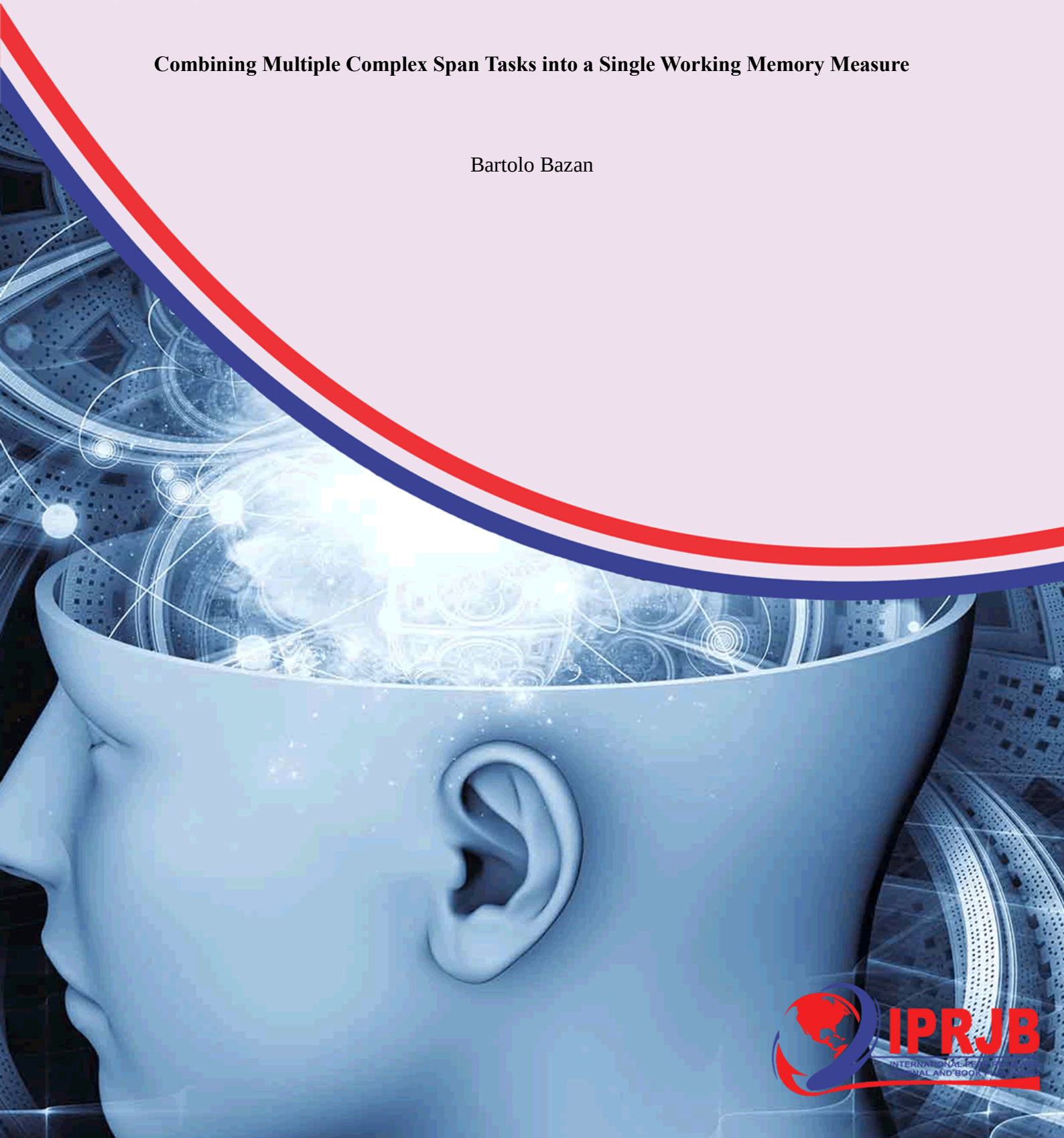
International Journal of
Psychology
(IJP)

Combining Multiple Complex Span Tasks into a Single Working Memory Measure

Bartolo Bazan



IPRJB

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026

www.iprjb.org

Combining Multiple Complex Span Tasks into a Single Working Memory Measure

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

Article History

Received 14th December 2025

Received in Revised Form 8th January 2026

Accepted 10th February 2026

How to cite in APA format:

Bazan, B. (2026). Combining Multiple Complex Span Tasks into a Single Working Memory Measure. *International Journal of Psychology*, *11*(1), 16–35. https://doi.org/10.47604/ijp.3628

Abstract

Purpose: The adequate measurement of working memory (WM) capacity presents some limitations. One of the main challenges is that WM assessment is confounded by the level of expertise of individuals in the particular domain required to perform the task, such as verbal fluency in the case of the speaking span task. Another drawback is that there is little psychometric evidence to support the use of complex span tasks as measures of WM capacity. Therefore, it is not clear if these assessment tools do in fact measure the theoretical construct they are intended to measure (i.e., WM) or something else. The purpose of this study was to address these shortcomings by developing a new measure that combines the listening and the speaking span tasks and collecting validity evidence for its use through the Rasch model.

Methodology: The participants were 290 Japanese high school students who were administered the speaking and the listening span tasks for which I collected validity evidence in Bazan (2020) and Bazan (2021), respectively. Both tasks were performed individually on a face-to-face basis with the stimuli being played on a computer that I operated. Performance was audio-recorded and scored dichotomously (i.e., right or wrong) using the same scoring system as in the two previous studies. That is, a credit was given for each item recalled successively in the order of appearance until memory failure to recall in order. Scores were put together and analyzed as if they belonged to a single test through the Rasch dichotomous model. The analysis involved an evaluation of whether later presented items within a set increased in difficulty as predicted by WM theory, person and item fit to the Rasch model, person and item reliability and separation, and the dimensionality of the combined WM measure.

Findings: The Wright map confirmed a hierarchy of item difficulty consistent with the theoretical expectation that the further the item appears within the set, the more difficult it should be. Over 96% of participants and almost all items fit the Rasch model, with person and item reliability indices demonstrating high replicability of the ordering of the persons' ability and item difficulty across similar samples. Person separation indicated that the measure is sensitive enough to separate participants into three levels of the construct (i.e., high spans, average spans, and low spans) whereas item separation showed that the items can be divided into 9 levels of difficulty, which is excellent according to the guidelines. The examination of dimensionality revealed that the combined measure taps into a single unidimensional construct, namely WM capacity.

Unique Contribution to Theory, Practice and Policy: This study provides evidence for the usefulness of the combining approach to mitigate the influence of domain-specific skills on WM measurement.

Keywords: *Rasch Model, Working Memory, Speaking Span Task, Listening Span Task, Complex Span Tasks, Working Memory Assessment*

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

INTRODUCTION

Working memory (WM) refers to a memory system with limited capacity that is responsible for the temporary maintenance of information in an active state while simultaneously processing the same or other information (Baddeley, 2007; Bayliss et al., 2005; Cowan, 2017). WM has been found to be allegedly involved in the performance of a wide array of cognitive skills, including mental arithmetic, reasoning, planning, and problem-solving (Conway et al., 2005).

