2014 ■ vol. 10 ■ no. 2

DOI: 10.20982/tqmp.10.2.p095

Psychometric validation of the Sentence Verification Technique to assess L2 reading comprehension ability

François Pichette [a], Sébastien Béland [b], Linda de Serres [c], Marc Lafontaine [d]

[a] UER Sciences humaines, lettres et communication, Téluq

[b] Département d'administration et fondements de l'éducation, Université de Montréal.

[c] Département de lettres et communication sociale, Université du Québec à Trois-Rivières

[d] Département de Langues, linguistique et traduction, Université Laval

Abstract ■ English teachers use the Sentence Verification Technique (Royer et al., 1979) to determine the readability of written material for their classes. This process requires students to read short passages from a book, followed by isolated sentences. These sentences can be either identical or different from the original passages, in their meaning as well as in their form. For each sentence, students must indicate whether or not its content corresponds to that of the original passage. This paper reports on the design and assessment of an SVT test created for measuring reading comprehension ability, based on four English texts. The instrument was administered to 171 adult English learners, of various levels of English proficiency. The data were analyzed using both traditional psychometric methods and the Rasch model. Results indicate that the test shows high internal consistency, that it respects the basic assumptions behind the Rasch model, and that it is in the recommended range of difficulty for that technique.

Keywords ■ Sentence Verification Technique; Reading comprehension; Reading test; Readability; Rasch model

francois.pichette@teluq.ca

Introduction

Sentence Verification Technique (hereafter, SVT) tests were first developed by psychologist Mike Royer and his colleagues in the late 1970s (Royer, Hastings, & Hook, 1979), and have since been used for a variety of languages and learning contexts. These tests are based on short text passages, normally consisting of 12 sentences each. Following the reading of each passage, and without going back to it, the reader is presented with 16 sentences, and must indicate by Yes or No for each one if what it says corresponds to information that was provided in the passage. The 16 sentence items of a traditional SVT test are prepared as follows: four of the sentences in the text are left intact, four others are paraphrased by changing as many words as possible while preserving their meaning, four others have their meaning transformed by changing only one or two words, and four other sentences are added as distractors, providing information not conveyed in the text.

SVT tests measure reading for meaning, assuming that comprehension involves the construction of meaning without necessarily recalling the exact words, and acknowledging the important role played by memory in reading (Baddeley, Logie, Nimmo-Smith, & Brereton, 1985; de Jonge & de Jong, 1996). SVT tests have thus been used to assess how well readers process the propositional content of specific text passages to form a coherent mental image reflecting correctly the events described in those passages. The reliability of a four-passage SVT test (64 sentence questions) is typically between .7 and .8 (Royer, 2005). For more details about the nature and validity of SVT tests, see the extensive overview by Royer (2004).

Until now, SVT tests had been used by non-psychometricians for testing the comprehensibility of reading material for a targeted audience (Royer, 2004). The goal of this study was to assess some of the psychometric qualities of an SVT-based reading test using both traditional psychometric methods and the Rasch model. Therefore, the SVT test developed in this study serves no pedagogical purposes such as facilitating learning or helping to develop reading skills; rather, it is aimed at assessing reading comprehension ability as defined above. Such a new instrument is intended to show at least three possible advantages for its users: (1) SVT tests are faster to administer than usual standardized tests; (2) They would be free of charge, and (3) They are motivating for students, since

they are invited to read short, interesting texts without having anything to write.

Since L2 competence has long been shown to be a determinant of L2 reading ability (see Laufer & Ravenhorst-Kalovski, 2010; Pichette, Segalowitz, & Connors, 2003; Shiotsu & Weir, 2007; Yamashita, 2001), scores on our new instrument should correlate with English competence. A second goal of our study is to confirm the fact that SVT scores are sensitive to language proficiency as measured through self-assessment.

Test Design

The lengthy process of designing our instrument warrants a separate section, considering the amount of information to be presented. The two studies aimed at evaluating our instrument will be described in the Method section.

Texts

Text selection. In order to obtain an instrument designed for a wide audience, the texts which served as the basis of our SVT test needed a topic of interest to almost anyone. Additional criteria were followed, so that no undesirable factor interferes with the normal process and assessment of reading comprehension. As recommended for such tests, stories had to be true and needed to contain a beginning and an end, while not being too culturally charged or biased (see Royer, 2004).

The many elements to be avoided include:

- stories with a high amount of data, numbers and figures. Not only would that be non-motivating, but the instrument would tend to tap rote memory rather than comprehension;
- well-known stories (e. g., *Three little pigs*), for which selecting correct answers would be made possible without the participant's actually reading the text;
- the presence of emotionally-charged topics, such as violence, sex, religion, politics, war, to avoid as much as possible interference from affective variables, where people may be offended or uncomfortable when taking our test; and
- humor, because of its dominant nonlinguistic components.

Once these criteria were met, eight texts were kept and modified so that they contained 12 sentences each, before measuring their readability.

Readability scales. For designing a test suitable for a wide variety of reading proficiencies, we needed texts of various readability levels, preferably equidistant on a readability continuum. It was decided that the texts selected would be submitted to two readability scales of a different nature before creating our instrument.

