed. As suggested by Linacre (2009), eigenvalues below 2 indicate that the observed residuals are not substantial enough to consider a secondary dimension in the data.

For the present study, the results of PCAR showed that the model accounts for 70.3% of the observed variance; 63.2% are accounted for by person measures, and 7.1% are accounted for by item measures. The observed model variance is very close to the model expectation of 70.9%; however, 29.7% of the variance are still unexplained. The first factor (contrast) explains 9.4% of the unexplained variance, with the eigenvalue equals to 1.6, which is less than the critical value (e.g., < 2). It must be noted that the variance accounted for by the first contrast is larger than the variance accounted for by the item difficulties. Therefore, a secondary dimension is more likely to explicate more variance in the data than the variance is explained by the Rasch item difficulties (Linacre, 2009), indicating the multidimensionality of the test.

Rating Scale Analysis

Table 3 at the end summarizes partial credit structures of the data; the first column shows the number of items; the second column gives the number of categories which vary across items. The number of categories for each item depends on the items' total scores, and categories with no observations are not listed; the third column demonstrates the frequency and percentage of each category; column four is the observed, sample-dependent, average-measure of persons in the sample who responded in this category. It is expected that averages to increase along with category values. S.E. Mean is the standard error of the average (mean) measure of the sample of persons who selected that category; column five provides outfit mean squares for observed responses in each category level. With an expected value of 1, outfit statistics are the average of the outfit mean-squares associated with responses in each category (Linacre, 2009); and the last column represents the point-correlation between the categories. As presented in Table 3, Categories 0, 1, 2, and 3 involve the largest portion of the response categories for all of the items, indicating the low performance of test takers in the speeded cloze-elide test. With regard to the average ability, all of the items include categories in which average abilities do not ascend with category values (shown by asterisk * in the table). For example, category 6 shows an unexpected ordering pattern for item 1, categories 4, 6, 9, and 17 for item 2, categories 4, 6, and 7 for item 3, category 14 for item 4, and categories 5 and 6 for item 5. The average abilities of the examinees observed in these categories are lower than average abilities of the examinees in the next lower category. This indicates the contradiction of the RM assumption that higher categories should have higher average abilities (Linacre, 2009). In relation to outfit mean squares for each category level, most of the categories are not within the acceptable boundary. The acceptable range for fit values are 0.60 and 1.40 (Bond & Fox, 2015; Linacre, 1999; Wright & Linacre, 1994), and any value below or above this range is considered troublesome. Finally, the values of point-measure correlations show that except for category 0 in all of the items and category 1 in item 5, all correlations are positive. Negative point measure correlations indicate that the responses to the item contradict the latent trait defined by the con-

Table 2 ■ Descriptive Statistics, Item Measures, Fit Statistics, and Point-measure Correlations

Item	M	SD	Item Difficulty	Model S.E.	INFIT MNSQ	OUTFIT MNSQ	PT-Measures
EL1-W	0.79	1.31	0.19	0.10	1.20	1.22	0.56
EL2-W	1.12	2.21	0.20	0.07	1.50	0.94	0.59
EL3-W	1.32	1.66	−0.14	0.09	1.05	1.02	0.68
EL4-W	1.52	2.34	−0.12	0.07	0.70	0.80	0.67
EL5-W	1.75	2.43	−0.12	0.07	0.83	0.84	0.72
	Person Separation = 1.46				Person Reliability = 0.68		
	Item Separation = 1.61				Item Reliability = 0.72		

Note. M = mean; SD = standard deviation; SE = standard error of measurement; PT-Measure = point-measure correlations.

sensus of the items.

Discussion

Cloze-elide test is conceived as a general measure of both L1 and L2 reading comprehension (Klein-Braley, 1997; Manning, 1987) and communication skills, including listening, speaking, and writing (Manning, 1987). As Davies (1975) argued, the performance on cloze-elide test is not solely a measure of reading ability, but it is a good measure of processing speed if test takers are imposed to cross out redundant words from a text under time pressure. In the present study, an attempt was made to examine the fit of a speeded cloze-elide test to a polytomous Rasch model known as partial credit model (PCM; Masters, 1982). Previous studies have investigated the psychometric quality of cloze-elide test using traditional quantitative methods such as exploratory factor analysis and regression models. This study expands this line of research by using PCM. To the best knowledge of the author, this study is the first attempt in the literature that apply PCM, as an ordinal Rasch model, to explore the fit of a speeded cloze-elide test. To examine the fit of the model, several statistics were examined. First of all, the overall fit of different scoring techniques were compared to choose the most appropriate scoring method. The results showed that the scoring based on wrong scores has the best fit, suggesting that the speeded cloze-elide test is more likely to fit the PCM better than the other scoring techniques (e.g., correct, miss, and correct - wrong). The scores based on wrong scoring technique were thus selected for running the further analyses.