At the forefront of WM measurement are complex span tasks, which are dual tasks hypothesized to tap into WM by requiring individuals to hold in memory a series of items for subsequent serial recall in the face of interference caused by a concurrent processing task (Munakata et al., 2007). In fact, complex span tasks, such as the listening and speaking span tasks, are among the most commonly employed measurement instruments in cognitive psychology (Miyake, 2001; Conway et al., 2005).

Despite their prominence however, the assessment of WM through complex span tasks up to date have presented some methodological drawbacks. One of the main limitations is that due to practical constraints regarding length of administration, researchers have used single complex span tasks to measure WM (Monteiro et al., 2025). This is problematic because a single measurement is likely to be contaminated by the confounding of the individual's level of expertise in the particular skill required to perform the task (e.g., speaking fluency on the speaking span task) with WM abilities (Dehn, 2008).

Moreover, little effort has been made to validate complex span tasks through the application of a robust psychometric approach such as the Rasch model. Although validity evidence has been collected for the use of complex span tasks through factor analysis and structural equation modeling (Engle et al., 1999; Kane et al., 2004; Miyake et al., 2000; Monteiro et al., 2025; Oberauer et al., 2003; Oswald et al., 2015; Schmiedek et al., 2014), these methods are sample-dependent and thus, the likelihood of replicating the same results with different datasets is low (Miyake et al., 2001; Wright, 1996). In Rasch measurement however, once calibrated, item difficulty estimates are stable across samples, and person ability estimates are stable across different item sets (i.e., estimates of ability are stable across other complex span tasks), which supports the generalizability of the WM measure in question across populations (Bond et al., 2021).

The purpose of this study is to account for the shortcomings of previous measures by validating through the Rasch model a new WM measure that I developed in Bazan (2024) by combining the speaking and the listening span tasks. The rationale underlying the measure is as follows. First, the speaking span task where individuals are asked to produce utterances using lists of random words, is influenced by verbal fluency because higher fluency can reduce the duration of the interval over which the words must be retained, resulting in longer short-term retention of the words. For this reason, the speaking span task is combined with the listening span task, which involves receptive rather than productive oral skills. Next, by combining the items from both tasks into a single analysis, reliability should increase because a larger number of items should separate individuals into the varying levels of the hypothesized construct, in this case WM, more precisely (Bond et al., 2021).

As complex span tasks, both the listening and speaking span tasks are assumed to place demands on effortful limited-capacity controlled processing by requiring the temporary maintenance of information in the face of processing interference. The measurement of

executive functioning (e.g., shifting, updating, and inhibition) is thus beyond the scope of this study.

The underlying assumption is that the combined measure taps into general WM capacity, which is a domain-free pool of cognitive resources. This view is supported by empirical evidence that have examined whether different classes of working memory tasks measure the same general construct of WM capacity (Daneman & Merikle, 1996; Wilhelm et al., 2013).

The theoretical justification for the use of a combination of complex span tasks derives from evidence regarding misclassifications of participants as high when they should have been classified as low or vice-verse. Conway et al. (2005) found that participants are more likely to be classified in the correct quartile when two complex span tasks are given than when only one is given. Based on this evidence, they recommend assessing WM through at least two complex span tasks.

LITERATURE REVIEW

The Rasch Model

As complementary to traditional psychometric approaches within an item response theory paradigm, the Rasch Model offers a measurement framework that supports and expands upon classical statistical analyses. The Rasch Model refers to a probabilistic framework for measurement to which psychometricians can fit their data in order to examine the validity of a test. Central to the Rasch Model is the concept of unidimensionality, that is, any test should involve a single latent trait (i.e., WM). The Rasch Model estimates the level of ability of each test-taker and the level of difficulty of each item on a common logit scale by mathematically transforming raw scores, where differences between consecutive data points do not represent equal amounts of the construct into equal-interval measures, where differences on the scale represent equal differences in the measured latent trait. The Rasch Model is a useful tool to establish the construct validity of a measure because it provides detailed information about different aspects of validity. Rasch analyses provide an item-person map, also known as a Wright map (Bond et al., 2021), that graphically displays the person ability-item difficulty relationship 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. That is, if the items hypothesized to be more difficult when designing the test are indeed more difficult. In addition, Rasch analyses produce item and person fit indices, which are useful for examining the contribution of the individual items to the measurement of the underlying construct and for exploring if the participants' performance is in accordance with the model expectations.

A further advantage is that Rasch analyses provide reliability indices for both items and persons that indicate the degree to which the replicability of the item difficulty hierarchy and the spread of the participants' ability levels is possible were the test administered to a similar sample. Moreover, the model uses separation indices, which show the number of ability levels and item difficulty levels into which participants and items can be reliably separated. Finally, the principal component analysis (PCA) of the Rasch residuals shows the extent to which the items adhere to the measurement of a single underlying construct, thus satisfying the unidimensionality criterion of Rasch measurement (Bond et al., 2021). The PCA is accompanied by a fit graph, which is useful for visually assessing the extent to which the items contribute to the measurement of a single latent trait.

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

The Rasch model offers two advantages that are particularly important for WM assessment: invariance and unidimensionality. Measurement invariance implies that persons' estimates of WM capacity remain consistent across WM tests regardless of the content of the test (e.g., words, numbers, letters). Similarly, the functioning of the items is consistent across samples. This is essential for comparisons between clinical groups and control groups, different age groups, or to track longitudinal changes in WM capacity. Without invariance, ability estimates may reflect a task artifact rather than differences or changes in WM. Unidimensionality means that all the items on the task measure the same construct. This does not imply that performance on a complex span task is due to a single cognitive process. In fact, a variety of cognitive processes such as storage, attentional control, or processing speed are involved in the performance of a complex span task. However, as long as these processes operate jointly, unidimensionality is maintained. Basically, unidimensionality is what makes the construct of WM exist.

Raw scores are particularly problematic for WM measurement because they are counts of correct responses (i.e., ordinal data) whereas Rasch-modeled estimates are measures (Bond et al., 2021) as they take into account item difficulty (i.e., interval-level data). That is, the more difficult the items are, the higher the demands they place in WM, which should in turn be reflected in higher estimates of WM capacity. For example, the difference in WM capacity on the listening span task between a span of 2 and a span of 3 is not equivalent to the difference between a span of 5 and a span of 6 as Items 5 and 6 should be more taxing because they have more items that need to be remembered preceding them than Items 2 and 3.

Wright Map

Winsteps (Linacre, 2018a), which is the Rasch software package used in this study, produces a visualization of the data called a Wright map. The Wright map shows the performance of each person on a given test and the test items, which are typically represented by an "X" and the item number, respectively. The logit scale, "which is the joint scale of person ability and item difficulty, is displayed down the middle of the map" (Bond et al., 2021, p.56). The logit measures, common to both persons and items, can be read on the far-left side of the map. The persons and items are spread along the logit scale in descending order of ability and difficulty, respectively. Thus, the higher a person's performance is on the map, the higher their ability and the higher an item is on the map, the higher its difficulty. The Wright map is thus useful to visually analyze the relations between persons and items such as the targeting of the items and to verify whether the difficulty hierarchy of the items reflects the theorized order. In the context of this study, items appearing later within each set should be displayed above earlier items, as they are hypothesized to tax WM to a greater extent. For example, in a set of three items in either task, the third item should be positioned above the second, which in turn should be positioned above the first.