The main scale selected was the Flesch-Kincaid (FK) formula, which is recognized as the most reliable formula among the readily available ones (see Schinka & Borum, 1993). Numerous authors, even among proponents of alternative formulas, have concluded that vocabulary frequency and sentence length are the best indicators of text difficulty (see DuBay, 2004). The Flesch-Kincaid formula, included in MS Word, is based on sentence and word length, given that word length strongly correlates with word frequency with r values of up to 0.997 (Strauss, Grzybek, & Altmann, 2007).

It is also recommended that components beyond the sentence level be taken into consideration when assessing text difficulty (Wagenaar, Schreuder, & Wijlhuizen, 1987). Since the Flesch-Kincaid scale only considers word length, a component beyond the word level that is also encountered in some readability formulas is the clause. To that effect, a secondary scale was selected: the Sentence Complexity Index (SCI), included in Corel's WordPerfect software. As its name suggests, the SCI considers the complexity of sentence structure in terms of their clauses by computing the ratio of subordinate clauses of all kinds to the number of sentences. Its purpose was to confirm the sequence of texts in terms of difficulty, ensuring that no major discourse-type factor would hamper readability. Although scores on both scales tend to correlate strongly (e.g., $r > .75$ in Barrio Cantalejo & Simón Lorda, 2003), a few texts were rejected when the SCI scale did not confirm the FK scale in its assessment of text difficulty. The rest of the study relied exclusively on Flesch-Kincaid scores.

Among the eight candidates, the four texts that were closest to the four targeted readability zones were retained. They later underwent a series of minor modifications to bring them even closer to the readability scores that were targeted. Such modifications consisted of replacing a few words by paraphrases or synonyms of a different length and/or frequency (e.g., replacing *smart* by *intelligent*), which affected the readability scores in the direction we wanted. Table 1 below shows the titles of the four texts along with the readability levels that were reached for both scales. The following step was the creation of the 16 sentence items for each of these four texts.

Table 1 ■ Texts and their readability scores

Text	FK (100 = very easy)	Interpretation	SCI (100 = very complex)
4. A Special Volunteer	78	Fairly easy	27
1. The Person That Caused the Titanic to Sink	66	Standard	40
3. A Puzzling Parrot	49	Difficult	42
2. The First Frog Without Lungs	28	Very difficult	75

Note: For all tables, the number before each text indicates its sequence in the test battery.

Items

Item ratio. In order to create the best instrument possible, we built our instrument on five meaning changes and five paraphrases, leaving two originals to which we added four distractors. This decision follows a recommendation by Royer, to the effect that such a ratio would offer "better reliability and validity than the balanced version of the test because paraphrases and meaning change sentences have better discriminatory properties than originals and distractors, which are more easily identified" (Royer, 2004: 10-11).

Item creation. In building SVT tests, meaning changes are normally the first items to be created, because relatively few sentences lend themselves to that kind of modification. Paraphrases are the next step, followed by the remaining sentences that are left intact and serve as originals. Distractors are the final additions to the list.

Item creation followed recommendations based on decades of SVT building and use (see Royer, 2004). For any modification performed, items must match the original text for vocabulary frequency, sentence length, and structure. A meaning change consists of substituting only one or two words to the sentence, leading to a different meaning inconsistent with the text, all the while avoiding any bizarre effect. A delicate balance has to be reached somewhere in between a modification that is too subtle, and one that is too obvious, both cases representing invalid items. A paraphrase consists of an item bearing the same meaning as the original sentence after changing as many words as possible. Finally, distractors are added which have to be different in meaning from any original sentence in the passage. Containing information not present in the text, they often reflect the result of incorrect inferences that a reader could potentially make. In the case of distractors too, the test builder must avoid the presence of any striking element that would have been noticed by the reader if it had been encountered in the text.

Once the 16 items are created, they are displayed randomly on the back of the text they pertain to, but with the usual precaution of having the items related to the first half presented in the first half of the items section. This measure is to prevent the participant from encountering among the first items a sentence that was just read at the end of the text, in which case choosing the correct answer could be explained by short-term memory.

Piloting

The instrument was first piloted with 21 university students. This procedure was aimed at confirming our expectations regarding test duration and the general profile of scores based on text difficulty. Comments were obtained from the participants immediately following the test, which allowed for fine-tuning by correcting typos, removing ambiguities, and solving minor unforeseen issues with some items. The resulting SVT test may be found in Appendix A.

Method

Participants

This new instrument was administered to 171 adults, French-speaking ESL learners, of various levels of English proficiency. However, there were no absolute beginners, for whom performing the reading tasks would have been too difficult. All participants were university students, and money was awarded post hoc as participation prizes in the form of a lottery.

Material

The SVT test was taken by all participants. Despite an equal number of sentences across texts, test length increased with text difficulty. The four texts were put together in the following sequence: (1) standard, (2)

Table 2 ■ *Detailed scores for each text (N = 168)*

Text	Flesch-Kincaid (100 = very easy)	Level	SVT Mean (%)	SD	SE	Conf. Interval (95%)	
						Lower	Upper
4. Dog	78	Fairly easy	86.1	11.25	.89	84.35	87.85
1. Titanic	66	Standard	90.6	8.50	.67	89.32	91.97
3. Parrot	49	Difficult	84.5	9.51	.75	83.03	85.99
2. Frog	28	Very difficult	74.1	11.67	.92	72.29	75.92

very difficult, (3) difficult, and (4) fairly easy, according to Flesch-Kincaid scores. The reason for that sequence is of a motivational nature: encountering a difficult or very difficult text at the beginning could have discouraged several participants and made them reluctant to carry on. A similar reason explains why the easiest was kept for last: encountering the same difficult texts at the end might entice them to either give up or not do it properly. It was felt that more data, and of a more reliable nature, would be collected by so proceeding.