The analysis of item parameters, fit statistics, and point-measure correlations indicated that observed data structure accord with the expectation of the Rasch model, although one passage or super item (e.g., item 2) out of the five passages has infit value above the acceptable range. “High infit mean squares indicate that the items are misperforming for the people on whom the items are targeted. This is a bigger threat to validity, but more difficult to diagnose than high outfit” (Linacre, 2009, p. 596). Also, person and item reliability coefficients as well as separation values were checked. The results of reliability coefficients for both items and persons were below the expected values of 3 and 2, respectively, indicating the lack of the representativeness of the items. Item and person reliability values were relatively low, showing a narrow range of person measures, or the presence of a small number of items (Linacre, 2009).

In addition, the unidimensonality of the speeded cloze-elide test was examined because infit and outfit statistics are not very sensitive to systematic threats to unidimensionality (Smith & Plackner, 2009). Although the eigenvalue of the first contrast (factor) was below the expected value (e.g., < 2), the variance accounted for by the first contrast was higher than the variance accounted for by the item difficulties. This suggests that a secondary dimension could account for more variance in the data. The multidimensionality of the test is more likely due to the inclusion of speed factor to test administration. Research has shown that when tests are administered under time pressure, speededness can affect the test performance of examinees and is detrimental to the intended functioning of the test (Bolt, Cohen, & Wollack, 2002). Here, because speededness is the main factor for measuring cognitive processing under time limits, we can maintain that the speeded cloze-elide test is unidimensional.

Finally, the results of partial credit analysis of data structure revealed that a number of categories do not increase with category values, and OUTFIT MNSQ values for most of the categories were beyond the ideal range. As Andrich (2011) argued, when items form a testlet, “there is no reason for the thresholds to be ordered. In fact, the more local dependence you have accounted for with the testlet form, the more the thresholds will be disordered” (p.1318). Average ability values further showed that averages do not ascend along with category values. This represents that the average measure for a higher score value is lower than for a lower score value, which repudiates the hypothesis that lower score value implies lower measure, and vice versa

(Linacre, 2009). It also suggests that the category engages a narrow interval on the construct, and there are substantive problems with the definitions of rating scale categories (Linacre, 2009, p. 336). Point measure correlations indicated that all correlations were positive, excluding category 0 across all of the items and category 1 for item 5.

Limitations and Suggestions for Future Research

When considering the results from this study, several limitations for future research are important to note. Different limitations of the study are associated to psychometric properties of ordinal Rasch models. First, polytomous Rasch models do not take into account examinees' response patterns to all individual items which lead to losing a great deal of information (Eckes, 2011). Second, as blanks within a passage are greatly dependent with relatively low number of independent blanks, polytomous models probably generate spurious reliability estimates and biased parameter estimation (Wainer & Wang, 2000; Wang & Wilson, 2005a, 2005b). Third, Forthmann et al. (2020)t argue that polytomous Rasch models such as PCM do not allow practitioners to incorporate and estimate time limits, irrespective of item difficulty, within the model, and that item difficulty parameters only relate to item main effects. They highlight that "if the time limit for an easy text is short and its total raw score is smaller than the total raw score for a difficult text with a longer time limit, the easier texts will turn out to have a higher difficulty parameter" (p. 2).