Person and Item Fit

Fit is a quality-control mechanism that is used to evaluate how well the data adheres to the Rasch model's expectations. The Rasch model provides two fit statistics, infit MNSQ and outfit MNSQ. Infit MNSQ is a weighted unstandardized statistic whose estimation is impacted by unexpected responses close to a person's level of ability or an item's level of difficulty, respectively. In contrast, outfit MNSQ is a non-weighted standardized statistic, which is affected by outliers. That is, unexpected responses far from a person's level of ability or an item's level of difficulty (Wright & Masters, 1982). 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. For this reason, infit MNSQ is usually the statistic that guides the evaluation of fit (Bond et al., 2021). In this investigation too, 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.

To evaluate infit and outfit MNSQ, I adopted the criteria put forward by Wright et al. (1994) and Linacre (2007), who consider a range of between 0.50 and 1.50 logits to be satisfactory for measurement. Although values above 1.50 flag misfit, values within the range of 1.51 and 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).

Reliability and Separation

In addition to the infit and outfit statistics, Rasch produces person and item reliability and separation indices, which can be used to further examine the performance of the persons and items in a dataset. 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 et al., 2021). That is, for example, if persons who scored highly on a particular test, such as a speaking span task, they would also score highly on other similar speaking span tasks or, conversely, if persons scored poorly, they would score poorly again.

In this study, the reliability estimates were interpreted following the guidelines proposed by Fisher (2007). According to Fisher's proposed guidelines, 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 those above .94 indicate excellent reliability.

Together with the reliability estimates, Rasch analysis provides person and item separation indices, which serve as additional tools for evaluating the spread of persons and items along the measured construct, respectively. An index of 1.50 discerns two measurably distinct levels of person ability or item difficulty, an index of 2.00 discerns three levels, and an index of 3.00 discerns four levels (Duncan, et al., 2003). 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.

In the context of WM measurement, high person separation indicates that the WM instrument can reliably distinguish individuals with different levels of WM capacity. Conversely, low separation suggests little variation in WM capacity. Unlike extreme-group designs or split quartiles in which individuals can be misclassified due to measurement error (Conway et al., 2005), the Rasch separation index takes into account measurement error and extreme scores (Bond et al., 2021) therefore, the observed value is likely to reflect variation in WM capacity rather than measurement error. Consequently, person separation has practical implication for educational and clinical settings such as enabling targeted instructional support or differentiating different levels of WM impairment, respectively.

PCA of Item Residuals

One basic requirement of the Rasch model is that the data adhere to the measurement of a unidimensional construct, that is, that the items tap into the same latent trait. To assess the unidimensionality requirement of a measure, Winsteps provides the item or person residuals

PCA. Because residuals are random noise, they should not form systematic patterns. In other words, they should not correlate with each other (Linacre, 1998). If the residuals show no systematic relationship, then the measure is fundamentally unidimensional. Conversely, a systematic relationship among the residuals indicates the existence of a second dimension.

A unidimensional measure should satisfy two criteria. First, it should explain at least 20.00% of the variance in the data (Reckase, 1979). Second, the first residual contrast of unexplained variance should have an eigenvalue below 2.00 (Linacre, 2018b) and represent less than 10.00% of the total variance (Linacre, 2007).

Variable Pathway

In addition to the PCA of item residuals, Winsteps provides a visual tool to graphically explore whether the items in a test adhere to the measurement of a unidimensional construct, the variable pathway or the fit map. The map displays a path delimited by two solid lines with a dotted line in the center, which represents the measured construct. The items represented by asterisks are spread along the path vertically. Items within the boundaries of the path are thought to assess the same single construct whereas items outside the boundaries indicate multidimensionality (Bond et al., 2021). Therefore, a visual inspection of the variable map supplements the numerical evidence from the PCA when assessing unidimensionality. It should be noted that although the map shows two different pathways, infit MNSQ and outfit MNSQ, I reported the pathway for infit MNSQ because in this investigation, decisions about fit were made primarily based on infit MNSQ.

The Rasch Model and Complex Span Tasks

To the best of my knowledge, there are only two complex span tasks for which Rasch validity evidence has been collected. These tasks are variants of Daneman and Green's (1986) speaking span task and Daneman and Carpenter 's (1980) listening span task. In Bazan (2020), I developed a new speaking span task in which vocabulary was controlled for by a) keeping word length constant (i.e., from two to three mora) across trials, b) including only 12 abstract words in the test, and c) having two Japanese speakers check their degree of familiarity with the words. The words were randomly arranged into two sets of two, three, four, five, and six sets totaling 40 items and unlike its predecessors, the task was administered in auditory form. The participants were 31 Japanese speakers aged between 13 and 14 years old, who were required to listen to the increasingly larger sets of words, hold in memory the words in the set, and produce and utterance for each word in the set in order of appearance. Data were scored dichotomously (i.e., right or wrong) using a new scoring system where a point was awarded for each utterance produced correctly (i.e., contained the target word) and in order until memory failure. For instance, if on a set of four items, participants recalled the first and the second items, failed to recall the third item, but successfully recalled the fourth item, they would get a score of two points (i.e., one for Item 1 and one for Item 2). A Rasch analysis of the data indicated that the items ranged on a continuum from less difficult (i.e., first items in the set) to most difficult (i.e., last items in the set), matching the predicted order based on theory. All items except for Item 3.2 (*yuubinkyoku*, post office) Infit MNSQ = 1.20, Outfit MNSQ = 9.90) showed good fit to the Rasch Model, providing further validity for the use of the measure. Additional validity evidence was provided by the reliability and separation indices, which revealed that the measure reliably (.81) separated participants (Rasch person separation = 2.10) into three levels of the construct (i.e., low spans, average spans, and high spans). Furthermore,

the Rasch principal components analysis of item residuals (PCA) and the item fit graph indicated that the measure tapped into a unidimensional latent trait.