On the cover page, participants were asked to indicate their estimated level of English proficiency ranging among the following choices: beginner, high beginner, intermediate, high intermediate, advanced, near-native, native speaker. The first six levels correspond to the six levels (A1, A2, B1, B2, C1, C2) of the Common European Framework of Reference for Languages (henceforth CEFR; Council of Europe, 2011), to which we added the native speaker status as a seventh option. We opted for a choice of labels that were more transparent for our students, who were not familiar with the acronyms used by CEFR for naming their reference levels.

Procedure

Participants took the test in class, during one of their courses. The instructions for the test were in French, the everyday language for all of them. There was no time limit for doing the test, and participants could leave once they were finished. The average total testing time was about 22 minutes, with durations ranging from 16 to 25 minutes. Prior to taking the test, a sample test was shown to the participants, with an explanation of how to complete the test, along with a short explanation of each of level of competence in view of their self-assessment.

Scores were compiled for each of the 64 items, for each of the four texts and, finally, for the test as a whole. One point was awarded for a correct answer and none for an incorrect one. Scores were also associated with self-evaluated proficiency, ranging from 0 for a hypothetical absolute beginner to 7 for being a native or native-like speaker. Research over the years has shown that self-ratings for language competence tend to be accurate and correlate moderately to highly with competence scores on standardized tests, as concluded by LeBlanc & Painchaud (1985) , Blanche and Merino (1989), and Ross (1998) from their extensive review of earlier research on self-assessment. Later studies lent further support to the value of self-assessment (e.g., Oscarson, 1997; Wilson & Lindsey, 1999; Yoshizawa, 2009).

Analysis

This paper presents two analyses. The first analysis consisted in evaluating the internal consistency of our instrument through Cronbach's alpha and calculating the proportion of good answers, seen as reflecting item difficulty in Classical test theory (Lord & Novick, 1968). For example, if the proportion of correct answers for item A is 0.80 and for item B is 0.30, we conclude that item A is easier than item B.

The second analysis is a Rasch-based analysis of our items. In this case, the eRm R package (Mair, Hatzinger, & Maier, 2010) was used to estimate the parameters of the Rasch model.

The Rasch model (1960) calculates the probability for a participant to obtain the correct answer for an item. The equation is as follows

$$P(\theta) = P(x_i = 1 \mid \theta, b_i) = \frac{\exp\left[(\theta - b_i)\right]}{1 + \exp\left[(\theta - b_i)\right]} \qquad (1)$$

where θ is the parameter for the participants' ability and b_i is the difficulty parameter for item *i*. Compared with models of Classical test Theory, this model had the advantage of considering on the same measurement unit the ability of participants who take a test as well as the difficulty of the test items. In other words, it allows

Table 3 ■ Readability measures vs. SVT scores (N = 168)

Text	SVT Mean (%)	Flesch-Kincaid (100 = very easy)	LEXILE
4. Dog	86.1	78	690
1.Titanic	90.6	66	640
3. Parrot	84.5	49	1100
2. Frog	74.1	28	1440

for comparing students without being constrained by a specific test sample.

Theoretically, two general conditions must be met for this model to be applied adequately to a set of dichotomous data. First, the items must be independent from one another. Second, there must be only one main ability assessed by the test, referred to as the unidimensional postulate. For example, in the context of a test of English as a second language, only ability in English should be tested.

Before conducting the analyses, the following two factors were considered.

Treatment of missing data. Even though no treatment method will ever eliminate entirely the bias caused by missing data (see Rousseau, 2006), we adopted the most common method in the field of human sciences (e.g., Raîche, 2002; see Enders, 2010) which consists of considering missing responses as being incorrect, thus assigning them a score of zero. Logically, the items that make participants hesitate are the ones that are most difficult for them, thus most likely to be missed. In addition, the number of missing answers is so low (42 out of 10944, or 0.38%) that a comparison made between different treatment options (replacement by zero, replacement by the participant's mean score for other items, replacement by the overall average for that item) had no impact on Cronbach's alpha or on the data profile.

Exclusion of three participants. It is important for participants to perform the tasks in good faith and with the effort and seriousness that are expected from them, so that we actually measure what we intend to measure, and for the collected data to be reliable and valid. However, despite all the precautions taken, some people will participate wrongly assuming that they will get benefits from the professor who is testing them, or that not participating will be prejudicial, and answering haphazardly to finish rapidly. In some cases, participants will lie and pretend to meet the requirements for participating, and their inclusion can bias our data. Such situations, among others, lead to the undue participation of certain people, or the presence of unwilling participants, and warrant their post hoc exclusions from our analyses. Three participants out of the 171 (i.e. less than 2%) were excluded for not performing at chance or above chance on the test, for having completed it too rapidly, and for not having completed or for performing poorly on the last text, which was the easiest. In fact, in another study (Pichette, Béland, Magis, & Raîche, 2010) the correctness of those decisions were confirmed by the use of person-fit indices.