More importantly, as noted earlier, in cloze procedures and C-tests, each text is most often viewed as a polytomous item, and the scores on gaps of a single passage are added to get a total score. In this case, a polytomous (ordinal) Rasch model can be used to examine the fit of data to the model. However, this method has some limitations. First, these models involve a large number of parameters to be estimated and consequently, a large sample size is required to obtain stable and accurate parameter estimations (Eckes, 2011). Second, when these tests are administered under time pressure, examinees typically fail to complete the gaps and most of them are left unanswered or with very few observations. This results in low accuracy of parameter estimation and biased person parameter estimation (Li, 2013). With this in mind, the current ordinal Rasch models are unable to model the responses of test takers. Instead, IRT models for count data would be more appropriate, including Rasch–Poisson Counts Model (RPCM; Baghaei & Doebler, 2019; Rasch, 1960) and the Conway–Maxwell–Poisson counts model (Forthmann, Gühne, & Doebler, 2019). Future research could examine the utility of these models for investigating the fit of speeded cloze-elide test.

Authors' note

We want to acknowledge Ms. Samaneh Zare and Dr. Ali Akbar Boori for sharing their dataset and allowing us to use it in our study.

References

Alderson, J. C. (1979). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading*, *2*(2), 108–119. doi:10.1111/j.1467-9817.1979.tb00198.x

Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning*, *30*(1), 59–76. doi:10.1111/j.1467-1770.1980.tb00151.x

Alderson, J. C. (1983). The cloze procedure and proficiency in english as a foreign language. In W. Oller (Ed.), *Issues in language testing research* (pp. 205–217). Rowley, MA: Newbury House.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Anderson, J. (1976). *Psycholinguistic experiments in foreign language testing*. St. Lucia, Queensland: University of Queensland Press.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. doi:10.1007/BF02293814

Andrich, D. (2011). Testlets and threshold disordering. *Rasch Measurement Transactions*, *251*(1), 1318–1399.

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System*, *29*(2), 209–219. doi:10 . 1016 / S0346 - 251X(01) 00012-4

Bachman, L. F. (1981). The trait structure of cloze test scores. (pp. 1–99). Champaign-Urbana: TESOL Midwest Regional Conference and Illinois TESOL/BE Convention.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletion. *TESOL Quarterly*, *19*(3), 535–56. doi:10.2307/3586277

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baghaei, P., & Doebler, P. (2019). Introduction to the rasch poisson counts model: An R tutorial. *Psychological Reports*, *122*(5), 1967–1994. doi:10 . 1177 / 0033294118797577

Baghaei, P., & Effatpanah, F. (2022). *Elements of psychometrics (2nd ed.)* Mashhad, Iran: Sokhan Gostar Publishing.

Baghaei, P., & Grotjahn, R. (2014). The validity of C-tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In R. Grotjahn (Ed.), *Der C-test: Aktuelle tendenzen/the C-test: Current trends* (pp. 163–171). Frankfurt/M.: Lang.

Baker, B. A. (2011). *Use of the cloze-elide task in high-stakes english proficiency testing* (tech. rep. No. 9). Spaan Fellow Working Papers in Second or Foreign Language Assessment.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348. doi:10.1111/j.1745-3984.2002.tb01146.x

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences (3rd ed.)* New York: Routledge.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension tests scores. *Journal of Reading*, *10*(5), 291–299.

Bowen, J. D. (1978). The identification of irrelevant lexical distraction: An editing task. *TESL Reporter*, *12*(1), 14–16.

Carroll, J. B. (1961). *Fundamental considerations in testing for english language proficiency of foreign students*. In Testing the English proficiency of foreign students. Washington, DC: Center for Applied Linguistics.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing*, *7*(2), 121–146. doi:10.1177/026553229000700201

Davies, A. (1975). Two tests of speeded reading. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 1–99). Washington, DC: Center for Applied Linguistics.

Davies, A. (1989). Testing reading speed through text retrieval. In C. N. Candlin & T. F. McNamara (Eds.), *Language learning and community* (pp. 1–99). Sydney, NSW: NCELTR.

Davis, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American english: Word sketches, collocates & thematic lists*. New York: Routledge.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton, FL: Chapman & Hall/CRC Press.

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the rasch poisson counts model. *Learning and Individual Differences*, *52*, 121–128. doi:10.1016/j.lindif.2015.01.013

Eckes, T. (2010). Rasch models for C-tests: Closing the gap on modern psychometric theory. In A. Berndt & K. Kleppin (Eds.), *Sprachlehrforschung: Theorie und empirie – festschrift f'ur r"udiger grotjahn* (pp. 39–49). Frankfurt, Germany: Lang.

Eckes, T. (2011). Item banking for C-tests: A polytomous rasch modeling approach. *Psychological Test and Assessment Modeling*, *53*(4), 414–439.