Another span task whose psychometric properties were evaluated through the Rasch Model was the shortened listening span task that I described in Bazan (2021). The task contained 40 short Japanese utterances, which ranged between three and five words. To account for possible knowledge biases of previous tasks, all utterances were casual. Half of the utterances were grammatical and the other half ungrammatical (i.e., incorrect word order) and they were randomly arranged into two sets of two, three, four, five, and six utterances. The task was given to the same 31 Japanese speakers that took the speaking span task, who were required to verify the plausibility of each utterance, while holding in memory the last word of each utterance in the set for serial recall at the end of the set. Using the same scoring system as in Bazan (2020), the words that were correctly recalled in order of appearance were awarded 1 point. A Rasch analysis of the data revealed that the farther in the set the item appeared, the more difficult it was as it was hypothesized. All items showed good fit to the Rasch Model and the Rasch person reliability was of .84, suggesting that the probability of obtaining a similar spread of participants' WM capacities in similar samples is high. The Rasch person separation was estimated at 2.28, indicating that the measure separated participants into three levels of the construct (i.e., low spans, average spans, and high spans). In addition, the examination of the results of the PCA and the item fit graph demonstrated that the items adhered to the measurement of a unidimensional trait. The validity evidence obtained in the present study is consistent with, and extends, previous validation findings, including the concurrent validity reported by Ivanova and Hallowell (2014) for their modified listening span task and the convergent validity reported by Ünal et al. (2020) for their Turkish adaptation.

Aside from the work that I conducted in Bazan (2020) and Bazan (2021), I employed the Rasch Model to combine the listening and the speaking span tasks into a single WM measure for subsequent statistical analyses in Bazan (2024). However, I did not examine the validity of the combined measure. The present study presents Rasch-validity evidence for the use of the combined WM measure that I designed in Bazan (2024).

Research Questions

The research questions (RQs) that guided the evaluation of the psychometric properties of the combined WM measure are as follows:

1. Do the items within the sets of the listening and speaking span tasks in this combined context gradually increase in difficulty as expected based on theory (i.e., the further the item position within the set, the more difficult the item should be)?
2. Do the persons fit the Rasch model?
3. Do the items fit the Rasch model?
4. Is the person reliability of the combined WM measure sufficient to suggest a similar spread of participants with higher and lower spans if they were given a different complex span task?
5. Does the combined measure separate participants into different levels of WM capacity?
6. Is the item reliability of the combined WM measure sufficient to suggest replicability of the item difficulty hierarchy if the listening and speaking span tasks were administered to a sample of similar ability?
7. Does the sample of participants separate the items into different levels of difficulty?

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

8. Is the combined WM measure unidimensional?

METHODOLOGY

Participants

The sample was composed of 290 students (41% female, 59% male) attending 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 Japanese speakers with no reported history of language or cognitive impairment. Ethical approval was obtained and the study was conducted according to the guidelines of the institution for research.

Instruments and Procedure

The instruments were the listening span task and the speaking span task that I developed in Bazan (2020) and Bazan (2021), however, the word *yuubinkyoku* (post office) on the speaking span task was replaced for *mizu* (water) because it was the longest word on the test and produced an extreme outfit MNSQ value (9.90) in the 2020 study.

Data collection took place over two fixed cycles, each of approximately 70 sessions, which were administered individually during a two-year span. In the first cycle, all participants completed the speaking span task and in the second cycle, they completed the listening span task. Data from three participants were collected on each testing session for a total time of approximately 40 minutes per session. On each testing day, the scheduled participants were given written and oral instructions in Japanese, two practice items, and the opportunity to ask for questions or clarification. Both tasks were presented auditorily via computer speakers on a Windows device that I operated. Performance was audio-recorded and the tasks were scored in the same manner as in Bazan (2020) and Bazan (2021).

Analysis

Scores for both tasks were input into an Excel spreadsheet as if they belonged to a single WM span task composed of 80 items (the 40 items of the listening span task and the 40 items of the speaking span task). Then, the spreadsheet was imported into Winsteps version 4.3.1 Rasch software (Linacre, 2018a) for an analysis using the Rasch dichotomous model (Rasch, 1960). The data of four participants who missed the listening span task, and those of another participant who missed the speaking span task, were entered into the analysis as missing. The data for the listening span task of three participants were lost due to a technical failure of the recording equipment and were also entered as missing. For the same reason, the speaking span data for Sets 8 through 10 of a participant were also included as missing.

RESULTS

The Wright Map

Both the speaking span task and the listening span task were developed under the theoretical assumption that the difficulty of the items increases as the sets lengthen. Specifically, as items need to successively be stored in WM in the face of the concurrent processing component (i.e., producing utterances or judging their grammatically), the duration of the retention interval is increased, thereby imposing greater demands on WM. Hence, Item 2 in any set should be more difficult than Item 1, subsequently Item 3 should be more difficult than Items 2 and 1 and so forth. This prediction is depicted by the Wright map in Figure 1.