Results

Analysis of Test Results Using Classical Test Theory

Descriptive statistics for global scores. Scores and related data for each text and for the SVT test components with shortened titles are presented in Table 2, along with their Flesch-Kincaid readability scores. In that table, the texts are displayed not according to their actual sequence in the test package, but from the easiest (*Dog*) to the most difficult (*Frog*) according to our readability scales, for ease of comparison. The number that precedes each title indicates the sequence in which they were done.

Discrepancies between scores and estimated readability. Means for our texts increased as their established levels of difficulty decreased, except for "A special volunteer" predicted to be the easiest, but yielding only the second lowest mean. This could be due to the fact that it was given as the fourth and final text, at a point where participants may have paid less attention and been anxious to finish. However, another explanation is that our scale for readability might not have been precise enough: it could be possible that the Titanic text was actually easier than the Dog text, or at least that they would not be statistically different in terms of readability.

To explore that possibility, all four texts were submitted to a different reading scale, the Lexile scale (Metametrics, 2009), that we discovered while our data

Table 4 ■ Correlations between texts (N = 168)

	4. Dog (Fairly easy)	1. Titanic (Standard)	3. Parrot (Difficult)	2. Frog (Very difficult)
4. Dog (Fairly easy)	1.000	.294**	.431**	.225**
1. Titanic (Standard)		1.000	.277**	.203**
3. Parrot (Difficult)			1.000	.390**
2. Frog (Very difficult)				1.000

Note: ** = Correlation is significant at p < .01 (bilateral)

were being analyzed. Like Flesch-Kincaid, the Lexile scale is based on vocabulary frequency and sentence length, the two variables shown to best reflect text readability. However, as was mentioned, FK only assumes word frequency based on word length. For example, FK would consider that *government* and *intelligent* are rare words based on their length. Such cases are probably negligible for long texts, but their impact could be significant in the case of very short texts such as these, with FK scores becoming misleading. On the contrary, Lexile assigns an actual frequency measure to each word, based on the number of occurrences in a reference corpus of 600 million words. Table 3 below shows our Lexile measures compared to FK. As evidenced in the Table, the Lexile score profile matches our SVT score profile, that is, the easier the text, the higher the SVT score.

The only means that are not statistically different are for Texts 3 (M=84.5) and 4 (M=86.1) as shown by a Tukey test performed using SPSS. The mean for Text 1 (*Titanic*; 90.6 %) is therefore significantly higher than for Text 4 (*Dog*; 86.1%). This difference was unexpected, because as evidenced in Table 3, expectations based on Flesch-Kincaid were that *Titanic* would be more difficult than *Dog*. However, our participants scored higher on *Titanic*. Later use of the Lexile scale confirmed that *Titanic* is indeed easier (640) than *Dog* (690). The hierarchy of our texts in terms of difficulty using Lexile matches the hierarchy based on SVT scores, which is not the case using Flesch-Kincaid. Means in Table 3 thus suggest that scores on our instrument vary with text readability as was shown in previous research. The overall mean for the test is 83.7%

Internal consistency. The internal consistency of our instrument was also examined. Table 4 below shows correlations between scores on the four texts that make up the SVT test. As shown in the Table, all paired correlations are significant below .01. The correlations are not very high, but this is not unexpected, since the texts are of a different nature and show different distributions. We also obtained a Cronbach's alpha of 0.82 for the whole test, which denotes high consistency.

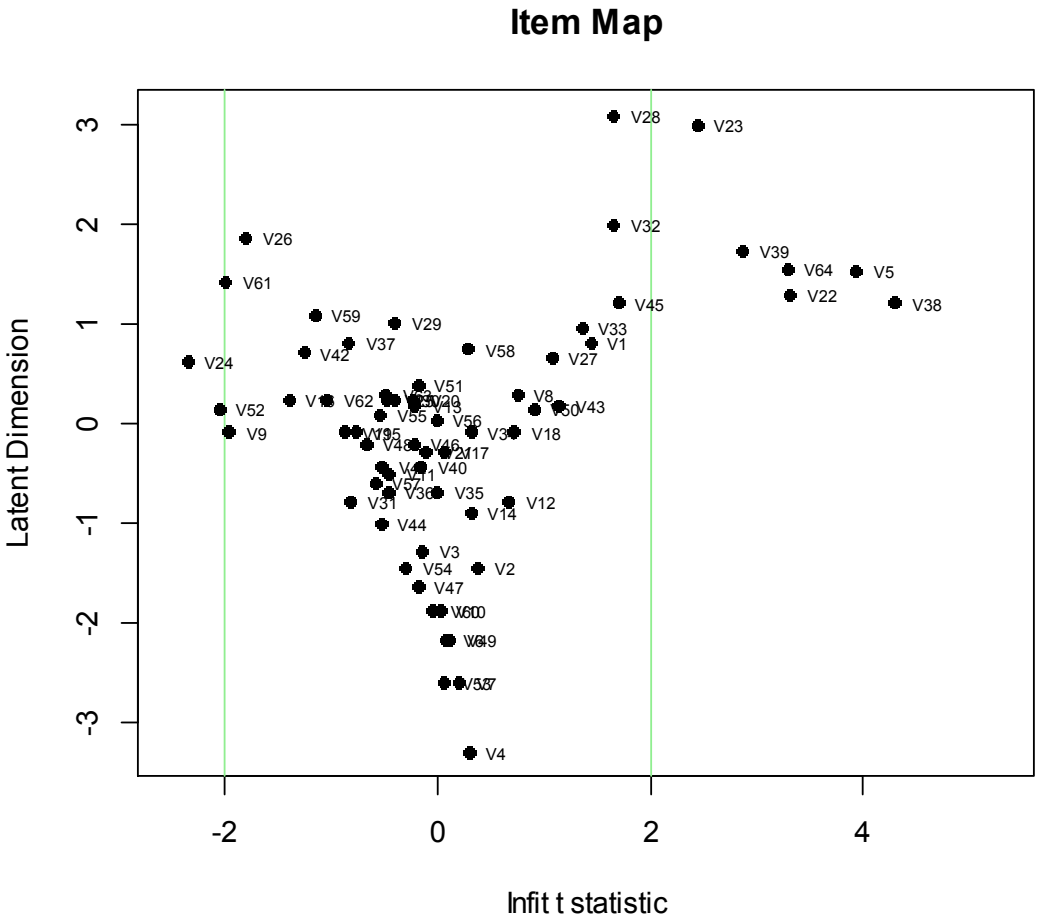
Item difficulty. Calculations of item difficulty based on Classical Test Theory suggest that the instrument is relatively easy, the mean difficulty being 0.82. While none of the 64 items was missed or achieved by all participants, the percentage of success for individual items ranges from 31 % to 99 %.

