International Journal of
Linguistics
(IJL)

Assessing Comprehension beyond Recognition: A Many-Facet Rasch Measurement Approach

Bartolo Bazan

International Journal of Linguistics
ISSN 2710-4788 (online)
Vol.7, Issue 1. No 1 Pp 1 - 20, 2026

www.iprjb.org

Assessing Comprehension beyond Recognition: A Many-Facet Rasch Measurement Approach

[1*]Bartolo Bazan
English, Ryukoku University Heian Junior & Senior High School, Japan

Article History

Received 13th November 2025

Received in Revised Form 16th December 2025

Accepted 19th January 2026

How to cite in APA format:

Bazan, B. (2026). Assessing Comprehension beyond Recognition: A Many-Facet Rasch Measurement Approach. *International Journal of Linguistics*, *7*(1), 1–20. https://doi.org/10.47604/ijl.3600

Abstract

Purpose: Comprehension is commonly assessed through single-task tests, particularly multiple-choice questions (MCQs). Although MCQs offer many advantages, a growing number of researchers have raised concerns that such measures may overestimate degree of understanding because they are based on recognition. A less-frequently used method are summaries, which are assumed to reflect a higher level of comprehension because they require learners to select and integrate information in order to create a mental representation of the input. This study proposes a combined comprehension measure that integrates summaries (assessing global comprehension) and MCQs (targeting detail-level comprehension) into a single measurement system. The purpose of the study is to collect validity evidence for the use of the combined measure through many-facet Rasch measurement (MFRM).

Methodology: Listening data were longitudinally collected from 290 EFL Japanese high school students over three separate waves, with each involving three measurement points (i.e., Pretest, Posttest 1, and Posttest 2). Comprehension was assessed twice at each measurement point through the combined measure consisting of a summary and a set of five MCQs, which were administered in paper-and-pencil format. The summaries were rated by two expert raters using a five-point rating scale in tandem with a list of main ideas and details that had been previously extracted from the target texts. The MCQs were dichotomously scored. All three waves of data were linked through a Rasch stacking design and were analyzed using MFRM with the analysis involving three facets: persons, items, and raters. Under the theoretical assumption that summaries are more difficult and entail a higher level of comprehension than MCQs, the summaries were given double weight when estimating learner ability.

Findings: The Wright map confirmed a hierarchy of item difficulty consistent with the theoretical expectation that the summaries were more difficult than the MCQs, providing support for the weighting scheme. The persons showed acceptable fit to the Rasch model with most participants being within parameters. Similarly, all summaries and multiple-choice items fit the Rasch model's expectations with the exception of only two multiple-choice items, which were slightly above the recommended criteria. The analysis revealed fair person reliability and excellent item reliability, suggesting that the replicability of the person ability and the item difficulty hierarchies were fair and high, respectively. In addition, rater severity did not negatively impact the measurement and the response thresholds suggested that the rating scale functioned as intended. These findings indicate that the combined measure is a valid instrument for comprehension assessment.

Unique Contribution to Theory, Practice and Policy: Practically, this study contributes to the literature by providing a combined measure that mitigates the weaknesses of summaries and MCQs when used separately. In addition, it demonstrates how MFRM can model productive and receptive tasks, which may be differently weighted tasks, within a single measurement system. Regarding policy, this study advocates for tests that move beyond single tasks to provide a more precise picture of learners' levels of comprehension across different educational settings.

Keywords: *Rasch Model*, *Many-Facet Rasch Measurement*, *Comprehension Assessment*, *Summaries*, *Multiple-Choice Questions*

INTRODUCTION

Assessing comprehension has traditionally relied on single-format instruments, most notably multiple-choice questions (MCQs). MCQs have been the most widely used assessment tool because a) they can be scored quickly (Bennett et al., 1991) or automatically using an optic reader or a computer (Lee et al., 2011), b) they are reliable and affordable (Rauch & Hartig, 2010), c) large groups of test-takers can easily be assessed in one sitting (Klufa, 2015), and d) they avoid the possibility of scorer bias (Zhou, 2019). However, there is a growing body of researchers who question the extent to which MCQs tap into the construct of comprehension (Chan & Kennedy, 2002; Hughes, 2000; Little et al., 2012, Martinez, 1999; Mee et al., 2024). These researchers have particularly criticized MCQs on the grounds that, given their receptive nature, they may represent a shallow level of comprehension (i.e., a minimal surface-level representation of the meaning of a text) because they emphasize recognition over coherent constructions of meaning. As a result, MCQs may overestimate learners' degree of comprehension.

In contrast, productive comprehension tasks, such as summaries, require learners to actively form a mental representation of the meaning and message of a target text by selecting and organizing relevant information, whether the input be auditory or written. Therefore, summaries should reflect a higher level of comprehension, particularly regarding global comprehension (i.e., a rich representation of the meaning of the text constructed by organizing and integrating information).

A key limitation of much of the existing research is the tendency to treat receptive (e.g., MCQs) and productive (e.g., summaries) tasks as separate or mutually exclusive assessment alternatives. In the present study, I address this shortcoming by proposing a hybrid approach to comprehension assessment that combines summaries and MCQs into a single measurement system. Summaries require meaning construction through the integration of information and are thus hypothesized to capture global comprehension. For this reason, they are complemented by a set of MCQs as they are mostly designed to target details. The theory suggests that details are less likely to be encoded as learners tend to focus on important information (Kintsch, 1998), which can consequently be targeted by the MCQs.

To operationalize the measure, I assigned double weight to the summaries for three reasons. First, if listening comprehension is defined as the ability to construct a coherent mental representation of the input (Kinstch, 1998), that coherent mental representation is more likely to be reflected in a summary that is entirely produced by learners than in a set of MCQs designed by the test designer. Second, writing a summary is more difficult than answering MCQs, and higher difficulty should predict higher ability (Bond & Fox, 2015). Finally, although summaries should represent a more holistic and complete record of listening comprehension, they cannot be relied upon completely because they require productive skills and learners can produce blank summaries. Therefore, to reflect these points in the overall comprehension measure, the summaries are complemented with MCQs but are accorded extra weight in the analysis when estimating student ability. It is important to note that weighting is not a Rasch assumption but rather a modeling decision. With this weighting scheme, the data of the combined measure, which came from a study I conducted in Bazan (2024), were analyzed through many-facet Rasch measurement (MFRM) (Eckes, 2015; Linacre, 1994) in order to evaluate its psychometric properties and collect validity evidence for its use.