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

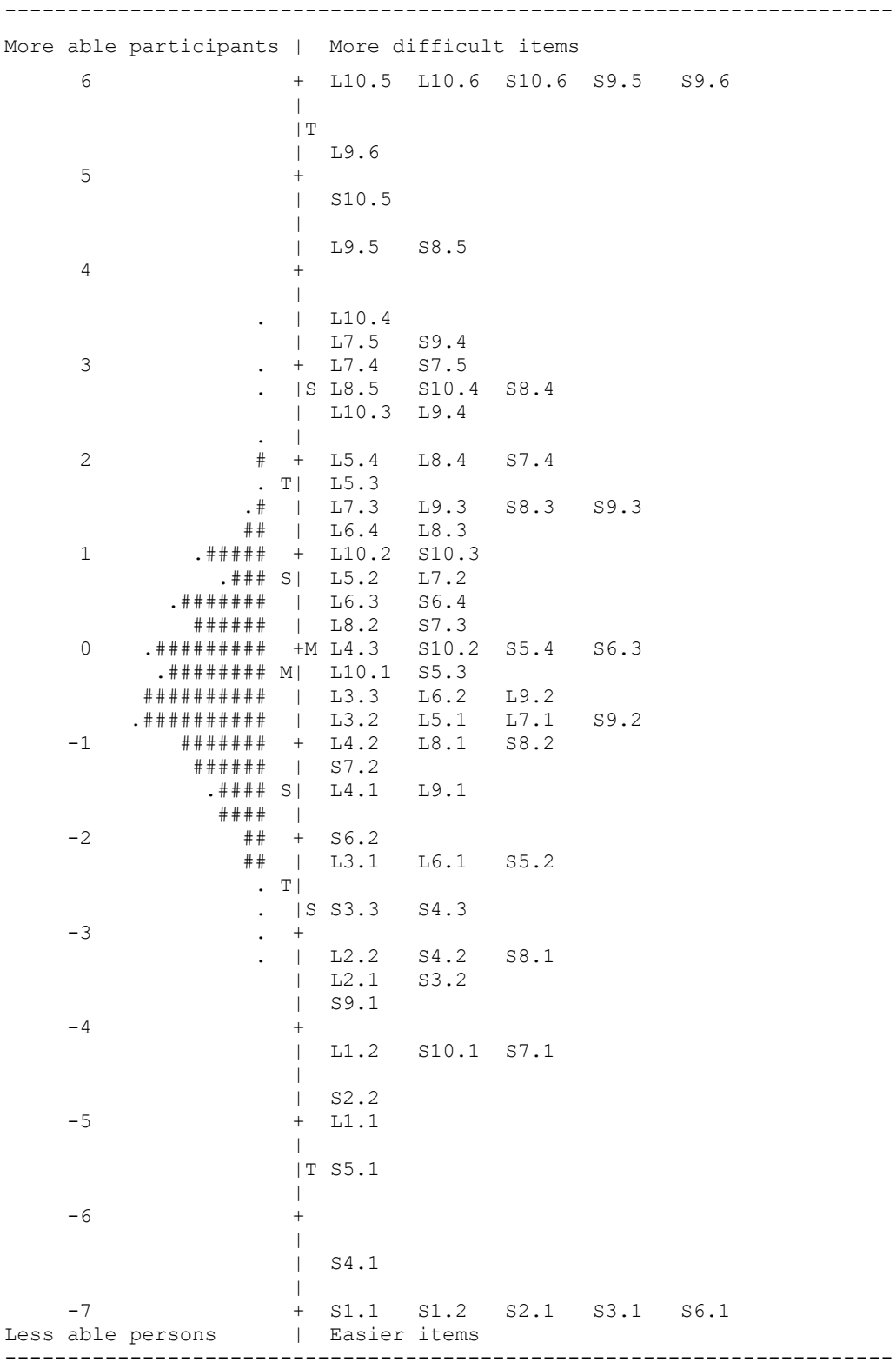


Figure 1. Wright Map for the Combined WM Measure. N = 290.

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

Items are labeled L and S, which refer to the listening and speaking span tasks, followed by the set number and the item number after a period. # = three participants, . = one or two participants.

The items starting with an L followed by the set number and the item number represent the listening span task items. In contrast, the items starting with an S followed by the set number and the item number represent the speaking span task items. As shown in the figure, the ordering of the items of both span tasks in this combined context matches the theoretically predicted hierarchy that the further the position for the item within the set, the more difficult the item should be to answer. This ordering can be illustrated by taking the example of the three items in Set 3 of the listening span task, where Item L3.3 is plotted higher than Item L3.2, which is plotted higher than item L3.1 (see Figure 1).

The participants, who are to the right of the logit scale, and are represented by # or a dot, are evenly spread out over approximately six logits forming a bell curve, suggesting a good spread of WM span. However, the locations of the participants in comparison to those of the items suggest that the WM span tasks were difficult for the sample because the items were spread out to a larger extent than the participants. Looking at the targeting, a number of items at the bottom of the distribution were well within the WM capacities of the participants. Most of these items were the first or second items in the sets, which reflected a primacy effect (Howieson & Lezak, 2012) where the first items in a list of words are easier to recall. In contrast, there is a cluster of items located at the higher end of the distribution, whose difficulty exceeds the participants' WM capacities. Most of these items correspond to the final items of the largest sets (i.e., six-item sets), which were hypothesized to be the most difficult items. This apparent targeting problem was, however, an artifact of the scoring system, in which 1 point is awarded to each word recalled in a string in the correct order of appearance until memory failure to recall in order. For example, if on a set of six items, a participant succeeded on the first and second items, failed the third item, but succeeded on the fourth, fifth, and sixth items, she would get a score of 2 in the set. Thus, to be able to score on the fifth or sixth item of the largest sets, the participant must succeed in all previous items, which only a handful of participants could do.

Person and Item Fit

Next, I conducted an analysis of the person fit. Table 1 presents a summary of the person fit statistics for the combined WM instrument. Of importance is that 96% of the participants satisfied the 0.50 to 1.50 criterion with respect to infit MNSQ and 67% did so with respect to outfit MNSQ. No participant had an infit MNSQ value above 2.00, which would distort the measurement. In contrast, 7% of the sample exhibited high levels of outfit. To investigate the source of such concerning values, I examined the table of poorly fitting persons provided by the Winsteps output. This analysis indicated that most of the misfitting participants had high WM spans who, possibly, due to a lack of concentration, failed to succeed on items that were within their level of ability. For example, Participant 51107, with an estimated ability of 2.74 logits, unexpectedly failed the first three-item set of the speaking span task, particularly on Items S3.2 and S3.3 with difficulty measures of -3.54 and -2.78 logits, respectively. Similarly, Participant 50319, with an estimated ability of 2.29 logits, was unexpectedly unsuccessful on the second two-item set of the listening span task, Items L2.1 (difficulty measure = -3.45 logits) and L2.2 (difficulty measure = -3.22 logits), respectively. All in all, the majority of participants behaved in accordance with the Rasch model's expectations. Similarly, Participant 51310's

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

(ability measure = 2.01 logits) unexpected erratic performance on Item S9.1, (difficulty measure = -3.83 logits) contributed to its large outfit estimate.

Table 1: Person Statistics for the Combined WM 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	1%	96%	3%	0%
	Outfit MNSQ	20%	67%	6%	7%

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

The item infit and outfit MNSQ estimates were similarly examined (see Table 2). This examination demonstrated evidence for notable fit with all of the infit MNSQ coefficients being inside the 0.50 to 1.50 range. The outfit MNSQ coefficients of 60 items (75%) also met the criterion. Only two misfitting items (S3.2 and S9.1) were observed with outfit MNSQ coefficients of 2.83 and 2.43, respectively. An inspection of the item individual responses revealed that the unexpected failure on these items by a few participants with high WM spans was the cause of the high outfit MNSQ values. For example, Participant 51107, who had the highest WM capacity with an ability measure of 2.74 logits, unexpectedly failed on Item S3.2, which had an estimated difficulty measure of -3.54 logits, causing the item misfit.