The two items that are markedly more difficult than the rest are two meaning changes: items 23 ("This remarkable discovery is an important step in the study of amphibians, according to the specialized magazine *Current Biology*") and 28 ("In addition, the specialists speculate that the absence of lungs helps the animal to float in the water and to move more comfortably in the rivers in which it lives").

The third item which is closer to the limit is item 32 ("The frog, no more than 70 mm (2.8 in) long, lives in warm, translucent, and slow moving rivers in remote areas of the rainforests of Kalimantan, the Indonesian part of the island of Borneo"). As in the case of items 23 and 28, item 32 belongs to the most difficult text of the four.

Involvement of language proficiency. Finally, our participants' estimation of their English proficiency confirms the fact that SVT scores are sensitive to language proficiency, as should be reading scores. Scores on our test show a correlation of r = .43 (p <.001, N = 161) with self-assessed language proficiency. Reading tests typically yield such a significant but moderate correlation with language proficiency, since strong relations are observed in L1 "between the size of [participants'] vocabularies and their reading comprehension scores" [...] For second language learners, this relation appears to be even

Figure 1 ■ Infit t statistics: Pathway Map of SVT scores



Note: Figure computed using the eMr package for R (Mair & al., 2010)

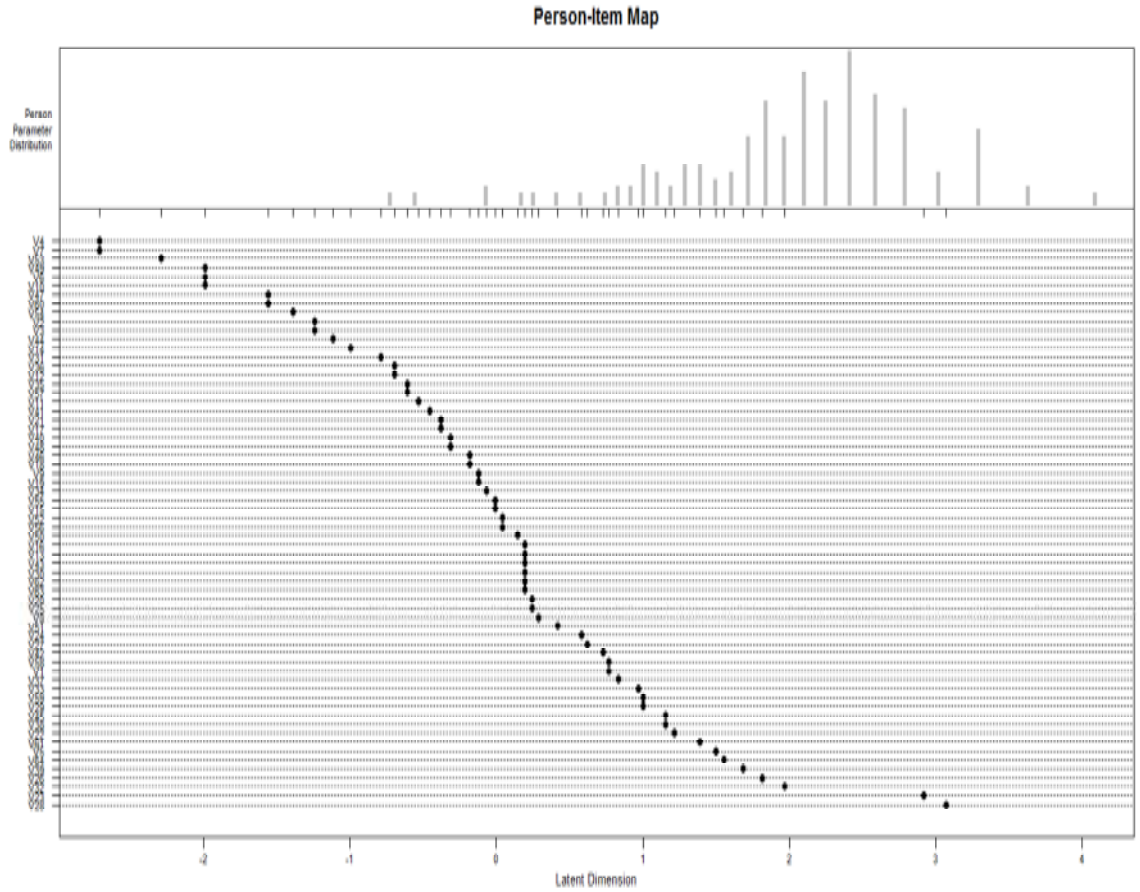
stronger." (Droop & Verhoeven, 2011: 81).