Statement of the Problem

Accurate assessment of comprehension is essential to education and research. Comprehension measures are used for educational decisions such as classroom placement or university admission. Regarding research, if comprehension is assessed imprecisely, measures might reflect artifacts of the measurement tool rather than true ability, potentially resulting in misleading findings. A long-standing problem in comprehension measurement is that comprehension is commonly assessed through MCQs, which are based on recognition skills and thus may reflect a shallow level of understanding. Summaries, in contrast, require learners to create a mental representation of the input by reconstructing meaning, reflecting higher levels of comprehension. A central gap in the literature is the lack of valid assessment tools, particularly measures that combine productive (i.e., summaries) and receptive (i.e., MCQs) as complementary into a single measurement system. In this hybrid approach, summaries are hypothesized to tap into global comprehension and MCQs into detail-level comprehension.

The findings of this study can benefit several stakeholders. For researchers, the study provides a novel approach for measuring comprehension through a combination of assessment tools, which are weighted differently based on theory. For educators and test developers, the study offers a more accurate and valid assessment of comprehension than single-task tests. For learners, this new approach can yield fairer estimations of ability by integrating global and detail-level comprehension into a single score.

LITERATURE REVIEW

Summaries as Comprehension Measures in Second Language (L2)

Summaries have long been used as an alternative method to MCQs in the assessment of L2 comprehension. This choice is grounded on empirical research. In a study comparing summaries against free recall protocols (i.e., writing down everything that is remembered from a text) as measures of global comprehension in reading, Riley and Lee (1996) had 80 early-stage L2 readers of French read a passage. Subsequently, half of the sample wrote a summary and the other half a free recall protocol, which were both analyzed for total number of main ideas and details. A two-way ANOVA showed that the summaries contained a higher number of main ideas than the free recall protocols. In addition, the summaries were found to have a larger percentage of main ideas and details. This finding supports the use of MCQs as tapping into detail-level comprehension in the combined context of this study.

A central methodological question for L2 comprehension assessment concerns the language in which the summaries should be produced: in the learners' first language (L1) or L2. However, relatively little research has been conducted in this area and to the best of my knowledge, no research has directly compared the two modalities. Much of the current research has been dominated by L2 summaries. For example, Qin and Groombridge (2023) sought to investigate the factors that impact L2 summarization. The participants were 46 Emirati university students, who were asked to write a summary of between 100 and 150 words after having read a 745-word text on the topic of *consumerism*. The summaries were given a global score using the prescribed university 4-point rubric as well as a content score using another rubric designed for the purpose of the study that contained the main ideas of the text and assessed how many main ideas were included in the summaries. A correlation analysis between the global and content scores and the participants' scores on the IELTs revealed a significant relation between the IELTS reading scores and both the global ($r = .55, p = .000$) and content summary

($r = .35$, $p = .009$) scores. Interestingly, the scores on the writing section of the IELTS significantly correlated with the global score ($r = .44$, $p = .001$) but not with the content score ($r = .24$, $p = .07$), which was based on the main ideas in the text. In addition, overall language proficiency as assessed by the IELTS was found to significantly correlate with the summaries' global scores ($r = .60$, $p = .001$), but no significant correlation was found between overall proficiency and the content scores ($r = .25$, $p = .07$). The researchers concluded that writing a summary in an L2 is a challenging task as it requires a combination of effective reading, writing, as well as a good level of language proficiency.

A similar conclusion was drawn by Alaofi (2020) from a qualitative perspective. He interviewed nine Saudi university students on an individual basis using semi-structured interviews. The interviews, which lasted between 15 and 26 minutes, were conducted in Arabic and were audio recorded for transcription purposes. A thematic analysis of the data revealed a number of challenges that learners encounter when summarizing such as language proficiency, writing style, and reading comprehension.

In sum, this line of research suggests that L2 summaries may not be an optimal measure of comprehension due to the potential confounding effect of L2 writing ability. For this reason, the participants in this study were required to write their summaries in their L1, Japanese.

Comprehension as Assessed through Summaries and MCQs

There is scant research that has investigated comprehension using both summaries and MCQs. Research conducted from this double-format assessment paradigm has usually treated summaries and MCQs as separate comprehension measures. One such study was conducted by Chen et al. (2014), who investigated the effects of mode of reading (i.e., reading on tablets, computers, or paper) on comprehension. Ninety Chinese college students were randomly assigned to a tablet, a computer, or a paper condition and were asked to read four texts extracted from the national college entrance examination in their corresponding mode. The texts were between 1,050 and 1,099 Chinese characters with the first text being used as a practice trial and the other three as formal tests. The participants completed the entire research procedure in a single session of 60 minutes. After reading each text, the participants answered a set of five MCQs and subsequently wrote a summary with both tasks given in a paper-based format. This order of task administration is a limitation of this study because MCQs tend to follow the same chronological sequence of events as is found in the target text, which can help learners construct a mental representation of the text even in cases of partial comprehension. In other words, the information contained in the MCQs can help learners write more detailed summaries than they would be able to write if they relied only on their text understanding. In any case, the summaries, which were rated by two experts using a rating scale and averaged into a single score, were assumed to represent deep comprehension and the MCQs shallow comprehension. A one-way ANOVA revealed no statistical differences among the groups in the summaries (i.e., deep comprehension) with the paper group outperforming the computer group on the MCQs (i.e., shallow comprehension).

In a second language context, Oded and Walters (2001) set out to examine whether higher levels of processing result in deeper comprehension. Sixty-five Israeli learners studying English as part of their majors were divided into two groups and were asked to read two texts under one of the following experimental conditions: a summary or an example condition. The summary condition involved writing a summary of the target text in English and the example

condition involved writing a list of examples given in the text to support the author's argument. It was hypothesized that the summaries would elicit higher levels of processing and hence deeper comprehension in comparison to the list of examples. After writing the summary or the list of examples, comprehension was assessed by a set of seven MCQs. Paradoxically, the summaries were overlooked as measures of comprehension. The results of a *t*-test suggested that the summary condition led to deeper comprehension as theorized. Overall, these studies attest to the methodological tendency to separate summaries from MCQs as measures of comprehension.