Eckes, T., & Grotjahn, R. (2006a). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325. doi:10.1191/0265532206lt330oa

Eckes, T., & Grotjahn, R. (2006b). C-tests als anker fur testdaf: Rasch-analysen mit dem kontinuierlichen ratingskalen-modell [C-tests as an anchor in testdaf: Rasch-analyses with the continous rating scale model]. In R. Grotjahn (Ed.), *Der C-test: Theorie, empirie, anwendungen* (pp. 167–193). Frankfurt am Main, Germany: Peter Lang.

Effatpanah, F. (2019). *Cognitive diagnostic assessment of iranian efl university students' l2 writing ability: Selecting the best model (unpublished master's thesis)*. Mashhad, Iran: Islamic Azad University.

Effatpanah, F., & Baghaei, P. (2021). Cognitive components of writing in a second language: An analysis with the linear logistic test model. *Psychological Test and Assessment Modeling*, *63*(1), 13–44.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based english language screening tool. *Language Assessment Quarterly*, *5*(3), 173–194. doi:10.1080/15434300802229334

Farhady, H. (1979). The disjunctive fallacy between discrete-point and integrative tests. *TESOL Quarterly*, *13*(3), 347–357. doi:10.2307/3585882

Farhady, H. (1996). Varieties of cloze procedure in efl education. *Roshd Foreign Language Teaching Journal*, *12*, 217–229.

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-tests. *Journal of Psychoeducational Assessment*, *38*(6), 692–705. doi:10.1177/0734282919889262

Forthmann, B., Gühne, D., & Doebler, P. (2019). Revisiting dispersion in count data item response theory models: The conway–maxwell–poisson counts model. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 32–50. doi:10.1111/bmsp.12184

Friedman, M. M. (1964). *The use of the cloze procedure for improving the reading comprehension of foreign students at the university of florida (unpublished doctoral dissertation)*. Miami: University of Florida.

Gaies, S. J. (1987). Validation of the noise test. In C. K.-B. R. Grotjahn & D. K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (pp. 41–74). Bochum: Brockmeyer.

Gaies, S. J., Gradman, H. J., & Spolsky, B. (1977). Towards the measurement of functional proficiency: Contextualization of the noise test. *TESOL Quarterly*, *11*(1), 51–57. doi:10.2307/3585591

Gradman, H. L., & Spolsky, B. (1975). Reduced redundancy testing: A progress report. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 59–70). Arlington, VA: Center for Applied Linguistics.

Grotjahn, R. (2010). *Der C-test: Beitrage aus der aktullen forschung the C-test: Contributions from current research*. Frankfurt/ M: Lang.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer- Nijhoff.

Harsch, C., & Hartig, J. (2010). Empirische und inhaltliche analyse lokaler abh"angigkeiten im C-test [empirical and content analysis of local dependencies in C-tests]. In R. Grotjahn (Ed.), *Der C-test: Beitr"age aus der aktuellen forschung [the C-test: Contributions from current research]* (pp. 193–204). Frankfurt, Germany: Lang.

Heckman, R. W., Tiffin, J., & Snow, R. E. (1967). Effects of controlling item exposure in achievement testing. *Educational and Psychological Measurement*, *27*(1), 113–125. doi:10.1177/001316446702700111

Hudson, T. (2007). *Teaching second language reading*. New York: Oxford University Press.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research*, *34*(2), 245–268. doi:10 . 1207 / S15327906Mb340205

Johansson, S. (1973). *Partial dictation as a test of foreign language proficiency*. Lund: University of Lund, Department of English.

Johansson, S. (1974). Controlled distortion as a language testing tool. In H. S. Qvistgaard & H. Spang-Hanssen (Eds.), *J* (pp. 397–411). Heidelberg: Applied linguistics, problems, and solutions: AILA proceedings Copenhagen. III . : Julius Groos Verlag.

Jonz, J. (1976). Improving on the basic egg: The multiple choice cloze. *Language Learning*, *26*(2), 255–65. doi:10.1111/j.1467-1770.1976.tb00276.x

Klein-Braley, C. (1981). *Empirical investigations of cloze tests (unpublished doctoral dissertation)*. Germany: University of Duisburg.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84. doi:10.1177/026553229701400104

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing*, *1*(2), 134–146. doi:10.1177/026553228400100202

Lee, L., & Gunderson, E. (2011). *Select reading*. London: Oxford University Press.

Lee, S. H. (2008). Beyond reading and proficiency assessment: The rational cloze procedure as stimulus for integrated reading, writing, and vocabulary instruction and teacher–student interaction in esl. *System*, *36*(4), 642–660. doi:10.1016/j.system.2008.04.002

Li, E. F. (2013). The impact of unobserved extreme categories on item and person estimates: A simulation study. In Zhang & H. Yang (Eds.), *Q* (pp. 117–128). Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding: Springer.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103–122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 1–99.

Linacre, J. M. (2009). *A user's guide to winsteps*. Chicago, IL: Winsteps.

Manning, W. H. (1987). *Development of cloze-elide tests of english as a second language*. Princeton, NJ: Educational Testing Service.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10 . 1007 / BF02296272

Norris, J. M. (2018). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7–33). Germany: Lang.

O'Reilly, R. P., & Streeter, R. E. (1977). Report on the development and validation of a system for measuring literal comprehension in a multiple-choice cloze format: Preliminary factor analytic results. *Journal of Literacy Research*, *9*, 45–69. doi:10 . 1080 / 10862967709547206

Oller, J. W., Jr. (1971). Dictation as a device for testing foreign language proficiency. *English Language Teaching Journal*, *25*(3), 254–259. doi:10.1093/elt/XXV.3.254

Oller, J. W., Jr. (1973). Cloze tests of language proficiency and what they measure. *Language Learning*, *23*(1), 105–18. doi:10.1111/j.1467-1770.1973.tb00100.x

Oller, J. W., Jr. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Spracen*, *75*(2), 165–174.

Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Oller, J. W., Jr., & Conrad, C. (1971). The cloze technique and esl proficiency. *Language Learning*, *21*(2), 183–95. doi:10.1111/j.1467-1770.1971.tb00057.x

Ozete, O. (1977). The cloze procedure: A modification. *Foreign Language Annals*, *10*(5), 565–568. doi:10.1111/j.1944-9720.1977.tb03033.x

Porter, D. (1976). Modified cloze procedure: A more valid reading comprehension test. *English Language Teaching Journal*, *30*(2), 151–155. doi:10.1093/elt/XXX.2.151

Raatz, U. (1985). Tests of reduced redundancy- the C-test, a practical example. In Klein-Braley & U. Raatz (Eds.), *C* (pp. 14–19). Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis . Bochum: AKS.

Raatz, U., & Klein-Braley, C. (1981). The C-test: A modification of the cloze procedure. In C. K.-B. T. Culhane & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 113–145). Colchester, UK: University of Essex.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53*(3), 349–359. doi:10.1007/BF02294217

Ruddell, R. B. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. *Improvement of Reading Through Classroom Practice*, *9*, 298–303.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. doi:10.1111/j.1745-3984.1991.tb00356.x

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in rasch measurement. *Journal of Applied Measurement*, *10*(4), 424–437.

Spolsky, B. (1968). What does it mean to know a language, or how do you get someone to perform his competence? In *The second conference on problems in foreign language testing* (pp. 1–24). U.S.A.: University of Southern California.

Spolsky, B. (1969). Reduced redundancy as a language testing tool. (pp. 1–18). Cambridge, England: Applied Linguistics.

Spolsky, B. (1971). Reduced redundancy as a language testing tool. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 383–390). Cambridge: Cambridge University Press.

Spolsky, B., Sigurd, B., Sato, M., Walker, E., & Arterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning*, *18*(3), 79–101. doi:10.1111/j.1467-1770.1968.tb00224.x

Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly*, *17*(1), 29–38. doi:10.2307/3586422

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. doi:10.1177/107769905303000401

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260. doi:10.1111/j.1745-3984.1989.tb00331.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score toefl. *Journal of Educational Measurement*, *37*(3), 203–220. doi:10.1111/j.1745-3984.2000.tb01083.x

Wang, W. C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296–318. doi:10.1177/0146621605276281

Wang, W. C., & Wilson, M. (2005b). The rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. doi:10.1177/0146621604271053

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370–399.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement (4th ed* (pp. 111–153). Westport, CT: Praeger.

Zare, S., & Boori, A. A. (2018). Psychometric evaluation of the speeded cloze-elide test as a general test of proficiency in english as a foreign language. *International Journal of Language Testing*, *8*(2), 33–43.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on the journal's web site.

Citation

Effatpanah, F., & Baghaei, P. (2022). Evaluating different scoring methods for the speeded cloze-elide test: The application of the Rasch partial credit model. *The Quantitative Methods for Psychology*, *18*(3), 241–254. doi:10.20982/tqmp.18.3.p241

Received: 08/06/2022 ∼ Accepted: 30/08/2022

Table 3 follows.

Table 3 ■ Summary of Category Statistics for the Speeded Cloze-elide Test

Items	Category	Observed		Average	Ability	Outfit MNSQ	PT-Measures
		Count (%)		(S.E. Mean)			
EL1-W	0	83	(55%)	−3.05	(0.15)	1.3	−0.51
	1	41	(27%)	−2.03	(0.16)	1.2	0.13
	2	13	(9%)	−1.20	(0.29)	1.6	0.24
	3	9	(6%)	−0.34	(0.20)	0.9	0.34
	4	1	(1%)	0.19		0.2	0.14
	5	1	(1%)	0.28		0.3	0.14
	6	1	(1%)	−0.03*		2.1	0.13
	10	1	(1%)	0.66		0.4	0.16
EL2-W	0	80	(53%)	−3.27	(0.13)	1.0	−0.65
	1	35	(23%)	−2.00	(0.16)	0.9	0.13
	2	19	(13%)	−1.16	(0.15)	0.6	0.30
	3	5	(3%)	−0.25	(0.15)	0.2	0.26
	4	3	(2%)	−0.33*	(0.31)	1.0	0.19
	5	3	(2%)	0.20	(0.11)	0.1	0.24
	6	1	(1%)	0.19*		0.1	0.14
	8	1	(1%)	0.66		0.5	0.16
	9	1	(1%)	0.10*		1.3	0.13
	13	1	(1%)	0.70		0.0	0.17
	17	1	(1%)	0.10*		10.0	0.13
EL3-W	0	53	(35%)	−3.55	(0.18)	1.1	−0.58
	1	52	(35%)	−2.32	(0.14)	1.1	0.02
	2	24	(16%)	−1.65	(0.19)	1.2	0.21
	3	7	(5%)	−0.68	(0.19)	0.5	0.25
	4	5	(3%)	−0.70*	(0.22)	1.5	0.20
	5	3	(2%)	0.27	(0.19)	0.2	0.25
	6	3	(2%)	0.22*	(0.10)	0.3	0.24
	7	2	(1%)	0.13*	(0.03)	0.9	0.19
	10	1	(1%)	0.70		0.4	0.17
EL4-W	0	61	(41%)	−3.54	(0.14)	0.9	−0.65
	1	39	(26%)	−2.31	(0.19)	1.5	0.02
	2	20	(13%)	−1.78	(0.12)	0.7	0.15
	3	12	(8%)	−0.89	(0.12)	0.4	0.29
	4	9	(6%)	−0.35	(0.15)	0.4	0.34
	5	4	(3%)	−0.17	(0.25)	0.7	0.24
	6	1	(1%)	0.19		0.0	0.14
	11	1	(1%)	0.41		0.0	0.15
	12	1	(1%)	0.66		0.0	0.16
	14	2	(1%)	0.49*	(0.21)	0.7	0.22
EL5-W	0	53	(35%)	−3.74	(0.14)	0.9	−0.68
	1	40	(27%)	−2.49	(0.14)	1.0	−0.05
	2	23	(15%)	−1.69	(0.14)	0.6	0.19
	3	11	(7%)	−1.03	(0.19)	0.5	0.25
	4	8	(5%)	−0.45	(0.16)	0.3	0.30
	5	6	(4%)	−0.46*	(0.34)	2.0	0.26
	6	2	(1%)	−0.53*	(0.38)	2.1	0.14
	7	3	(2%)	−0.08	(0.17)	1.0	0.22
	10	1	(1%)	0.06		1.5	0.13
	12	2	(1%)	0.55	(0.14)	0.1	0.22
	15	1	(1%)	0.66		0.4	0.16

Note. S.E. Mean = standard error of mean; MNSQ = mean-square; PT-Measure = point-measure correlations.