Rasch-Based Analysis

Testing the assumptions behind the Rasch model. Three tests of hypothesis were used to verify if the data violated the assumptions that underlie the Rasch model. Simply speaking, the main idea in using goodness-of-fit tests is to evaluate the null hypothesis H_0 that the data fit (i.e., if Rasch's homogeneity assumptions hold) against the alternative H_1 that they do not. Here, we make use of three well-known tests: Rost Deviance, Casewise Deviance, and Collapsed Deviance. First, the Rost's Deviance test (Mair, Reise, & Bentler, 2008) uses a saturated model to explain thoroughly and perfectly (with $p = 1$) the data matrix to analyze. This model can then become a reference for testing many other models of interest (i.e., 2PL or 3PL models). Second, Casewise Deviance is a widely used method for analysing the data predicted by the fitted model and those predicted by a "perfect model" (i.e., the most complete model to fit the data under investigation). Third, Collapsed Deviance is a test based on likelihood ratio test procedures. For these three methods, rejecting the null hypothesis H_0 would imply that Rasch may not be the best model to fit our data. Corresponding p-values are used to determine the level of compatibility of the data for H_0 (for a comparison of these tests and their equations, see Mair, Reise, & Bentler, 2008).

Results from the Rost Deviance test (6275.944, $p = 1.00$) and the Casewise Deviance test (7658.812, $p = 1.00$) strongly suggest that the Rasch model assumptions are met. However, the Collapsed Deviance test (1833.939, $p = 0.038$) tends to suggest otherwise. Therefore, we consider that the Rasch model assumptions are generally met, since two of the three

Figure 2 ■ Person-item Map



Note: Figure computed using the eMr package for R (Mair & al., 2010)

tests we used support that claim.

To further investigate that point, we calculated the percentage of variance accounted for by the Rasch model. In this case, we obtained a Pearson R of 0.22 and a McFadden R of 0.30. These figures are within an acceptable range, suggesting that 22% and 30% of the variance is explained by the model. This percentage of variance explained for SVT tests is comparable to those of other well-known tests such as the NSF survey (four categories) and the CAT test (Linacre, 2008).

Goodness of fit. The results above allow us to use the Rasch model to validate our SVT test. A deeper analysis of the items' Goodness of fit is encouraging: less than 10% of the items are considered to be potentially problematic for an analysis using the Rasch model.

Furthermore, the location of each item or each person against its infit *t* statistic is shown on a scatter plot in Figure 1 below, where the majority of items are within the acceptable range of -2 to 2 infit *t* statistics. Only 8 scores out of 64 are considered outliers. Nevertheless, items 24 and 52 are borderline cases for rejection. Such outliers are problematic, as they can reflect either type I or Type II errors. Items 5, 22, 23, 38, 39 and 64 are convincing outliers, showing high *t* statistics that leave little chance for Type 1 errors.

Estimation of parameters. Mair, Hatzinger, and Maier's (2010) R package allowed us to obtain a Conditional log-likelihood of -3513.986 after 91 iterations. A mean difficulty of -0.01 (S.D.=1.28) was also obtained for all 64 items of the test.

Figure 2 displays the distribution of items according to their estimated degree of difficulty on a logit scale. The plot shows a majority of points, which correspond to the participants' estimated ability, located between -1 and 1 logits. These results indicate that our participants obtained relatively high scores on the SVT test. Furthermore, the figure shows that the estimated ability of our participants tends to be higher than the estimated difficulty for items.

Discussion and conclusion

Items 23 and 28, which consist of two meaning changes,

were markedly more difficult than the rest of the items. As for the reason why these items may prove to be more difficult, one could argue that the meaning change for item 23 (replacing the word *evolution* by *amphibians*) may not have been significant enough. Item 28 seems to represent a valuable item, since the change (replacing the word *submerge* by *float*) not only affects the meaning of the sentence, but renders it illogical, given that it is air-filled lungs that make animals float. Such an item seems to tap the general understanding of the text. In the case of Item 32, close to the limit, it is surprising to see an intact sentence as being among the three most difficult items of the whole test. It could be hypothesized that, faced with a sentence containing a series of data, many test-takers could assume that any of those could have undergone an unnoticeable change. This item argues for the above-mentioned guideline that consists of selecting text that contains as little data as possible.

One could argue that our new instrument may prove too easy, given the overall mean of 83.7% which is slightly above the ideal range of 70 to 80% (Royer, 2004). However, we may consider the fact that our participants were university students whose high education suggests they have above-average reading skills. We could expect the scores to fall within the ideal range when it comes to testing younger and/or non-university participants. These means allow us to consider our instrument to be adequate for testing people of various reading abilities, and whose competence in English is sufficient for the reading of short texts (A2 and above on the CEFR). Testing a non-university population that is more representative of the general public will yield a more heterogeneous sample and is considered for future steps in this research project.