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

Table 2: Item Statistics for Combined WM Measure

Item	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
S1.1	-7.72	1.83		[a] MINIMUM MEASURE		
S1.2	-7.72	1.83		MINIMUM MEASURE		
S2.1	-7.72	1.83		MINIMUM MEASURE		
S2.2	-4.67	0.42	1.00	0.10	1.35	0.80
S3.1	-7.72	1.83		MINIMUM MEASURE		
S3.2	-3.54	0.27	1.06	0.40	2.83	3.60
S3.3	-2.78	0.20	1.07	0.60	1.87	2.90
S4.1	-6.51	1.00	0.96	0.30	0.13	-1.70
S4.2	-3.29	0.24	1.02	0.20	0.91	-0.20
S4.3	-2.63	0.19	1.03	0.30	1.04	0.30
S5.1	-5.39	0.58	0.99	0.20	1.06	0.30
S5.2	-2.22	0.17	1.03	0.40	1.04	0.30
S5.3	-0.31	0.13	1.01	0.20	1.01	0.20
S5.4	0.11	0.13	0.96	-0.80	0.95	-0.60
S6.1	-7.72	1.83		MINIMUM MEASURE		
S6.2	-1.98	0.16	1.05	0.60	0.99	0.00
S6.3	-0.09	0.13	1.04	0.80	1.16	2.00
S6.4	0.49	0.14	0.99	-0.20	0.99	-0.10
S7.1	-4.24	0.35	1.03	0.20	1.45	1.00
S7.2	-1.13	0.14	1.05	0.90	1.08	0.80
S7.3	0.34	0.14	1.04	0.80	1.07	0.70
S7.4	1.97	0.19	0.93	-0.50	0.94	-0.20
S7.5	2.94	0.27	1.02	0.20	0.75	-0.60
S8.1	-3.28	0.24	1.02	0.20	0.98	0.00
S8.2	-0.99	0.14	1.14	2.60	1.19	2.00
S8.3	1.46	0.17	1.08	0.80	1.22	1.20
S8.4	2.67	0.25	0.97	-0.10	0.78	-0.60
S8.5	4.24	0.48	1.06	0.30	0.55	-0.60
S9.1	-3.83	0.31	1.10	0.50	2.43	2.70
S9.2	-0.74	0.14	1.15	2.90	1.25	2.80
S9.3	1.46	0.17	1.07	0.70	1.21	1.20
S9.4	3.27	0.31	0.92	-0.20	0.90	-0.10
S9.5	5.91	1.02	1.02	0.30	0.46	-0.70
S9.6	7.13	1.83		[b] MAXIMUM MEASURE		
S10.1	-4.23	0.35	1.03	0.20	0.94	0.00
S10.2	-0.05	0.14	1.16	3.20	1.27	3.30
S10.3	1.03	0.16	1.14	1.70	1.37	2.50
S10.4	2.67	0.25	1.06	0.40	1.13	0.50
S10.5	4.78	0.59	0.92	0.00	1.30	0.60
S10.6	7.13	1.83		MAXIMUM MEASURE		
L1.1	-5.08	0.51	0.99	0.10	0.67	-0.30
L1.2	-4.22	0.35	0.87	-0.40	0.45	-1.30
L2.1	-3.45	0.25	1.01	0.10	1.97	2.30
L2.2	-3.22	0.23	0.99	0.00	1.81	2.20
L3.1	-2.34	0.18	0.94	-0.60	0.88	-0.50
L3.2	-0.66	0.13	0.92	1.70	0.87	-1.70
L3.3	-0.42	0.13	0.95	1.10	0.90	-1.30
L4.1	-1.54	0.15	0.95	-0.80	0.88	-0.90
L4.2	-0.93	0.14	0.96	-0.80	0.92	-0.90
L4.3	0.00	0.13	0.90	2.10	0.85	-2.00
L5.1	-0.71	0.14	1.09	1.80	1.16	1.80
L5.2	0.77	0.15	1.00	0.00	1.03	0.30
L5.3	1.86	0.19	0.91	-0.60	0.78	-1.00
L5.4	2.10	0.20	0.87	-0.90	0.66	-1.40
L6.1	-2.28	0.17	0.94	-0.60	0.99	0.00
L6.2	-0.52	0.13	0.96	-0.70	1.04	0.50
L6.3	0.49	0.14	0.94	-0.90	0.87	-1.30
L6.4	1.23	0.16	0.97	-0.30	0.86	-0.90
L7.1	-0.64	0.13	1.03	0.70	1.02	0.30

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

L7.2	0.64	0.14	0.97	-0.50	0.92	-0.70
L7.3	1.60	0.18	0.90	-0.90	0.74	-1.40
L7.4	3.06	0.29	0.88	-0.50	0.60	-1.00
L7.5	3.34	0.32	0.84	-0.50	0.32	-1.90
L8.1	-1.06	0.14	0.96	-0.60	0.92	-0.70
L8.2	0.24	0.14	0.97	-0.60	0.95	-0.60
L8.3	1.31	0.16	0.96	-0.40	0.85	-0.90
L8.4	2.10	0.20	0.98	-0.10	0.76	-1.00
L8.5	2.64	0.25	0.93	-0.30	0.66	-1.10
L9.1	-1.41	0.14	0.96	-0.60	0.91	-0.70
L9.2	-0.40	0.13	1.00	0.00	0.97	-0.40
L9.3	1.54	0.17	0.97	-0.30	0.77	-1.30
L9.4	2.58	0.24	0.90	-0.50	0.63	-1.20
L9.5	4.21	0.46	0.99	0.10	0.34	-1.20
L9.6	5.18	0.72	0.97	0.20	0.28	-1.10
L10.1	0.13	0.13	1.02	0.50	1.00	0.00
L10.2	0.15	0.15	1.06	0.80	1.08	0.70
L10.3	0.23	0.23	1.01	0.10	0.75	-0.80
L10.4	0.35	0.35	0.93	-0.10	0.56	-0.80
L10.5	1.01	1.01	1.01	0.30	0.34	-1.00
L10.6	1.83	1.83		MAXIMUM MEASURE		

Note. $N = 290$. Items are sorted in entry order. Items are labeled L and S, which refer to the listening and speaking span tasks, followed by the set number and the item number after a period. MNSQ = mean-square, ZSTD = Standardized z-scores.
[a] MINIMUM MEASURE = extreme minimum score.
[b] MAXIMUM MEASURE = Extreme maximum score.

In examining the table, Items S1.2, S2.1, S3.1, and S6.1 were identified as minimum measures and Items S9.6 and S10.6 as maximum measures. That is, Items S1.2, S2.1, S3.1, and S6.1 were answered correctly by the entire sample. On the contrary, no participant succeeded on Items S9.6 and S10.6. This finding corroborates the intuitions gained from looking at the targeting in the Wright map, where these items with minimum measures were identified as being too easy for the sample and the items with maximum measures as being too difficult, respectively.

Reliability and Separation

Following the checks of person and item fit, I examined the Rasch person and item reliability and separation estimates. The person reliability estimate (.87) indicated good reliability (Fisher, 2007), which was supported by a good separation index of 2.56 (Duncan et al., 2003). These results suggested that the replicability of the person ordering across other items measuring WM was high and that the measure separated the persons into three distinct WM spans: high, average, and low.

Furthermore, the Rasch item reliability (.99) and separation (8.47) estimates were excellent on the basis of the adopted criteria (Fisher, 2007; Duncan et al., 2003). These findings indicated that the probability of obtaining a similar hierarchy of item difficulty if these WM span tasks were given to a sample of comparable ability was high and that the items could be separated into eight distinct levels of difficulty.

Rasch PCA of Item Residuals

I next conducted an inspection of dimensionality via the Rasch PCA of item residuals to support the hypothesized underlying EWM construct. The findings are presented in Table 3, from which it can be seen that the raw variance explained by the measures (variance = 54.9%, eigenvalue = 87.7) was above the recommended minimum of 20% (Reckase, 1979). This

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

positive result was muddied by a potential off-dimension cluster of items as indicated by the high eigenvalue of the first contrast (3.53), which exceeded the cut-off value of 2.00 (Linacre, 2018b). However, as the first contrast (variance = 2.2%) explained less than 10% of the variance (Linacre, 2007), I consulted the standardized residual loadings in the Winsteps output and found that the cause of the high eigenvalue was task format (i.e., speaking span task vs. listening span task) rather than a second dimension. Table 4 illustrates how the listening span task items, represented by an L, load positively onto the measured construct whereas the speaking span task items, represented by an S, load negatively.