Rasch Modeling of Multi-Faceted Data

The Rasch model (Rasch, 1960) is a psychometric measurement model against which empirical data can be compared to assess the validity of an instrument. It is based on the assumption that a construct is represented by the relationship between the ability of the persons and the difficulty of the items in a sample. Many-facet Rasch measurement (MFRM) is an extension of the Rasch model that allows for the incorporation of variables or facets beyond persons and items, such as raters evaluating performance on the basis of a rating scale. In the context of this study, where the summaries were assessed by two raters in terms of the comprehension criteria specified by the five-point rating scale (i.e., four thresholds), the assigned level of comprehension is a function of the ability of the participants, the difficulty of the text, the difficulty of the threshold (e.g., the threshold of being awarded a 3 over a 2), and the severity of the raters.

The Rasch model provides several benefits over traditional analytic techniques. First, the Rasch model transforms nonlinear raw scores, where differences between consecutive data points do not represent equal amounts of the construct, into equal interval data (Bond & Fox, 2015). The unit of measurement that results from this transformation is expressed in logits (i.e., logarithm odd units), which are based on probabilistic calculations that take into account a person's ability, item difficulty, and other facets potentially having an impact on the test outcome. Second, the Rasch model provides detailed information about different aspects of validity, which are described in the following subsections. MFRM is particularly suitable for this study because it can accommodate the different facets involved (i.e., summaries, MCQs, and raters) under a common measurement framework while allowing for the assignment of extra weight to the summaries.

Wright Map

MFRM produces a variable map, also known as a Wright map (Bond & Fox, 2015), that graphically displays the relationship of the test facets on a single equal interval logit scale. The Wright map is useful for examining the difficulty hierarchy of items along a measured construct, which can reveal if the construct has been operationalized as intended. Specifically, if the items hypothesized to be more difficult when designing the test are indeed more difficult. In the context of this study, the summaries should be displayed above the MCQs on the map as they are hypothesized to be more difficult. The Wright map is also useful to visually inspect the degree of severity exercised by multiple raters in contexts of rater-mediated assessment performance. Large between-rater distances indicate that there is great variance in rater severity whereas smaller distances suggest a closer degree of severity when applying the rating criteria (Linacre, 2023).

Person and Item Fit

MFRM compares observed responses against a theoretical ideal model, which is mathematically represented by a straight line (Bond & Fox, 2015). The degree to which item and person performance conforms to the model's theoretical expectations is reported as item and person fit indices. Sufficient fit to the Rasch model indicates that the items on the test assess the same underlying construct. MFRM provides two mean-square (MNSQ) fit statistics: infit and outfit MNSQ. Infit MNSQ is a weighted unstandardized form of fit that is sensitive to unexpected performances by participants who fail items that have a difficulty value that is close to their ability level. In contrast, outfit MNSQ is a non-weighted standardized fit statistic that is sensitive to outliers (e.g., participants who manage to succeed on items above their abilities). Because the calculation of the infit MNSQ statistic involves giving more weight to the performances of participants whose ability level is near the item difficulty level, infit MNSQ provides more insightful information about item and person performance than outfit MNSQ. Consequently, infit MNSQ is usually the statistic that guides the evaluation of fit (Bond & Fox, 2015). As such, in this investigation, decisions about fit were made based on infit MNSQ, but problematic outfit MNSQ values were also explored to investigate unexpected performances of items and persons. The Rasch model's expected value of infit and outfit MNSQ is 1.00. Of the several guidelines that have been proposed based on the ideal 1.00, I adopted the generic range of 0.50 and 1.50 recommended by Wright et al. (1994) and Linacre (2007) when making decisions about fit for the summaries and the stricter 0.70–1.30 range suggested by Wright et al. for the MCQs. There are two important points to note here. First, although values above the adopted guidelines flag misfit, values up to 2.00 do not degrade measurement (Linacre, 2007) and, for this reason, values found to be slightly above or below the criteria were accepted as tolerable (Wright et al., 1994). Second, although these criteria apply equally to infit and outfit MNSQ values of both items and persons, their application to persons tends to be less strict than to items because persons are expected to behave less consistently than items (Linacre, 2007; Wright et al., 1994). Therefore, the fit criteria were applied more flexibly to persons than to items.

Person and Item Reliability and Separation

The Rasch person reliability index indicates the degree to which replicability of the person hierarchy is possible if the sample were given a similar test measuring the same underlying construct (Bond & Fox, 2015). In contrast, the Rasch item reliability index indicates the degree to which the replicability of item location along the measured construct is possible were the test administered to a similar sample. In this study, the reliability estimates were interpreted following the guidelines proposed by Fisher (2007), which state that values below .67 indicate poor reliability, values between .67 and .80 indicate fair reliability, those between .81 and .90 indicate good reliability, those between .91 and .94 indicate very good reliability, and values above .94 indicate excellent reliability.

Together with the reliability estimates, Rasch analysis provides person and item separation indices. The person separation index estimates the number of ability levels into which a sample of participants can be reliably separated by a measure. In contrast,

the item separation index estimates the number of difficulty levels into which the items can be reliably separated by a sample. According to Duncan et al.'s (2003) guidelines for person separation, an index of 1.50 represents an acceptable separation, an index of 2.00 represents

good separation, and an index of 3.00 represents excellent separation. With respect to item separation, Linacre (2007) recommended indices of 3.00 or above.

Rater Severity

Four diagnostics were used to evaluate rater severity (Linacre, 2023): the rater column in the Wright map as explained above, rater severity measures, rater reliability and separation, and the ratios of observed and expected agreement.

Rater Severity Measures

In addition to the visual inspection of the rater column on the Wright map, rater severity was explored numerically by looking at the rater Rasch measures. The larger the differences between the estimated measures, the wider the level of severity that the raters exercised. Statistically significant differences are revealed by a significant χ^2 value (Linacre, 2023).

Rater Reliability and Separation

The rater reliability and separation estimates indicate the level of replicability of the raters' severity measures if they rated a similar pool of essays and the degree to which their level of severity is statistically separable from one another (Sick, 2013). The larger the reliability index, the more variation there is between the level of severity exercised by the raters and conversely, the lower the reliability index, the higher the degree of rater agreement. Similarly, the higher the separation value, the larger the discrepancy between the level of severity of the raters and the lower the separation index, the closer the raters are in level of severity.

Ratios of Observed and Expected Agreement

The level of severity between raters is reflected in the ratios of observed and expected agreement (Sick, 2013) reported in the Facets output, which were also explored. Large differences between observed and expected agreement indicate inconsistent rater behavior, meaning that the raters are not severe or lenient in a consistent manner. Therefore, values that closely approximate to one another are desired (Linacre, 2023).

Rater Fit

Rater infit and outfit MNSQ statistics show the degree of consistency to which raters use a rating scale and thus, the extent to which they adhere to the Rasch model requirements. Although there are no definite guidelines to evaluate rater infit and outfit MNSQ values, I adhered to Wright et al. (1994) and Linacre's (2007) guidelines, which I adopted for the summaries.

Response Category Functioning

Because a rating scale is used in this study, its response category functioning must be examined. Four essential criteria have been proposed (Linacre, 2002). First, the scale should be oriented with the latent variable. Second, there should be at least 10 responses on each category function with higher scores representing higher measures on the latent variable. Next, because higher categories should reflect higher measures, the observed average measures of the persons in the category are expected to increase monotonically. Finally, Outfit MNSQ values should be below 2.00. In the context of this study, where the L1 summaries were assessed by two raters in terms of the comprehension criteria specified by the five-point rating scale with descriptors (i.e., 0 = *No comprehension*, 1 = *Minimal comprehension*, 2 = *Good comprehension*, 3 = *Very good comprehension*, and 4 = *Excellent comprehension*), MFRM produces four category response

thresholds (i.e.., the thresholds between the scores of 0 and 1, 1 and 2, 2 and 3, and 3 and 4), which must be analyzed.

Research Questions

Based on the different aspects of validity reviewed in the literature, the study addressed the following research questions (RQs):

1. Are the summaries more difficult than the multiple-choice items to support the weighting approach (i.e., giving double weight to the summaries)?
2. Do the persons fit the Rasch model?
3. Do the items fit the Rasch model?
4. Is the person reliability of the combined comprehension measure sufficient to suggest a similar spread of participants with higher and lower levels of comprehension if they were given a similar test containing summaries and MCQs?
5. Does the combined comprehension measure separate participants into different levels of comprehension?
6. Is the item reliability of the combined comprehension measure sufficient to suggest replicability of the item difficulty hierarchy if the test were administered to a sample of similar ability?
7. Does the sample of participants separate the items into different levels of difficulty?
8. Do the raters fit the Rasch model?
9. Do the raters exercise a similar level of severity when awarding ratings to the summaries?
10. Does the rating scale function adequately?

METHODOLOGY

Participants

The participants were 290 students (119 girls and 171 boys) learning English as a foreign language in a private high school in Western Japan, of whom 113 were first-years (aged 15-16 years old), 141 were second-years (aged 16-17 years old), and 36 were third-years (aged 17-18 years old). All participants were native speakers of Japanese, ranging in English proficiency from level A1 to B1 in the Common European Framework (Council of Europe, 2020), which indicates basic to low-intermediate proficiency. Ethical clearance was obtained and the study was conducted according to the guidelines of the institution for research.

Materials, Procedure and Design

Six listening texts were selected as the stimuli on which the summaries and MCQs were based. All texts were 400-word extracts of the Easy-starts graded readers from the Penguin collection and contained a narrow range of vocabulary (i.e., between 94.3% and 98.3% of the vocabulary came from the 1,000 most frequent words in English) and grammatical structures (i.e., 81.56% of the forms were in the simple present and 6.17% in the present progressive). The audio files were recorded at 160 words per minute in a private studio by a native English teacher in the school, a Canadian male. After listening to the stories, which were played once through a Bluetooth speaker connected to a Windows laptop, the participants wrote their summaries in their first language, Japanese, before they turned to the MCQs, which were written in English. Both the summary and the MCQs tasks were given as a handout before each listening. Looking

at the summary when answering the MCQs was prohibited as was returning to the summary when answering the MCQs.

The summaries were rated by two expert raters, who had previously gone through the texts in order to create a list of main ideas and important details for each text against which the summaries could be compared. They scored the data independently using a five-point rating scale in tandem with the list of main ideas and details developed for the study.

The study was a longitudinal study encompassing three waves of data collected from three different cohorts of students over three subsequent school years. There were three measurement points separated by two nine-session cycles of listening training. Thus, two of the six tests served as Pretest, two as Posttest 1, and another two as Posttest 2. The administration of the two tests at each timepoint was separated by one or two days but the data were combined as if they belonged to a single test each time (see Bazan, 2024 for a detailed explanation of the materials, procedure, and design).

Analysis

The three waves of data were linked through a Rasch stacking procedure and were analyzed using Facets version 3.80.0 (Linacre, 2017) giving double-weight to the summaries. It is important to note that in the original study (Bazan, 2024), the weighted analysis was performed to produce person measures for subsequent statistical analysis. However, because weighting tends to inflate fit and reliability indices, an analysis in which the ratings and the items were equally weighted is reported here. In this respect, it is noteworthy that the two independent raters rated a subset of the summaries, which corresponded to 70% of the total, obtaining both a separation and a reliability estimate of .00 and an 87.3% exact-agreement rate when the model expected 50.4%. These results indicated that the raters were not statistically different from each other and thus, the remaining 30% of the summaries were rated only by one rater (i.e., pseudonym of Barney) and the data from the second rater for this subset was entered in the analysis as missing. Moreover, the data of five participants who missed one of the two tests at one of the three timepoints were entered into the analysis as missing.

RESULTS

RQ1: Are the summaries more difficult than the multiple-choice items to support the weighting approach?

To answer RQ1, I examined the Wright map, which portrays the three facets in the analysis, as shown in Figure 1. The column on the far left shows the logit scale on which the person, item, and rater measures are positioned. The column next to the logit scale shows the persons with each star representing nine persons and a dot representing one or two persons. The persons appear in descending order of comprehension with higher-scoring participants located at the top and lower-scoring persons located at the bottom. As can be seen, the persons were widely distributed along the construct with an ability range from -4.00 to 6.00 logits. This spread of persons is large enough to demonstrate a hierarchy of ability on the measured construct. However, a number of participants are positioned above the most difficult item and others are below the easiest item, indicating a ceiling and a floor effect, respectively. In any case, the spread of participants relative to that of the items demonstrates that the targeting of the items was satisfactory overall.

The raters, with pseudonyms Barney and Simpsons, appear in the column to the right of that of the persons in descending order of rating severity. As illustrated by the figure, the variability between the raters in their degree of severity is almost non-existent because they are located next to each other around the zero origin of the scale, which suggests that they exercised the same level of middling severity when rating the L1 recall summaries. This finding indicates that both raters used the rating scale consistently and thus, rater severity did not impact the participants' ability estimates.

The column on the far right, which is the column of interest regarding RQ1, displays the items in descending order of difficulty. The dichotomous items (i.e., MCQs) are represented by the name of the test with the letters A to D followed by the question number, whereas the ratings are represented by an R for rating followed by the test name, A to D (the titles of the stories together with a detailed description of their profiles is presented in Bazan, 2024).

This instrument was developed under the assumption that the ratings or summaries are more difficult than the dichotomous items or MCQs. Thus, the ratings should appear above the dichotomous items in the map. As illustrated by the figure, the ordering of the items matches the initially predicted results based on the hypothesized higher difficulty of the ratings, with the six ratings clustering together at the top of the distribution. Only two (i.e., Items B5 and D5) of the 30 MCQs used in the study appear to have a similar level of difficulty to that of the summaries. This match between the empirical hierarchy of the items and the expected hierarchy is supportive of the construct validity of this listening measure.

International Journal of Linguistics
ISSN 2710-4788 (online)
Vol.7, Issue 1. No 1 Pp 1 - 20, 2026
www.iprjb.org

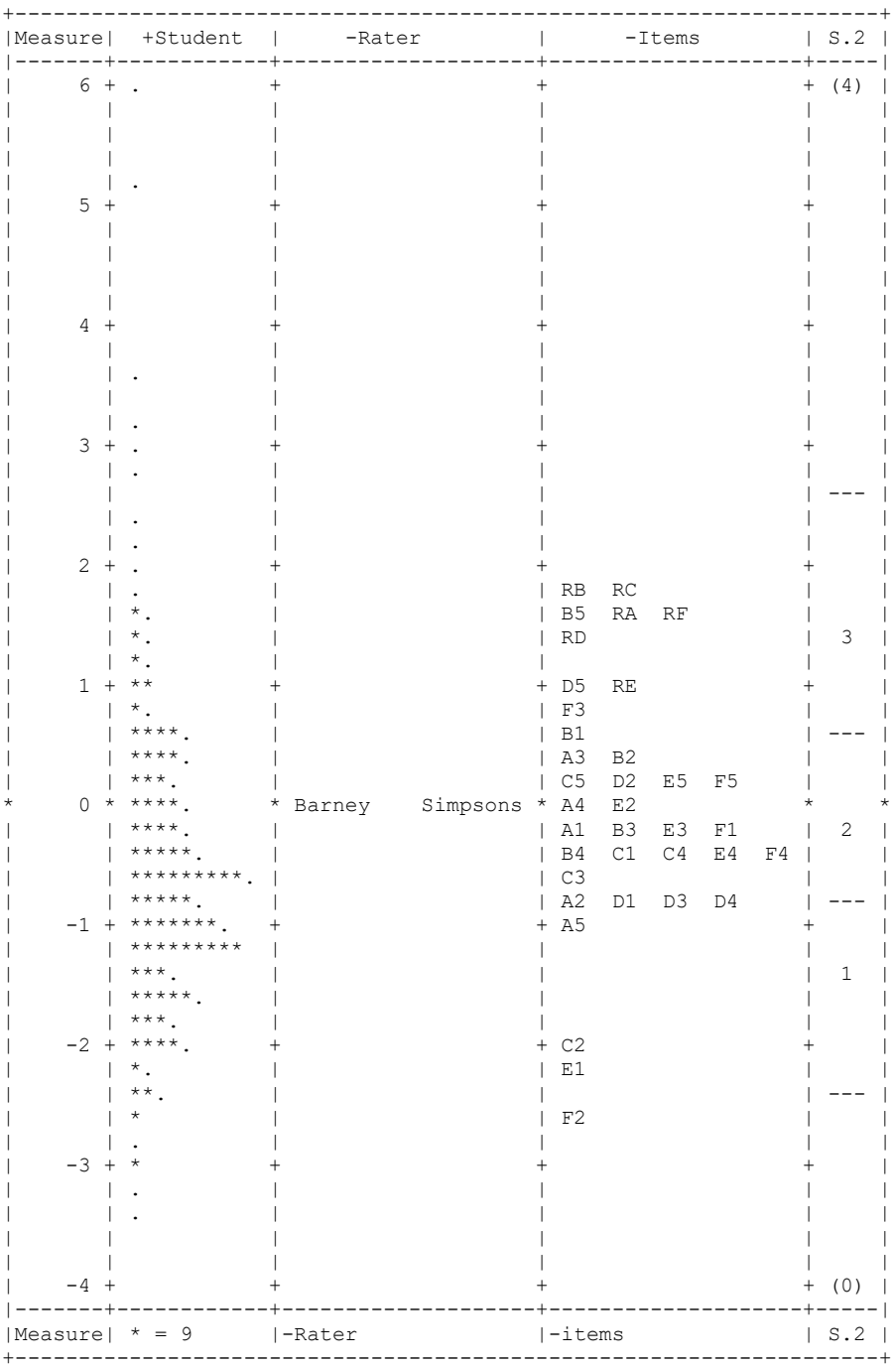

Figure 1: Wright Map for the Comprehension Measure. *N* = 290.

Items labeled R are summaries rated using a five-point rating scale and items labeled A-F are MCQs scored dichotomously according to an answer key. * = 9 participants, . = one or two participants. A = *Troy*, B = *Pete and the Pirates*, C = *Simon and the Spy*, D = *Maisie and the Dolphin*, E = *Dino's Day in London*, F = *Billy and the Queen*.

International Journal of Linguistics
ISSN 2710-4788 (online)
Vol.7, Issue 1. No 1 Pp 1 - 20, 2026
www.iprjb.org

RQ2: Do the Persons Fit the Rasch Model?

To answer RQ2, I conducted an inspection of the persons' infit and outfit MNSQ values, which are presented in percentages in Table 1. As can be seen, the majority of the sample had fit values within the acceptable 0.50 to 1.50 criterion. As illustrated by the table, 92% of the participants met the benchmarks with respect to infit MNSQ and only 1% of the participants were above the recommended criterion of 2.00. Similarly, 85% of the participants fit the Rasch model with regard to outfit MNSQ and only 5% of the sample displayed values above 2.00. These results demonstrated that the participants adhered sufficiently well to the Rasch model.

Table 1: Person Statistics for the Combined Comprehension Measure

		Criteria			
		Not degrading	**Within parameters**	**Not degrading**	**Degrading**
		0.00–0.49	0.50–1.50	1.51–1.99	>2.00
% of participants	Infit MNSQ	5%	92%	2%	1%
	Outfit MNSQ	2%	85%	8%	5%

Note. $N = 290$. All statistics are based on Rasch logits. MNSQ = mean-square.

RQ3: Do the Items Fit the Rasch Model?

To address RQ3, an examination of the item fit statistics was conducted. Table 2 presents the item fit statistics. As shown in the table, most of the dichotomous items (i.e., A1 to F5) fit the 0.70 to 1.30 criterion for the infit MNSQ statistic. The only exceptions were Items A4 and E5, which were slightly above the 1.30 upper boundary with values of 1.31 and 1.32 respectively. These values were not of concern for the measurement because they were within the range so as not to distort measurement (Linacre, 2017). Furthermore, all of the rated items (i.e., RA to RF) also fell within the 0.50 to 1.50 infit MNSQ criterion (range from 0.60 to 0.81) adopted for the summaries, with respect to infit MNSQ.

The outfit MNSQ statistics for the dichotomous items were similarly acceptable. Although five items, A4 (outfit MNSQ = 1.37), B5 (outfit MNSQ = 1.49), C1 (outfit MNSQ = 1.31), D5 (outfit MNSQ = 1.41), and E5 (outfit MNSQ = 1.54) showed slight underfit, these indices were unproblematic because they did not distort the measurement. Item F5, however, displayed a value of 2.90, which was above the 2.00 criterion recommended by Linacre (2007) for non-degrading items. To identify the source of the extreme outfit, I examined the unexpected responses of the item. This analysis revealed that the high outfit was caused by three participants who, despite having low estimated ability measures (-2.39, -2.39, and -2.47 logits, respectively), managed to succeed on the item (difficulty measure = 0.11 logits), which was above their ability level. Additionally, an inspection of the item did not suggest that editing the item or changing its wording should be considered. As for the outfit MNSQ of the rated items, the table shows that the items fit the criterion outlined above for performance-based tasks with ranges from 0.58 to 0.66. These results can be interpreted as evidence to support the fact that the items functioned well together.

International Journal of Linguistics
ISSN 2710-4788 (online)
Vol.7, Issue 1. No 1 Pp 1 - 20, 2026
www.iprjb.org

Table 2: Item Statistics for the Combined Comprehension Measure

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
A1	-0.12	0.14	1.14	2.10	1.14	1.20
A2	-0.76	0.13	0.93	-1.30	0.85	-1.30
A3	0.33	0.15	1.15	2.00	1.27	2.00
A4	0.08	0.14	1.31	4.30	1.37	2.90
A5	-1.02	0.14	1.02	0.40	1.12	0.90
B1	0.50	0.15	1.10	1.20	1.19	1.40
B2	0.37	0.15	1.08	1.10	1.13	1.10

Table 2. (continued)

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
B3	-0.21	0.13	0.98	-0.20	0.97	-0.30
B4	-0.32	0.13	1.09	1.60	1.10	1.20
B5	1.56	0.19	1.23	1.60	1.49	1.90
C1	-0.43	0.13	1.17	3.30	1.31	3.80
C2	-2.03	0.15	0.96	-0.50	0.93	-0.40
C3	-0.58	0.13	1.06	1.30	1.08	1.00
C4	-0.45	0.13	1.19	3.80	1.24	3.10
C5	0.24	0.14	1.15	2.30	1.19	1.80
D1	-0.86	0.13	1.01	0.10	1.02	0.20
D2	0.16	0.14	1.01	0.10	1.08	0.80
D3	-0.90	0.13	1.10	1.90	1.15	1.90
D4	-0.74	0.13	1.07	1.40	1.11	1.40
D5	1.04	0.16	1.13	1.30	1.41	2.30
E1	-2.21	0.16	0.95	-0.50	1.03	0.20
E2	0.06	0.14	1.27	4.20	1.36	3.60
E3	-0.15	0.13	1.28	4.60	1.39	4.30
E4	-0.34	0.13	0.93	-1.30	0.94	-0.70
E5	0.11	0.14	1.32	4.80	1.54	5.10
F1	-0.26	0.13	1.15	2.60	2.13	8.40
F2	-2.66	0.18	0.98	-0.10	0.92	-0.10
F3	0.71	0.15	1.13	1.60	1.22	1.50
F4	-0.44	0.13	1.14	2.60	1.61	4.90
F5	0.11	0.14	1.18	2.90	1.29	2.40
RA	1.69	0.07	0.81	-2.70	0.66	-4.40
RB	1.79	0.07	0.69	-4.60	0.60	-5.20
RC	1.74	0.07	0.61	-6.10	0.58	-5.80
RD	1.33	0.07	0.62	-6.50	0.59	-6.60
RE	1.08	0.06	0.60	-7.00	0.58	-7.00
RF	1.63	0.07	0.65	-5.60	0.61	-5.40

Note. $N = 290$. items are sorted in entry order. Items labeled R are summaries rated using a four-point rating scale and items labeled A-F are multiple-choice questions scored dichotomously according to an answer key. MNSQ = mean-square, ZSTD = Standardized z-scores.

RQ4: Is the person reliability of the combined comprehension measure sufficient to suggest a similar spread of participants with higher and lower levels of comprehension if they were given a similar test containing summaries and MCQs? and RQ5: Does the combined comprehension measure separate participants into different levels of comprehension?

The reliability of the person ability estimates was .72, which is considered to be fair in terms of the replicability of the person ability hierarchy (Fisher, 2007). This Rasch person reliability coefficient corresponded to a separation value of 1.62 which, according to Duncan et al.'s (2003) guidelines, is acceptable as it discerns two strata of ability: high and low comprehension. It should be noted that weighted analysis, which is the analysis that should be conducted to estimate persons' comprehension measures, produced a good Rasch person reliability of .84 with an equally good separation of 2.31. Readers should be reminded that weighting inflates the reliability and separation indices so, for this reason, an unweighted analysis is reported.

RQ5: Is the item reliability of the combined comprehension measure sufficient to suggest replicability of the item difficulty hierarchy if the test were administered to a sample of similar ability? and RQ7: Does the sample of participants separate the items into different levels of difficulty?

In contrast, the Rasch item reliability estimate was .98, which is excellent according to criteria recommended by Fisher (2007). This result indicated that confidence can be placed in this order of item difficulty being replicated across similar samples. The high reliability was accompanied by an equally high separation index of 7.88, which is also excellent according to the criteria (Linacre, 2007) and indicates eight distinct levels of item difficulty. These reliability and separation values provide further validity evidence for the use of the instrument as a measure of comprehension.

As with the Rasch person reliability and separation, it is noteworthy that the weighted analysis revealed a Rasch reliability of .99 and a separation of 8.51, which represent excellent values according to the adopted criteria.

RQ6: Do the Raters Fit the Rasch Model?

As can be seen in Table 3, both raters had infit and outfit MNSQ indices that fell within the lower-control limit of 0.50 and the upper-control limit of 1.50 as proposed by Linacre (2002), with infit MNSQ values of 0.60 and 0.70 and outfit MNSQ values of 0.56 and 0.64, respectively. Thus, the raters performed consistently, well within the Rasch model's expectations.

RQ7: Do the raters exercise a similar level of severity when awarding ratings to the summaries?

In response to RQ9, an inspection of the rater measurement information was carried out. Detailed measurement results for each individual rater are shown in Table 3, from which it can be seen that the rater severity measures of Barney (severity measure = -0.01 logits) and Simpsons (severity measure = 0.01 logits) are homogeneous. The variance between the raters in their degree of severity was almost non-existent because their logit severity difference was only 0.02 logits. These data were corroborated by the χ^2 value ($\chi^2 = 0.30$, $df = 1$, $p = .61$), which indicated that the raters were equally severe. These results lined up with the evidence gained

from looking at the Wright map, which showed that both raters were located adjacent to each other around the zero-logit point. In addition, the raters displayed a high level of agreement with 1,001 exact agreements (87%) out of the possible 1,150 inter-rater agreement opportunities. In fact, they had a higher inter-rater agreement than that expected by the Rasch model (48.9%). The separation index was .00 and was accompanied by a reliability statistic of .00. These estimates indicated that the raters formed a homogeneous group in which they functioned interchangeably by having a similar understanding of how to interpret and use the rating scale as well as having a similar degree of severity.

Table 3: Raters Measurement Report

Raters	Obs. av.	Fair av.	Mea.	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PM	PM ex.
[a] B	0.75	1.61	-0.01	0.04	0.70	-8.80	0.64	-9.00	.77	.64
[b] S	0.72	1.59	0.01	0.05	0.60	-9.00	0.56	-9.00	.80	.66

Note. *N* = 290. Obs. av. = observed average measure, fair av. = fair average measure, mea. = measure, MNSQ = mean-squared, ZSTD = Standardized z-scores, PM = point-measure correlation, PM ex. = expected value of point-measure correlation.
[a] B = Barney.
[b] S = Simpsons.

RQ8: Does the rating scale function adequately?

The functioning of the rating scale was evaluated by examining the adherence of the scale to the three essential criteria highlighted by Linacre (2002). The preliminary criterion was satisfied as the scale was oriented with the latent variable such that higher scores indicated a higher level of ability on the measured construct (i.e., comprehension). As seen in Table 4, all of the category frequencies had more responses (i.e., counts used) than the recommended minimum of 10, with the smallest number of used category counts being 18. Therefore, Criterion 1 was met. Also, the thresholds increased in order along with category (i.e., scores of 0–4) from -1.91 at Category 1 to 1.39 logits at Category 4, indicating that the higher the performance category, the higher the measure of comprehension. That is, a score of 1 corresponded to higher average ability than a score of 0 (-1.91 and -2.97 logits, respectively), a score of 2 corresponded to higher ability than a score of 1 (-0.94 and -1.91 logits, respectively), and so on until the maximum score of 4 (average ability measure = 1.39 logits). In other words, the average measures increase monotonically across the response categories (i.e., Categories 0, 1, 2, 3, 4), corroborating Criterion 3. More evidence for the threshold estimations was provided by the outfit mean-square values where the highest value is 0.80 logits for a score of 0 and 4 on the rating scale, which is below the cut-off criterion of 2.00 logits (Linacre, 2002).

Table 4: Rating Scale Combined Comprehension Measurement Report

Score	Counts used	Average measure	Expected measure	Outfit MNSQ	Rasch-Andrich thresholds	*SE*
0	1,419	-2.97	-2.80	0.80	—	—
1	935	-1.91	-2.03	0.60	-2.00	0.04
2	348	-0.93	-1.19	0.50	-0.63	0.06
3	131	0.40	-0.23	0.60	0.25	0.10
4	18	1.39	1.17	0.80	2.38	0.27

Note. *N* = 290. MNSQ = mean-square.

Discussion

Summary of Results

The psychometric properties of the combined comprehension measure were established. The Wright map indicated that the summaries were more difficult than the MCQs as was hypothesized. The percentage of participants demonstrating acceptable fit was high: 92% for infit MNSQ and 85% for outfit MNSQ, respectively. Only a small percentage of participants (1% for infit MNSQ and 5% for outfit MNSQ) substantially deviated from the Rasch model's expectations concerning fit. In the same vein, the items displayed good fit to the model because all of the items fell within, or were slightly above, the adopted bounds of 0.70 to 1.30. Item F5, however, displayed an outfit MNSQ value of 2.90, which was caused by the unexpected responses of three participants.

Fair Rasch person reliability and separation estimates were found at .72 and 1.62, respectively, which indicated that the participants could be separated into two levels of listening automaticity (i.e., high and low) fairly reliably. In contrast, the items had excellent reliability and separation at .98 and 7.88, respectively, indicating that there were eight distinguishable levels of item difficulty and that the replicability of the item ordering was high.

Both raters behaved as expected by the Rasch model with fit values ranging from 0.56 to 0.70. Additionally, the rater reliability and separation were both .00, which indicated that the raters were not statistically different from each other. In addition, the rating scale showed good functioning because the ratings given by the raters increased with ability.

General Discussion

The results of the MFRM analysis indicate that the combined measure is a valid approach to assessing comprehension. The fit indices revealed that both summaries and MCQs can jointly measure a common latent trait, namely comprehension, despite being different task formats, which likely elicit distinct cognitive processes. This finding provides supporting evidence for the integration of productive and receptive assessment tools into a single measurement system.

The results showed that the item difficulty hierarchy was topped by the summaries matching the initial prediction that the summaries would be more difficult than the MCQs. This finding supports the weighting approach to estimate persons' comprehension measures. Summaries require learners to determine what information is important and how the ideas relate to one another in order to integrate the information into a mental representation of a target text. In contrast, MCQs may demonstrate a lower degree of comprehension because they require learners to recognize the correct answer from the given choices.

This study provides evidence to suggest that the combined measure functions reliably in this multi-task context where summaries provide global comprehension and MCQs complement them by tapping into detail-level comprehension. When used together, these two formats of assessment are likely to provide a more balanced and theoretically grounded representation of comprehension than either one alone. Kinstch (1998) argues that comprehension is the ability to construct a coherent mental representation of the input but that details are less likely to be encoded as learners tend to focus on important information. Through this approach, learners recall the mental representation that they have created upon listening to or reading a text in the form of a summary and they subsequently answer MCQs as an assessment of details based on recognition. The combined approach mitigates the limitations of using summaries (e.g.,

difficulty to tap into detail-level comprehension), or MCQs (e.g., relying on recognition or using the information contained in the questions to construct a mental representation of meaning) as a stand-alone measure of comprehension.

The findings of this study have methodological implications regarding Rasch-based research, particularly MFRM. A clear implication is that MFRM can accommodate differently weighted tasks into a common measurement framework. Therefore, researchers can weigh the level of contribution of different tasks measuring the same construct based on theory, particularly when computing person measures for subsequent statistical analyses. From a classroom-based assessment perspective, where raw scores are used, the weighting approach could be applied by multiplying the summary ratings by two.

While this study provides valuable insights into comprehension measurement, several limitations should be acknowledged. A primary limitation concerns lack of practicality. It takes time to rate the summaries as the raters have to read the summaries and award ratings using a rating scale in tandem with a previously created list of main ideas and important details. In addition, the administration of the combined assessment takes longer than a single-task test. A further limitation is that a certain level of expertise is required to operate the assessment. Although some research has been conducted on Artificial Intelligence as an automated assessment tool, the evidence argues against removing human raters because they can provide personalized feedback, which can have a larger impact on motivation, engagement, and self-efficacy (Alshehri, 2025).

Conclusion

In this study, I proposed a novel approach to assessing comprehension by combining summaries, a productive task, with MCQs, a receptive task. Validity evidence was collected for the combined comprehension measure through MFRM. This novel approach offers promise in assessing comprehension through the integration of meaning reconstruction as assessed by summaries, and recognition-based skills as assessed by MCQs. In this integrated measurement system, summaries provide evidence of global comprehension, whereas MCQs capture detail-level comprehension. As summaries require more cognitive demands and represent a higher level of comprehension than MCQs, they are assigned double weight for the estimation of persons' comprehension measures.

Although the data for the present study were collected in a foreign language-learning context, the proposed combined measure is intended to model comprehension as a general cognitive construct underpinning not only listening and reading performance but also content learning across educational disciplines.

From a practical point of view, the proposed approach offers a useful tool to researchers and educators searching for more psychometrically robust and theoretically-driven assessment methods. It is hoped that the present approach is used by researchers to continue to explore multi-faceted tests with different facet weight and their implications for comprehension assessment. By moving beyond single-task tests, comprehension can be more accurately measured.

Implications

This study has implications for practice by offering a practical framework to improve comprehension assessment. The findings show that summaries and MCQs can function as

complementary components of the measurement, with summaries encapsulating global comprehension and MCQs targeting detail-level comprehension. By using both tasks in tandem, the weaknesses of each task when used in isolation are mitigated. In addition, the weighting scheme provides a more theoretically-driven approach to assessment, in which tasks that have a stronger impact on the measurement are given extra weight. This study illustrates how different tasks, which are differently weighted, can be integrated into a single analysis using MFRM.

Regarding policy, the study challenges the widespread reliance on single-task assessments. Using MCQs alone may overestimate learners' level of comprehension and thus, this study advocates for the inclusion of productive tasks such as summaries to obtain more precise estimates of ability.

REFERENCES

Alaofi, A. O. (2020). Difficulties of summarizing and paraphrasing in English as a foreign language (EFL): Saudi graduate students' perspectives. *International Journal of English Language Education*, *8*(2), 193–211. https://doi.org/10.5296/ijele.v8i2.17788

Alshehri, A. (2025). AI's effectiveness in language testing and feedback provision. Social Sciences & Humanities Open, 12, 101892. https://doi.org/10.1016/j.ssaho.2025.101892

Bazan, B. (2024). *Listening automaticity: A reduction of dual-task Interference and working memory demands*. IPR Journals and Book Publishers.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*(1), 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Erlbaum.

Chan, N. & Kennedy, P. E. (2002), Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*, 957–971. https://doi.org/10.1002/j.2325-8012.2002.tb00469.x

Chen, G., Cheng, W., Chang, T-W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journals of Computers in Education, 1*, 213–225. https://doi.org/10.1007/s40692-014-0012-z

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment–companion volume*. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, *84*, 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095. http://www.rasch.org/rmt/rmt211a.htm

Hughes, A. (2000). *Testing for Language Teachers*. Cambridge University Press.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Klufa, J. (2015). Multiple choice question tests: advantages and disadvantages. *Mathematics and Computers in Sciences and Industry Journal*, *3*, 91–97. https://files.eric.ed.gov/fulltext/EJ1272114.pdf

Lee, H-S., Liu, O., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115–136. https://doi.org/10.1080/08957347.2011.554604

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Mesa Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106. https://europepmc.org/article/med/11997586

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program.* MESA.

Linacre, J. M. (2017). FACETS (Version 3.80.0) [Computer Software]. Winsteps.com.

Linacre, J. M. (2023). *A user's guide to Facets Rasch-model computer programs.* https://www.winsteps.com/a/Facets-Manual.pdf

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344. https://doi.org/10.1177/0956797612443370

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

Mee, J., Pandian, R., Wolczynski, J., Morales, A., Panigua, M., Polina, H., Baldwin, P., & Clauser, B. E. (2024). An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. *Adv in Health Sci Educ 29*, 783–801. https://doi.org/10.1007/s10459-023-10266-3

Qin, J., & Groombridge, T. (2023). Deconstructing summary writing: Further exploration of L2 reading and writing. *Sage Open*, *13*(4). https://doi.org/10.1177/21582440231200935

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danmarks Paedagogiske Institut.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379. https://files.eric.ed.gov/fulltext/EJ1272114.pdf

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, *13*(2), 173–189. https://doi.org/10.1177/026553229601300203

Sick, J. (2013). Rasch measurement in language education part 8: Rasch measurement and inter-rater reliability. *Shiken*, *17*(2), 23–26. https://teval.jalt.org/sites/default/files/SRB-17-2-Sick-RMLE8.pdf

Wright, B. D., Linacre, M., Gustafsson, J.-E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. https://rasch.org/rmt/rmt83.htm

Zhou, Q. (2019) The feasibility of measuring reading ability by the format of short answer questions. *US-China Foreign Language*, *17*(1), 1–9. https://doi.org/10.17265/1539-8080/2019.01.001