For tests that require the selection of an answer, such as multiple choice tests, it is always difficult for researchers to determine whether incorrect answer choices stem from guessing, or whether participants actually believed their choice was correct. It is assumed that guessing played a limited role in our test, since mean scores were significantly above simulated chance scores for the test as a whole, as well as for each of the four texts. In all cases, t-tests yielded extremely low p values (see also Pichette, de Serres, & Lafontaine, in press). A modified version of the test, relying on pseudo guessing, with the same titles and items but without the texts, yields scores that were equal to chance (see Pichette, Béland, & Raîche, in press).

Few items seemed either too easy or too difficult, but a limitation of this study is its small size sample. A higher number of participants will be needed to confirm the need to increase the difficulty of certain items. A larger participant sample will also make it possible to calculate estimates using two- and three-parameter models in light of Item response theory. However, it must be kept in mind that the format requirements (ratio for item types, and the nature of the sentences from which they are built) render such modifications difficult to make. Changing a 'weak' item from a paraphrase to a meaning change will force us to transform another meaning change item into a paraphrase in order to keep the same recommended ratio, and the latter item may not necessarily lend itself to such a change, and/or may even lose in quality and difficulty.

The Rasch model that we used includes one item parameter, which is item difficulty. Other existing models consider additional item parameters, such as 2PL models, which include item difficulty and item discrimination, and 3PL models, which include item difficulty, item discrimination, and pseudo guessing. It will be interesting to use the item parameters that are not estimated by the Rasch model to investigate differential item functioning and person-fit statistics. In addition, a comparison between the item parameters yielded for this paper-and-pencil version and a computerized online version will be of great interest.

A wide field of investigation in statistics is devoted to the various ways to treat missing data. In this study we considered missing data to be incorrect answers and replaced them with a zero in our matrices. This method is the most common way to deal with missing data, although some other methods have been put forward (Allison, 2001; Little & Rubin, 1987; Shafer & Graham, 2002). It would be interesting to perform the same analyses by treating missing data through likelihood estimation and multiple imputation, which seem to be two promising approaches.

This new instrument seems promising as an assessment method for second-language reading comprehension ability. First, it represents a valuable addition to teachers' resources for assessing reading ability. As experienced in our studies, it is administered swiftly and it does not require that the participants write down short answers or elaborate on the meaning of given parts of the text. Thus, it is simple to mark, and scores are easy to sum up. For teachers, it also draws on reading dimensions that are to be taught: what can

be understood from a text (same sentence), what could be inferred, correctly (paraphrases) and incorrectly (meaning changes). Increased research interest in this technique shall lead to investigations on a variety of text types and structures, as well as on numerous topics. Research outcomes would not only find applications to testing *per se*, but could emerge in the form of pedagogical applications with SVT as a potential tool for improving text comprehension in areas where students could show weaknesses. The sentence verification technique, after a long period of limited notoriety in the field of education, may end up helping teachers revisit reading comprehension and prove a useful tool for their practice.

References

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage.

Baddeley, A. D., Logie, R.H., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24*, 119-131.

Barrio Cantalejo, I., & Simón Lorda, P. (2003). Measurement of the legibility of written texts. Correlation between the Flesch manual method and computer methods. *Atención primaria, 31*(2), 104-108.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning, 39* (3), 313-338.

Council of Europe (2011). *Common European Framework of Reference for Language: Learning, Teaching, Assessment*. Council of Europe.

De Jonge, P., & de Jong, P. F. (1996). Working memory, intelligence, and reading ability in children. *Personality and Individual Differences, 21*(6), 1007-1020.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement, 41*, 261-270.

Droop, M., & Verhoeven, L. (2011). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly, 38*(1), 78-103.

DuBay, W. H. (2004). The principles of readability. Costa Mesa, CA: Impact Information. Retrieved from http://www.impact-information.com/impactinfo/readability02.pdf

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.

LeBlanc, R., & Painchaud, G. (1985). Self-Assessment as a Second Language Placement Instrument. *TESOL Quarterly,* 19(4), 673-687.

Linacre, J. M. (2008). Variance in Data Explained by Rasch Measures. *Rasch Measurement Transactions, 22*(1), 1164.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Mair, P., Hatzinger, H., & Maier, M. (2010). eRm: Extended Rasch Modeling. R package version 0.13-0. Retrieved from http://CRAN.R-project.org/package=eRm.

Mair, P., Reise, S. P., & Bentler, P. M. (2008). *IRT goodness-of-fit using approaches from logistic regression*. UC Los Angeles: Department of Statistics.

Metametrics (2009). *The Lexile Framework for Reading*. Retrieved from http://www.lexile.com.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language and education: Vol. 7: Language testing and* assessment (pp. 175-187). Dordrecht, The Netherlands: Kluwer.

Pichette, F., Béland, S., Magis, D., & Raîche, G. (2010). *From language assessment to research testing: Lz as a promising method for data elimination*. Conference « LTIME Language Teaching in Increasingly Multilingual Environments: From Research to Practice ». Warsaw, Poland, September.

Pichette, F., Béland, S., & Raîche, G. (in press). Application de l'indice lz pour l'élimination de données de recherche en langues. In G. Raîche, N. Loye and H. Meunier (Eds.), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation*. Volume 4. Québec : Presses de l'Université du Québec.