Table 3: Standardized Residuals in Eigenvalues for the Combined EWM Measure

	Eigenvalue	Observed	Expected
Total raw variance in observations	159.69	100%	100%
Raw variance explained by measures	87.70	54.9%	55%
Raw variance explained by persons	20.65	12.9%	13%
Raw variance explained by items	67.04	42%	42.1%
Raw unexplained variance (total)	72.00	45.1%	45%
Unexplained variance in 1st contrast	3.53	2.2%	—

Note. $N = 290$.

Table 4: Standardized Residual Loadings for the Combined EWM Measure

Loading	Item	Loading	Item
.45	L8.3	-.39	S10.3
.43	L8.4	-.38	S7.3
.42	L5.4	-.37	S7.4
.41	L8.5	-.30	S10.2
.40	L5.3	-.29	S10.4

Note. $N = 290$. Items are labeled L and S, which refer to the listening and speaking span tasks, followed by the set number and the item number after a period. Table reports factor loadings above .40 and below -.40 logits.

Variable Pathway

In addition to the PCA of item residuals, Winsteps provides a visual tool to graphically explore whether the items in a test adhere to the measurement of a unidimensional construct, the variable pathway or the fit map. The map displays a path delimited by two solid lines with a dotted line in the center, which represents the measured construct. The items represented by asterisks are spread along the path vertically. Items within the boundaries of the path are thought to assess the same single construct whereas items outside the boundaries indicate multidimensionality (Bond et al., 2021). Therefore, a visual inspection of the variable map supplements the numerical evidence from the PCA when assessing unidimensionality. It should be noted that although the map shows two different pathways, infit MNSQ and outfit MNSQ, I reported the pathway for infit MNSQ because in this investigation, decisions about fit were made primarily based on infit MNSQ.

Additional evidence to support the unidimensionality of the measure is provided by the item fit graph. As illustrated by the infit mean-square part of Figure 2, the items lie along the center dotted line, which represents the measure construct, and no item was outside the boundaries (i.e., the solid vertical lines) of the unidimensional path. These locations indicated that the items contributed to the measurement of a unidimensional construct.

International Journal of Psychology
ISSN 2957-6881 (Online)
Vol 11, Issue 1, No. 2, pp 16 - 35, 2026
www.iprjb.org

DISCUSSION

The purpose of this study was to evaluate the psychometric properties of a newly developed WM measure that combines the listening and the speaking span tasks while addressing the flaws of previous complex span tasks. The results of the study provide validity evidence for the use of the combined WM measure.

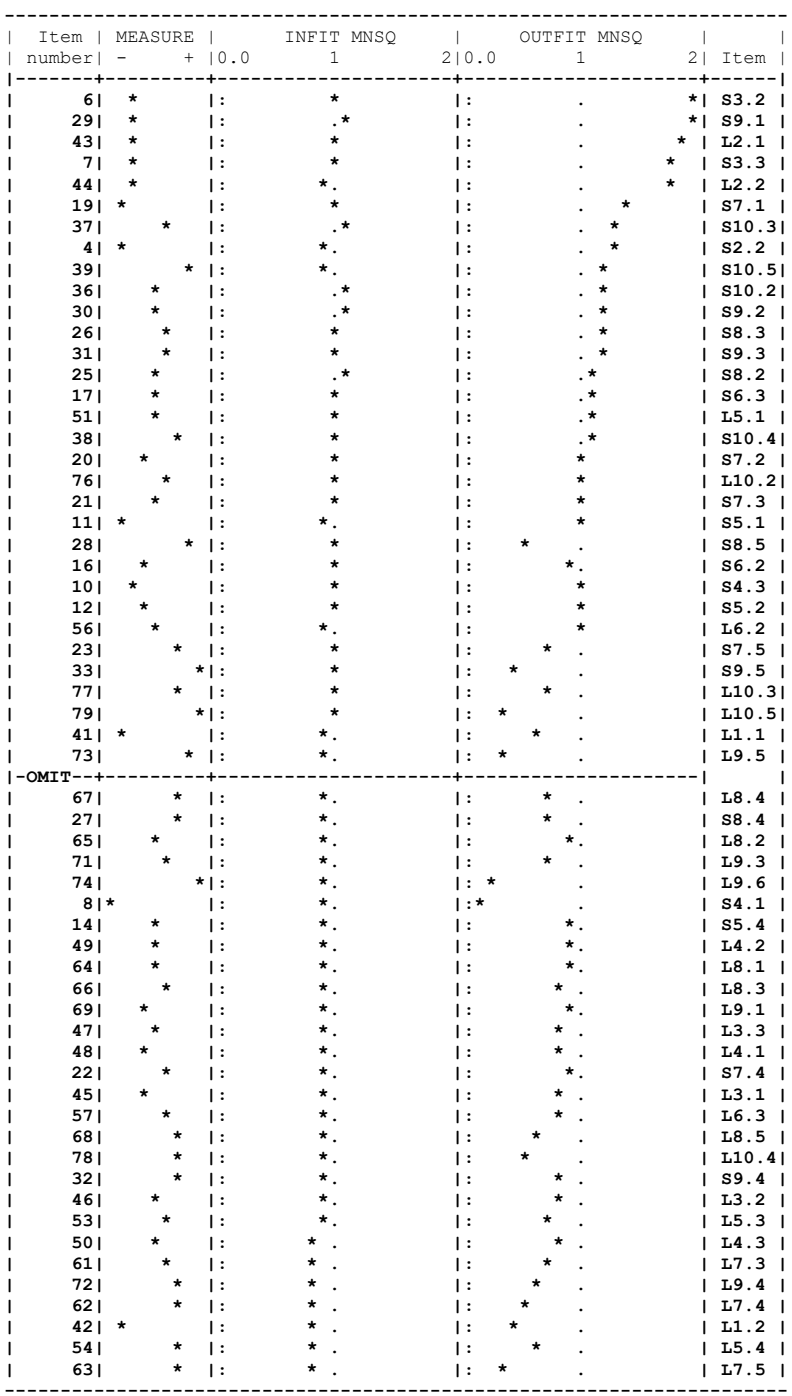


Figure 2. Item Fit Graph for the Combined EWM Measure. *N* = 290. Items are labeled L and S, which refer to the listening and speaking span tasks, followed by the set number and the item number after a period.

RQ2 asked whether the persons fitted the Rasch model. Most participants (96% of the sample) showed infit MNSQ values within the acceptable range (0.50-1.50), indicating that they performed in accordance to the expectations of the Rasch model. The proportion of persons with inflated outfit was small (7%) and was likely due to unexpected errors that participants with high WM spans made on a few easy items. Such errors have been observed in the performance of previous complex span tasks and may be caused by momentarily lapses in attention or fatigue (Bazan, 2020; Howieson & Lezak, 2012). An alternative explanation is that this small subset of misfitting cases represents participants who did not fully understand the tasks requirements. Although explicit instructions were given in written and oral form and the participants watched me perform a set of 2 items for each task, no practice trials were included on any of the tasks. Future studies may benefit from providing practice trials with sets of different length, in which the participants are closely monitored and feedback is provided in order to guarantee a complete understanding of the task requirements. In any case, no participant had infit MNSQ values above the 2.00 cut-off, which would have distorted the measurement system (Linacre, 2007). All in all, these results support the use of the combined measure.