Pichette, F., de Serres, L., & Lafontaine, M. (in press). Élaboration d'une mesure de la compréhension de textes en anglais. *Revue pour la Recherche en Éducation, 3*.

Pichette, F., Segalowitz, N., & Connors, K. (2003), Impact of Maintaining L1 Reading Skills on L2 Reading Skill Development in Adults: Evidence from Speakers of Serbo-Croatian Learning French. *The Modern*

Language Journal, 87, 391–403.

Raîche, G. (2002). *Le dépistage de sous-classement aux tests de classement en anglais, langue seconde, au collégial* [The detection of under-performance to college aptitude tests of English, as a second language]. Gatineau, Québec, Canada: Collège de l'Outaouais.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1- 20.

Rousseau, M. (2006). *L'impact des méthodes de traitement des valeurs manquantes sur les qualités psychométriques d'échelles de mesure de type Likert.* Unpublished Ph.D. thesis , Université Laval.

Royer, J. M. (2004). *Uses for the sentence verification technique for measuring language comprehension*. Amherst, Massachusetts: Reading Success Lab.

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior, 11*, 355-363.

Royer, J. M., Lynch, D. J., Hambleton, R. K., & Bulgareli, C. (1984). Using the sentence verification technique to assess the comprehension of technical text as a function of level of expertise. *American Educational Research Journal, 21*, 839-869.

Schafer. J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schinka, J. A., & Borum, R. (1993). Readability of adult psychopathology inventories. *Psychological Assessment, 5*(3), 384–386.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99-128.

Strauss, U., Grzybek, P., & Altmann, G. (2007). Word length and word frequency. In P. Grzybeck (ed.), *Contributions to the Science of Text and Language Word Length Studies and Related Issues*. Dordrecht, Netherlands: Springer.

TAKS (2004). Grade 9 Reading TAKS Information Booklet. Retrieved from http://ritter.tea.state.tx.us/student.assessment/taks/booklets/reading/g9.pdf.

TAKS (2006). *The Texas Assessment of Knowledge and Skills. Texas Education* Agency. Retrieved from http://ritter.tea.state.tx.us/student.assessment/resources/online/2006/grade9/read/9reading.htm.

Wagenaar, W. A., Schreuder, R., & Wijlhuizen, G. J. (1987). Readability of instructional text, written for the general public. *Applied Cognitive Psychology, 1*(3), 155-167.

Wilson, K. M., & Lindsey, R. (1999). *Validity of Global Self-Ratings of ESL Speaking Proficiency Based on an FSI/ILR-Referenced Scale*. Princeton, NJ: ETS Research Report RR-99-13.

Yamashita, J. (2001). Transfer of L1 reading ability to L2 reading: An elaboration of the linguistic threshold. *Studies in Language and Culture, 23*, 189-200.

Yoshizawa, K. (2009). To what extent can self-assessment of language skills predict language proficiency of EFL learners in school context in Japan? *Journal of Foreign Language Education and Research, 17*, 65-82.

Appendix

SVT sample : Text and related items

Lisez l'histoire suivante lentement et attentivement, une seule fois, en vous concentrant.

A special volunteer

Barkley, our dog, came to me when he was three years old after living with a family that could no longer take care of him.

I took him to visit the school for the blind where I worked as a teacher. He would walk over to the children and wait for a child to pet him.

One day, he started bumping into the walls of our house. When we played ball in the yard, I

noticed that he could not catch it. I took him to the veterinarian, who found that he had an eye illness. Barkley had to have several operations. He soon learned to function with his weak eyes.

When he got better, he stood at the door, blocking my way, trying to tell me that he wanted to go to school with me and visit his friends. I started taking him to school again. Everyone was happy. Barkley was the happiest of all.

UNE FOIS TERMINÉ, TOURNEZ LA PAGE ET RÉPONDEZ AUX QUESTIONS.
NE REVENEZ PAS À L'HISTOIRE

Lisez attentivement chacune des phrases suivantes, dont l'ordre peut être différent de celui du texte.
-Écrivez "YES" si la phrase lue signifie la même chose que dans le texte.
-Écrivez "NO" la phrase a un sens différent du texte, ou si cela n'a pas été dit explicitement dans le texte.
Les mots n'ont pas à être les mêmes.

1. I took him to visit the old-age home where I worked as a nurse.
2. I received Barkley, our dog, when he was three years old after he lived with people who could not take care of him anymore.
3. The dog would go near the children and wait to be petted.
4. Barkley was so happy that he pulled the leash on the way to school.
5. I took him to the veterinarian, who found that he had an eye illness.
6. One day, he started falling.
7. Barkley was not happy after the operation because he could not see his friends.
8. When we played outside, I realized that he could not catch the ball.
9. He never learned to cope with his sick eyes.
10. The vet had to operate on Barkley several times.
11. That dog was intelligent and eager to please.
12. When he got better, he stood at the door, blocking my way, trying to tell me that he wanted to go to school with me and visit his friends.
13. I started playing with him at school again.
14. Barkley was always very affectionate to the family with whom he lived.
15. The blind children were happiest of all.
16. All the kids were joyful.

Citation

Pichette, F., Béland, S., de Serres, L., & Lafontaine, M. (2014). Validation of the Sentence Verification Technique to assess L2 reading comprehension ability. *The Quantitative Methods for Psychology*, *10* (2), 95-106.

Received: 29/12/13 ~ Accepted: 12/02/14