RQ3 examined the degree of fit of the items. The item fit statistics demonstrated notable fit with all infit MNSQ values falling within the adopted criteria for fit and 75% of the outfit MNSQ also being within the acceptable range. These findings indicate that the items contribute to the measurement of the construct. There were only two items (S3.2 and S9.1), which displayed poor fit with outfit MNSQ coefficients of 2.83 and 2.43, respectively. However, this large misfit was again caused by the unexpected lack of success of a number of participants with high WM spans on these items, which were estimated to be easy with difficulty measures of -3.54 and -3.83 logits, respectively. However, given the general good item fit, these misfitting items are unlikely to compromise the estimations of person ability or the validity of the combined measure.

RQ4 and RQ5 concerned person reliability and separation, respectively. The person reliability estimate indicated good reliability (.87), which was accompanied by an equally good person separation index of 2.56, suggesting that the ordering of persons if they were given a similar WM span task was highly replicable and that the measure separated the participants into three WM spans; high, average, and low. These findings provide evidence supporting the idea that combining the listening and the speaking span tasks is a useful way to increase measurement precision by expanding the item pool. They also address a major weakness of single-task measurement tools, which may be confounded by the level of expertise in the particular domain of performance (Dehn, 2008), such as verbal fluency.

RQ 6 and RQ7 related to the item reliability and separation, respectively. In addition to the good person reliability and separation, the item reliability (.99) and item separation (8.47) estimates were excellent, indicating a high level of replicability of the item ordering across similar samples and eight different difficulty strata within the measure, respectively. These findings provide further support for the use of the measure because they ensure that difficulty estimates are not sample-dependent. Therefore, the task could be administered to other samples with confidence that the ordering of item difficulty would remain consistent.

RQ8 targeted the unidimensionality of the measure. The results of the PCA of item residuals and the examination of the variable pathway indicate that the combined WM measure assesses a single unidimensional construct. This finding suggests that the combined measure

overshadows level of expertise in the modality in which each task is performed (i.e., speaking or listening), supporting the approach of integrating the two tasks into a single measure.

CONCLUSION

In this study, I sought to validate a newly developed measure that combined the listening and the speaking span tasks using the Rasch model. The findings provide convincing evidence for the use of the measure as a psychometrically sound WM instrument. The evidence also demonstrates that combining the listening and the speaking span task through the Rasch model is a promising approach to improving WM assessment. The combined offers a generalizable, reliable, and theoretically coherent approach for estimating WM capacity and represents a step forward toward more precise WM measurement. It is hoped that the combined approach presented in this study can be adapted by researchers to mitigate the influence of domain-specific skills on WM performance and further improve the measurement of WM capacity.

Implications

From a practical point of view, the combined measure strengthens the measurement of WM capacity by reducing domain-specific variance, thus supporting a more accurate identification of individual differences than a single complex span task. This allows psychologists and teachers to devise interventions and programs that are suitably adapted to the individual's cognitive capacity.

Dedication

This article is dedicated to my advisor Dr. James Sick, who suddenly passed away on January 6, 2026. What I know about methodology and Rasch analysis in particular, comes from him. Though he could have been far more widely known, he chose to invest his time in guiding students rather than in his own research. I hope that his ideas, which profoundly shaped my work, live on through this article.

REFERENCES

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford University Press.

Bayliss, D. M., Jarrold, C., Baddeley, A. D., & Gunn, D. M. (2005). The relationship between short-term memory and working memory: complex span made simple?. *Memory*, *13*(3–4), 414–421. https://doi.org/10.1080/09658210344000332

Bazan, B. (2020). A Rasch-validation study of a novel speaking span task. *Shiken*, *24*(1), 1–21. Retrieved from http://teval.jalt.org/node/95

Bazan, B. (2021). The construction and validation of a new listening span task. *Shiken*, *25*(1), 39–56. https://doi.org/10.37546/JALTSIG.TEVAL25.1-4

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

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Erlbaum.

Conway, A. R. A., Kane, M. J., Buntig M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, 12(5), 769–786. https://doi.org/10.3758/BF03196772

Cowan N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin and review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior, 19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25*(1), 1–18. https://doi.org/10.1016/0749-596X(86)90018-5

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review, 3*(4), 422–433. https://doi.org/10.3758/BF03214546

Dehn, M. J. (2008). *Working memory and academic learning: Assessment and intervention.* John Wiley & Sons, Inc.

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

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology*, *128*(3), 309–331. https://doi.org/10.1037//0096-3445.128.3.309

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

Howieson, D. B., & Lezak, M. D. (2012). Separating memory from other cognitive disorders. In A. D. Baddeley, M. D. Kopelman, M. D., & B. A. Wilson (Eds.), *The handbook of memory disorders* (2nd ed., pp. 637–654). Wiley.

Ivanova, M. V., & Hallowell, B. (2014). A new modified listening span task to enhance validity of working memory assessment for people with and without aphasia. *Journal of Communication Disorders*, *52*, 78–98. https://doi.org/10.1016/j.jcomdis.2014.06.001

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, *12*(2), 636.

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

Linacre, J. M. (2018a). WINSTEPS (Version 4.3.1) [Computer Software]. Winsteps.com.

Linacre, J. M. (2018b). *Winsteps Rasch measurement computer program user's guide*. Beaverton.

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology*, *130*(2), 163–168. https://doi.org/10.1037/0096-3445.130.2.163

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A.H., Howerter, A, Wager, T. D. (2000) The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology*. *41*(1), 49-100. https://doi:10.1006/cogp.1999.0734

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Monteiro, F., Nascimento, L. B., Leitão, J. A., Santos, E. J. R., Rodrigues, P., Santos, I. M., Simões, F., & Nascimento, C. S. (2025). Optimizing working memory assessment: development of shortened versions of complex spans, updating, and binding tasks. *Psychological Research*, *89*(2), 65. https://doi.org/10.1007/s00426-025-02083-7

Munakata, Y., Morton, J.B., & O'Reilly, R.C. (2007). Variation in working memory due to typical and atypical development. In A. R. A Conway, C. Jarrold, M. J. Kane, M. Akira, & J. N. Towse (Eds.), *Variation in working memory* (pp. 162–193). Oxford University Press.

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343–1355. https://doi.org/10.3758/s13428-014-0543-2

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230. https://doi.org/10.2307/1164671

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in psychology*, *5*, 1475. https://doi.org/10.3389/fpsyg.2014.01475

Ünal, G., Özge, D. and Marinis, T. (2020). Assessing complex working memory in Turkish-speaking children: The listening span task adaptation into Turkish. *Frontiers in Psychology*, *11*, 1–6. https://doi.org/10.3389/fpsyg.2020.01688

Wilhelm O, Hildebrandt A, Oberauer K. (2013) What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*(433), 1–22. https://doi:10.3389/fpsyg.2013.00433

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, *3*(1), 3–24. https://doi.org/10.1080/10705519609540026

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA.