

---

Title	A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research
Author(s)	Vahid Aryadoust, Li Ying Ng and Hiroki Sayama
Source	<i>Language Testing</i> , (2020)
Published by	SAGE Publications

---

Copyright © 2020 SAGE

This is the author's accepted manuscript (post-print) of a work that was accepted for publication in *Language Testing*.

**Notice:** Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source.

The final publication is also available at <https://doi.org/10.1177/0265532220927487>

## **A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research**

### **Abstract**

Over the past decades, the application of Rasch measurement in language assessment has gradually increased. In the present study, 215 papers using Rasch measurement published in 21 applied linguistics journals were reviewed and coded for multiple features. It was found that seven Rasch models and 23 software packages were adopted in these papers, with many-facet Rasch measurement (n=100) and *Facets* (n=113) being the most frequently used Rasch model and software, respectively. Significant differences were detected between the number of papers that applied Rasch measurement to different language skills and components, with writing (n=63) and grammar (n=12) being the most and least frequently investigated, respectively. In addition, significant differences were found between the number of papers reporting person separation (n=73, not reported: n=142) and item separation (n=59, not reported: n=156) and those that did not. An alarming finding was how few papers reported unidimensionality check (n=57 vs 158) and local independence (n=19 vs 196). Finally, a multilayer network analysis revealed that research involving Rasch measurement has created two major discrete communities of practice (clusters), which can be characterized by features such as language skills, the Rasch models used, and the reporting of item reliability/separation vs person reliability/separation. Cluster 1 was accordingly labelled the production and performance cluster, whereas cluster 2 was labelled the perception and language elements cluster. Finally, guidelines and recommendations for analyzing unidimensionality, local independence, data-to-model fit, and reliability in Rasch model analysis are proposed.

**Keywords:** fit; language assessment; local independence; networks analysis; modularity maximization method; Rasch measurement; reliability and separation; unidimensionality

## Introduction

Rasch measurement refers to a family of probabilistic models that are used to predict the outcome of encounters between persons and assessment/survey items (Fischer, 1995; Rasch, 1960/1980; Wright & Stone, 1979). Rasch (1960/1980) conceptualized the basic Rasch model for tests comprising of dichotomous items measuring one latent attribute (Wright & Stone, 1979). The basic Rasch model was gradually extended to parameterize polytomous scales. The resultant models were named the rating scale model (Andersen, 1977; Andrich, 1978<sup>1</sup>; Wright & Masters, 1982) and the partial credit model (Masters, 1982; Wright & Masters, 1982). These models are applicable to polytomous scoring systems wherein “one or more intermediate levels of performance on an item [are identified] and [...] partial credit [is awarded] for reaching these intermediate levels” (Wright & Masters, 1982, p. 40). Examples of polytomous scales include Likert scales that are widely used in surveys and self-appraisals.

Next emerged many-facet Rasch measurement (MFRM), which was formulated by Linacre (1994) to accommodate different facets or variables that exert an influence on the probability of persons receiving a particular score on test items from judges or raters. This model was readily adopted in language assessment (Eckes, 2015; McNamara, 1991, 1996) and is now considered one of the most useful validation tools in studies of rater effects and bias in performance assessments (e.g., Batty, 2014; Engelhard, 2013). The unidimensional Rasch models were then extended to multidimensional models (Ackerman, 1994; Embretson, 1991; Wang, Wilson, & Adams, 1997; Wu, Adams, & Wilson, 1998) and merged with latent class models to develop the mixture Rasch models (Rost, 1991; von Davier, 1996). Fischer and Molenaar (1995) discussed other developments of the model that may be less familiar to the language assessment community (see Fischer & Molenaar, 1995).

McNamara and Knoch (2012) recognized the general importance of Rasch measurement and to the field of language testing in particular, extensively discussing the adoption and growth of Rasch measurement from 1984 to 2009 with a focus on research in the US, the UK, Australia, and the Netherlands. Rasch measurement rarely came into the picture until the collaboration between Georg Rasch and Benjamin Wright, an American advocate of Rasch’s method. Subsequently, the start of annual courses and conferences on the theory and practices of Rasch

---

<sup>1</sup> Rasch (1963) proposed a polytomous conceptualization of the model, which seems to be less known to the community.

measurement further propagated interest in the method (see Aryadoust, Tan, & Ng, 2019). According to McNamara and Knoch (2012), these events spurred interest among psychometric experts in major centers, such as the Central Institute for Test Development (CITO) in the Netherlands, the Educational Testing Service (ETS) in the US, the National Foundation of Educational Research (NFER) in the UK, and the Australian Council for Educational Research (ACER), and Rasch measurement began to be adopted on a wider scale.

The underlying idea of Rasch measurement is that variation in test takers' performances is caused by a latent attribute or trait that the test sets out to measure (see Linacre, 2005). Since this is also the assumption of item response theory (IRT) models, many have presumed that Rasch measurement is a special case of IRT. However, in debates between IRT and Rasch measurement scholars, the Rasch model was presented as a prescriptive model that demands a data-to-model fit (Bond & Fox, 2015) whereas IRT models are descriptive and require a model-to-data fit. In a correspondence with Benjamin Wright in 1965, Fred Lord wrote that "Rasch's model for unspeeeded tests [the Rasch dichotomous model] can be considered as a special case of the normal-ogive model, as Rasch himself points out extremely briefly at the end of Section 8 of his Chapter VII" (Lord & Wright, 2010, p. 1289), where the normal-ogive model refers to IRT models. In response, Benjamin Wright stressed that the Rasch model is fundamentally different from IRT, stating "I think he [Georg Rasch] would be horrified to learn that you regard his model as a special case of the normal-ogive model. The special feature of his model is that it allows for separating parameters of objects and agents, that is of children and test items. This is not possible with the normal-ogive model [...] the Rasch item analysis model is the only model which retains parameter separability. From Rasch's point of view this separability is a sine qua non for objective measurement" (Lord & Wright, 2010, p. 1289). Similarly, the differences between Rasch measurement and IRT has piqued the interest of language assessment researchers (see Holster & Lake, 2016, and Stewart, McLean, & Kramer, 2017).

Rasch measurement was not immediately well received by the language assessment community. To some scholars, the unidimensionality assumption (and, to a lesser extent, local independence) renders Rasch models inappropriate due to the complexities of language proficiency and psychological constructs (Buck, 1994). Unidimensionality refers to the assumption that the test measures only one underlying latent trait, while local independence means that, after conditioning for the latent trait, performance on one test item does not covary with

performance on other items (Borsboom, 2005). In language assessment, the assumption of unidimensionality was regarded as too stringent (e.g., Buck, 1994). Nevertheless, advocates of Rasch measurement—such as McNamara (1991, 1996)—argued that unidimensionality in Rasch measurement is a psychometric property and that a psychometrically unidimensional test can incorporate varied psychological dimensions.

As discussed earlier, the growth of Rasch measurement over the years has led to the development and application of more sophisticated models such as MFRM (Linacre, 1994). Specifically, MFRM allows for a shift away from dichotomous scored tests and provides a paramount tool to examine various rater characteristics, such as their severity and leniency, consistency, and the influence of rater training and professional background (Engelhard, 2013). It also allows examination of other aspects of the rating situation, such as the mode of test delivery, and the effect of interactions of various factors on the quality of the eventual assessment (Myford & Wolfe, 2003, 2004). Multidimensional and mixture Rasch models have also been adopted in language research to examine research problems like the separability of reading competencies (Baghaei, Kemper, Reichert, & Greiff, 2019; Min & He, 2014). With the need to cope with increasing complexities of measurement have come new and specialized Rasch-based programs such as *Facets* (Linacre, 2019b), *Winsteps* (Linacre, 2019a), *Winmira* (von Davier, 2001), *RUMM* (Andrich, Sheridan, & Luo, 2009), and *ConQuest* (Adams, Wu, & Wilson, 2015), to name a few.

### **Research Gap**

Since its emergence, Rasch measurement has made significant contributions to diverse fields of research (Aryadoust et al., 2019). However, there has not been any comprehensive review of the application of Rasch measurement in language assessment. A comprehensive review is defined as an “attempt to integrate empirical research for the purpose of creating generalizations” (Cooper & Hedges, 2009, p. 6). According to Stratton (2016), the quality of a comprehensive review hinges on several criteria: (i) having clearly defined research question(s) to be answered by the review, (ii) inclusion and exclusion criteria that are clear and objective in relation to the research question, and (iii) a conclusion that is based on what the data objectively show. Accordingly, while the review by McNamara and Knoch (2012) is extensive, it has several limitations. First, the article is a historical account of the uptake of Rasch measurement in language assessment and, as such, the results do not have quantitative generalizability. Second, the paper reviewed studies published in few journals, whereas Rasch measurement has been adopted in a number of language assessment

and validation studies published in a wide array of peer-reviewed journals (e.g., Yamashita & Shiotsu, 2017). Third, McNamara and Knoch (2012) did not investigate whether the reviewed studies verified the requirements of Rasch measurement, which include data-to-model fit, unidimensionality, and local independence (Bond & Fox, 2015).

### **Theoretical Framework**

Several studies have established specific frameworks for Rasch model analysis in different research fields such as language assessment (e.g., Eckes, 2015; McNamara, 1996), educational measurement (Bond & Fox, 2015; Boone, Staver, & Yale, 2014; Engelhard, 2013; Salzberger, 2012; Smith, Linacre, & Smith, 2003), and validation in medicine (Tennant & Conaghan, 2007). In this study, we adapted Wright and Stone's (1999) comprehensive framework which comprises of (i) metrics of psychometric validity, consisting of unidimensionality, local independence, and fit statistics, and (ii) metrics of reliability, consisting of reliability and separation coefficients for items and persons. For clarity, we briefly review these concepts in this section in the following order: unidimensionality, local independence, reliability, and fit.

First, unidimensionality refers to whether a test measures the delineated latent trait that it purports to measure and not unintended constructs (Wright & Stone, 1999). Several methods are used to investigate unidimensionality, the most common of which in language assessment are principal component analysis of residuals (PCAR) and factor or principal component analysis (EFA/PCA) (see Hattie, 1985, for an in-depth review of relevant methods). Residuals are the discrepancies between the observed data and the data expected by the Rasch model. PCAR investigates whether there are any significant and substantive secondary structures in the residuals, whereas EFA and PCA investigate the underlying structure of the raw data. Secondary dimensions create either auxiliary or nuisance dimensions, with the former being relevant to the main construct under assessment and the latter causing variance in data that may adversely affect the unidimensionality of the data (Ackerman, Gierl, & Walker, 2003).

Second, in Rasch measurement, items are regressed on the latent variable; therefore, it is essential that unexplained variances in the items do not correlate with each other (Borsboom, 2005). This is called local independence and it is determined by assessing the correlations between the residuals of the test items (also known as the Q3 coefficient) (Fan & Bond, 2019; Lee, 2004; Wright, 1996a; Yen, 1984). A Q3 coefficient larger than  $|\cdot 3|$  indicates a respectable degree of local dependence. The investigation of local independence in Rasch measurement is analogous to the

investigation of multicollinearity in linear regression. There are several methods for examining local independence in IRT research, including the  $G_2$  statistic (Chen & Thissen, 1997),  $\chi^2$  statistic (Chen & Thissen, 1997), and Cramer's V statistic (Baldonado et al., 2015). Despite their popularity in IRT, these methods have not been widely adopted in Rasch measurement publications.

Third, reliability is a necessary, but insufficient, criterion to assess the quality of measurement. Reliability indicates the reproducibility of the item measures if the items were administered to another sample drawn from the same population, or the reproducibility of person measures if they were tested on another occasion (Bond & Fox, 2015). Separation, which refers to the number of statistically different levels of item difficulty or person ability in the data (Linacre, 2019a), provides another index for reliability. High separation ( $>2$ ) indicates that the test was able to differentiate between difficulty/ability groups of items/persons (Linacre, 2019a).

Fourth, infit and outfit statistics are computed based on the Rasch model residuals (Smith, Schumacker, & Bush, 1998; Smith, 2000), although as Smith et al. (1998) noted they were first formulated by Wright and Panchapakesan (1969) based on person raw scores. Infit and outfit statistics, respectively, are sensitive to on-target and off-target response patterns: Erratic responses to items located near person ability measures are identifiable by infit metrics, whereas aberrations far from person or item measures are detected by outfit metrics (Linacre, 2019a). The mean square (MnSq) index, which has an expected value of 1.00, indicates the size of the anomalies in the measurement. For example, a MnSq of 1.2 indicates 20% noise in the data whereas 1.1 indicates less distortion (Linacre, 2002a).

There is no universal agreement on fit statistics in Rasch measurement. For the MnSq metrics, liberal and stringent ranges of 0.5–1.5 and 0.8–1.2, respectively, have been suggested (Linacre, 1994; 2002a). In addition, Smith et al. (1998) recommended a formulas to determine the upper bound for MnSq metrics:  $\text{infit MnSq} = 1 + \frac{2}{\sqrt{x}}$  and  $\text{outfit MnSq} = 1 + \frac{6}{\sqrt{x}}$ , where  $x$ =sample size (see also DeMars, 2017; Karabatsos, 2000; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). For standardized (Zstd) metrics, which provide a  $t$ -test, a range between -1.96 and +1.96 has been proposed (Linacre, 2002a). There are two additional methods of establishing fit that are fairly well researched: (i) the  $lz$  person-fit index, which is a likelihood-based index with a sampling distribution (Dragow, Levine, & Williams, 1985; Hulin, Dragow, & Parsons, 1983; see Armstrong, Stoumbos, Kung, and Shi, 2007, and Linacre, 1997, for critiques of the  $lz$  index), and (ii) the Rasch bootstrap fit (RBF), a computer macro for SAS software to estimate the confidence

intervals (CIs) of fit indices generated by Winsteps (Wolfe, 2008, 2013; see also Baghaei & Aryadoust, 2015; Hodge & Morgan, 2017).

In light of these various criteria for assessing measurement quality, this study aims to investigate whether, and the degree to which, these criteria were addressed in previous research in language assessment. In addition, the study aims to provide a descriptive summary of, for example, the language skills and components investigated and the different Rasch models that were used by researchers. The research questions of the study are as follows:

1. In language assessment research where Rasch measurement was used, what language skills or components did the authors investigate?
2. What Rasch models were used by the authors to fulfil their research goals?
3. What methodologies did the authors use to investigate unidimensionality and local independence?
4. What reliability coefficients did the authors report?
5. What fit statistics did the authors use to explore the quality of the data? What fit criteria were applied to interpret the fit indices?
6. From a networks system perspective, are there any distinct communities of practice that adopted Rasch measurement in language assessment? If so, what can they be characterized by?

To answer these research questions, a coding scheme was developed to code Rasch-based publications. The data were then analyzed using conventional univariate methods and networks analysis. The networks approach is a complex systems methodology that allows for the detection of communities of practice and the analysis of their distinctive properties (Freeman, 2004). It also offers the advantage of identifying patterns and influential nodes (e.g., methods, skills, or other facets of interest) in the data.

## **Methodology**

### **Dataset**

In the present study, we utilized a total of 215 studies that used Rasch models from 21 journals in language assessment and applied linguistics indexed in Scopus to generate statistical results. Scopus is the largest available database of published research (Schotten, Aisati, Meester, Steinginga, & Ross, 2018). First, we chose 21 journals in applied linguistics from the Scimago's

list of the top 100 journals in linguistics as of November 2019 (Scimago, 2018). Next, the Scopus database was used to conduct a document search using the ‘Source title’ method, which generated 441 publications (Appendix 1 in the supplemental document on the *Language Testing* website provides the Scopus search code). Next, the principal investigator (PI) and four research assistants (RAs) read the papers to identify the studies where Rasch measurement was used as a primary or secondary data analysis method. Methodological IRT papers were removed since the focus of this study was application of the method. Studies that referred to Rasch measurement as IRT, Rasch IRT, or similar terms were included in the review. After this screening process, 215 empirical studies remained.

We found that the total number of examinees across all studies was 839,837. The sample sizes of the existing studies varied largely in their nature and magnitude; sample sizes included examinees, raters, recordings, etc., and ranged from 4 raters to 14,089 examinees. As shown in Figure 1, there has been a general upwards trend since 2010 in the number of papers using Rasch measurement in language assessment, peaking at 21 in 2017. The number of papers shown for 2019 is not representative of the total number of papers published that year as the data were collected prior to the end of 2019. (Please also see Supplemental Figure A on the *Language Testing* website for a summary of the papers by language skills and country or region where the studies were carried out.)

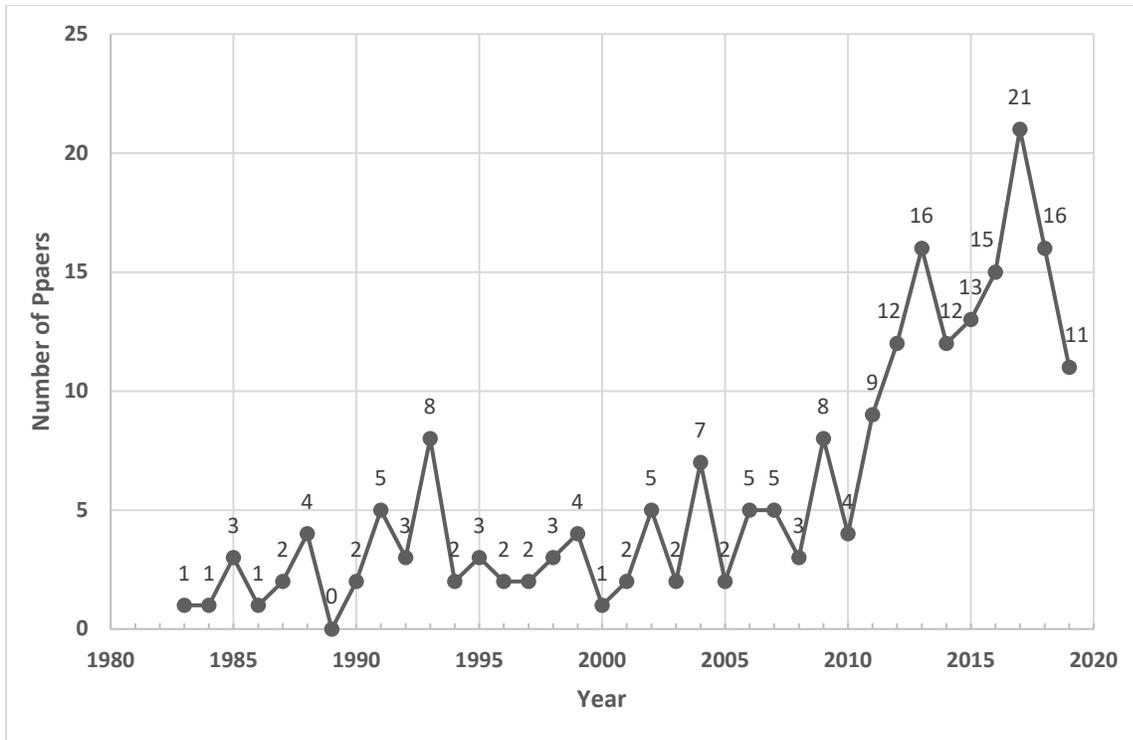


Figure 1. Line graph representing the number of papers on Rasch measurement in language assessment over the years. The number of papers each year is specified by the data points.

Table 1 summarizes various characteristics of the 215 Rasch measurement articles and the various Rasch models adopted. Overall, seven varieties were adopted: the Rasch model, many-facet Rasch measurement (MFRM), the Rasch-Andrich rating scale model, the partial credit model, the mixed Rasch model, the general polytomous Rasch model, and the general item response theory (IRT). Most Rasch measurement articles were published in *Language Testing* (n=97; 45.12%), followed by *Assessing Writing* (n=29; 13.49%), and *Language Assessment Quarterly* (n=28; 13.02%). The remaining journals accounted for smaller percentages of the Rasch measurement articles used in the present study. The Rasch model (Rasch, 1960) was the most commonly used model, appearing in articles from 16 journals, followed by the MFRM (Linacre, 1994) (n=13) and the Rasch-Andrich rating scale model (n=6) (Anderson, 1977; Andrich, 1978). The one-parameter logistic IRT model (Birnbaum, 1968) adopted in an article in *Language Testing* was considered to be interchangeable with the Rasch model (Linacre, 2005). Articles from *Language Testing* used the widest variety of Rasch models (n=7; 100.00%). By contrast, articles in *Assessment Writing*, the journal with the second-most Rasch-based studies, adopted the smallest variety of Rasch models (n=3; 42.86%) (MFRM, the Rasch Model, and the Rasch-Andrich rating scale model). This indicates that the

different proportions of articles originating from different journals have no proportional relationship with the variety of Rasch models adopted.

Table 1  
*Descriptive Statistics of the Database*

Journal	# of Rasch measurement papers	% of Rasch measurement papers	Model applied	# and (%) of the Rasch model applied
Language Testing	97	45.12	The Rasch Model, Mixed Rasch Model, MFRM, Partial Credit Model, Rasch-Andrich Rating Scale Model, Polytomous Rasch Model, IRT	7 (100.00%)
Assessing Writing	29	13.49	The Rasch Model, MFRM, Rasch-Andrich Rating Scale Model	3 (42.86%)
Language Assessment Quarterly	28	13.02	The Rasch Model, MFRM, IRT	3 (42.86%)
System	13	6.05	The Rasch Model, MFRM, Rasch-Andrich Rating Scale Model	3 (42.86%)
Language Testing in Asia	9	4.19	The Rasch Model & MFRM	2 (28.57%)
TESOL Quarterly	8	3.72	The Rasch Model, MFRM, Partial Credit Model	3 (42.86%)
Language Learning	7	3.26	The Rasch Model & MFRM	2 (28.57%)
Language Teaching Research	3	1.40	The Rasch Model & MFRM	2 (28.57%)
Modern Language Journal	3	1.40	The Rasch Model & Rasch-Andrich Rating Scale Model	2 (28.57%)
ReCALL	3	1.40	MFRM, Rasch-Andrich Rating Scale Model, Polytomous Rasch Model	3 (42.86%)
RELC Journal	3	1.40	The Rasch Model & MFRM	2 (28.57%)
Studies in Second Language Acquisition	3	1.40	The Rasch Model & MFRM	2 (28.57%)
Applied Linguistics	1	0.47	The Rasch Model	1 (14.29%)
Applied Psycholinguistics	1	0.47	The Rasch Model	1 (14.29%)
Computer Assisted Language Learning	1	0.47	Rasch-Andrich Rating Scale Model	1 (14.29%)
International Journal of Applied Linguistics	1	0.47	The Rasch Model	1 (14.29%)
Iranian Journal of Language Teaching Research	1	0.47	The Rasch Model	1 (14.29%)
Iranian Journal of Language Testing	1	0.47	MFRM	1
Journal of Second Language Writing	1	0.47	MFRM	1 (14.29%)
Language Learning in Higher Education	1	0.47	The Rasch Model	1 (14.29%)
The Language Learning Journal	1	0.47	Not reported	-
Total	215	100.00		7

*Note:* MFRM = many-facet Rasch measurement. IRT = item response theory, TESOL = Teachers of English to Speakers of Other Languages, ReCALL = Journal of the European Association for Computer Assisted Language Learning (EUROCALL), RELC = Regional Language Centre. As the author(s) either did not identify what IRT or specific polytomous Rasch model was used, or several different models were used (e.g. 1 parameter logistic (PL), 2PL, and 3PL), a generic label is used.

## Coding Scheme

*Defining the codes.* As previously described, to evaluate the publications, we developed a checklist (Table 2) informed by Wright and Stone's (1999) Rasch-based validation framework. The metrics in Table 2 were grouped as: (i) metrics of reliability, consisting of reliability and

separation coefficients for items and persons, and (ii) metrics of psychometric validity, consisting of fit statistics, unidimensionality, and local independence.

*Applying the codes.* An Excel spreadsheet was developed to code the journal papers involving Rasch measurement. This spreadsheet was circulated among the RAs from a major university in Singapore who were trained to code the data. The RAs were English native speakers, psychology majors, and high-performing university students. They were instructed to direct their questions related to uncertainty about the coding to the PI. The articles were independently coded by the PI and RAs. The inter-coder agreement coefficient was computed per variable and ranged approximately between 90% and 100%. Observed discrepancies were mainly related to the type of Rasch model adopted and sample size in cases where MFRM was used. The final codes were standardized and automatically converted to numerical codes using SPSS. The quantified variables were then used in the statistical and complex networks analyses.

Table 2  
*Criteria Used to Code the Publications*

Analysis	Description	Example reference
<i>Descriptive feature</i>		
Location	Country or region and language skill	NA
Features of the sample	Sample size	NA
Features of the instrument	Test used	NA
The Rasch model used	The Rasch Model, Many-Facet Rasch Measurement (MFRM), Rasch-Andrich Rating Scale Model, Partial Credit Model, Mixed Rasch Model, General Polytomous Rasch Model, and General Item Response Theory (IRT)	Ackerman (1994); Andrich (1978); Masters (1982); Rasch (1960/1980); Linacre (1994)
Software	Facets, Winsteps, Winmira, RUMM, ConQuest, etc.	Andrich et al. (2009); Linacre (2019a); Linacre (2019b); von Davier (2001)
<i>Reliability</i>		
Item and person reliability	<ul style="list-style-type: none"> <li>An index of precision and reproducibility of items / person measures</li> </ul>	Fisher (1992); Linacre (1995); Mallinson & Stelmack (2001); Wright (1996b);
Item and person separation	<ul style="list-style-type: none"> <li>Another representation of reliability to classify items and persons into separable clusters</li> <li>The ratio of the true standard deviation over the error of standard deviation</li> </ul>	As above
<i>Psychometric validity</i>		
Infit mean square (MnSq) and standardized (ZStd)	<ul style="list-style-type: none"> <li>MnSq: An inlier-sensitive fit index to capture anomalous response patterns on items targeted on persons and vice versa</li> <li>ZStd provides a t-test to investigate whether the data have a perfect fit to the model</li> </ul>	Linacre (2002a, b); Wright (1994a); Wright & Masters (1990)

Outfit mean square (MnSq) and standardized (ZStd)	<ul style="list-style-type: none"> <li>• MnSq: An outlier-sensitive fit index to capture anomalous responses to items far away from persons and vice versa</li> <li>• ZStd provides a <i>t</i>-test to investigate whether the data have a perfect fit to the model</li> </ul>	As above
Unidimensionality	<p>Whether the test measures one psychometric dimension. Secondary structures create either auxiliary or nuisance dimensions, with the former being relevant to the main construct under assessment and the latter one causing adverse variation in the data. Common methods to check unidimensionality are:</p> <ul style="list-style-type: none"> <li>• Principal component analysis of residuals</li> <li>• Factor or principal component analysis</li> <li>• Infit and outfit statistics</li> </ul>	Ackerman et al. (2003); Raïche (2005); Tennant & Pallant (2006); Wright (1994b);
Local independence	<p>Whether items covary after conditioning for the latent construct under investigation (i.e., the Rasch dimension). The most common method to investigate local independence is correlation analysis or Q3 index.</p>	Fan & Bond (2019); Lee (2004)

*Note:* Analysis refers to the specific criterion being investigated; description defines what the criterion is and/or provides examples of the categories in the criterion.

## Data Analysis

For research questions 1 and 2, the data were arranged by the Rasch model used per language skill and per publication to identify frequencies of model usage with reference to the language skill investigated. For research questions 3 and 4, the data were arranged by the various statistics reported per paper (unidimensionality, local independence, and reliability/separation coefficients) and sorted by language skills to identify trends of reported statistics by language skills. Where appropriate, chi-square tests were conducted to test for differences between the categories of papers. To address research question 5, a frequency analysis was carried out to investigate trends in using infit and outfit indices.

To address research question 6, community detection analysis, which is a type of networks analysis, was conducted to group the publications into distinct communities (clusters) based on how closely related the papers were to each other. Network analysis is a collection of system analytics techniques that have been developed and widely used in social sciences, physical/natural sciences, and engineering (Barabási, 2003, 2016; Easley & Kleinberg, 2010; Sayama, 2015; Wasserman & Faust, 1994). Community detection in networks analysis is an exploratory technique used to identify communities in data. Community detection is differentiated from model fitting typically adopted in statistics, and, accordingly is not founded on the notion of fit statistics. The fundamental assumption is that having more connections between papers (nodes) makes them more relevant (connected) to each other, which is the basis of all community detection algorithms (Fortunato & Hric, 2016).

In the present study, the data were organized into a bipartite (two-mode) network by connecting each paper to the specific properties of the Rasch method used in that paper. Then, this bipartite network was projected to a weighted unipartite (one-mode) network made of papers, by using the number of attributes shared by two papers as the weight of the connection between them. Communities in this weighed network of papers were detected using the two most popular methods: (1) the modularity maximization method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Newman & Girvan, 2004) that heuristically finds the best way to split the network into multiple communities such that the "modularity" metric (Newman, 2006) of the network becomes maximal, and (2) the spectral partitioning method (Chung, 1997; von Luxburg, 2006) that partitions the network into multiple communities based on the eigenvalues and eigenvectors of the network's Laplacian matrix (Sayama, 2015). The above-mentioned network modeling, community detection, and visualization were all done in Wolfram Research Mathematica version 12.0.0. After identifying communities via these methods, we inspected the content of the publications in each community to detect distinguishing patterns.

## Results

### Research Questions 1 and 2

Supplemental Table A presents the proportions of the different language skills or components alongside the various Rasch models and the software used in the papers in this analysis (Please see the online supplement on the *Language Testing* website). The total number of Rasch papers was the count of individual unique papers. Overall, the papers investigated 15 distinguishable language skills and components, with writing (n=63), speaking (n=59), and integrated skills (n=35) being the subject of investigation in the largest numbers of papers. The category "Others" includes competency/proficiency (n=16), communication (n=2), control (n=2), teaching (n=2), memory (n=1), pragmatics (n=1), pronunciation (n=1), translation (n=1), as well as papers that did not report the skills investigated (n=2).

In addition, a total of 23 Rasch software packages were utilized in the papers, with *Facets* (n=113; writing=45; speaking=43; others=19), *Winsteps* (n=39; integrated=13; vocabulary=11; writing=7), and *ConQuest* (n=10; listening=4; vocabulary=3; reading=2) being the top three packages. Some papers (n=27) did not report the Rasch software used.

Fifty Rasch measurement papers investigated more than one language skill or component. The percentage of Rasch measurement papers studying each linguistic component in the third column of Supplemental Table A was calculated based on a total of 215 papers. The papers investigating integrated skills, speaking, vocabulary, and listening adopted the widest variety of Rasch models ( $n = 5$ ; 62.50%). A series of  $\chi^2$  tests showed that the categories of language skills and components occurred with unequal probabilities ( $\chi^2(20, n=283)=373.339, p < 0.001$ ), with writing being the most researched skill ( $n=63$ ) followed by speaking ( $n=59$ ), integrated skills ( $n=35$ ), reading ( $n=32$ ), and vocabulary ( $n=31$ ). Research on memory, pragmatics, pronunciation, and translation had the lowest number of articles involving Rasch measurement ( $n=1$ ).

The 32 papers investigating reading utilized the largest variety of Rasch software ( $n=12$ ), but adopted only four models (Rasch-Andrich rating scale model, Rasch model, MFRM, & IRT).

### **Research Question 3**

Table 3 provides a summary of the various methodologies used in the analysis of unidimensionality and local independence for the different groups of Rasch measurement papers that investigated varying language skills or components. The table also presents the varying proportions of papers in each group that checked for unidimensionality and local independence using those methodologies. The percentage proportions were calculated based on the total of 215 Rasch measurement papers. A total of 13 methodologies and criteria were used for unidimensionality analysis to varying degrees, namely fit statistics analysis, Bejar's (1980) method, factor analysis, Reckase's (1979) criteria of unidimensionality, PCA(R), reliability coefficients analysis, DIMTEST, analysis of strength, analyzing similarities of estimates of item difficulty parameters, scalability analysis, linearity and equality tests,  $t$ -value analysis, and confidence interval analysis. Fit statistics analysis was the most widely used unidimensionality method among seven groups of papers investigating different linguistic components (grammar, vocabulary, integrated skills, listening, speaking/oral, reading, and writing). The largest proportion of papers that tested for unidimensionality investigated the component of integrated skills ( $n=10$ ; 4.63%), followed by vocabulary ( $n=8$ ; 3.70%), and writing ( $n=7$ ; 3.24%).

By contrast, only six Rasch measurement papers, which studied vocabulary ( $n=2$ ; 0.93%), integrated skills ( $n=2$ ; 0.93%), listening ( $n=1$ ; 0.46%), and reading ( $n=1$ ; 0.46%), tested local independence with reported methodologies. There was a significant difference between the papers

that reported the local independence check (n=19) and those that did not (n=196), ( $\chi^2(1, n=215)=145.716, p < 0.001$ ). Notably, no Rasch measurement papers investigating the components of syntactic complexity, writing error detection, and word derivation knowledge reported unidimensionality or local independence analysis.

Table 3  
*Investigation of Unidimensionality and Local Independence*

Language skill or component	# of papers using the unidimensionality analysis method	%	# of papers using the local independence method	%
Grammar	4	1.86	2	0.93
Vocabulary	16	7.44	5	2.33
Integrated	15	6.98	6	2.79
Listening	7	3.26	1	0.47
Speaking/Oral	9	4.19	2	0.93
Reading	9	4.19	5	2.33
Writing	14	6.51	1	0.47
Others	5	2.33	0	0.00
Total	57/215	36.74	19/215	10.23

*Note:* A paper may appear multiple times if it investigated multiple skills. The total number counts each paper only once so as to accurately represent the percentage of papers reporting these statistics.

#### Research Question 4

Table 4 presents a summary of the proportions of Rasch measurement papers that reported person reliability (PR), person separation (PS), item reliability (IR), and item separation (IS) grouped by the linguistic component investigated. Of the total of 215 papers, 109 papers (50.70%) reported PR, 73 papers (33.95%) reported PS, 109 papers (50.70%) reported IR, and 59 papers (27.44%) reported IS. Papers that investigated writing constituted the largest proportion of papers reporting PR (n=39; 18.14%) and PS (n=30; 13.95%). By contrast, the largest proportions of papers reporting IR and IS investigated integrated skills (n=24; 11.11%) and speaking (n= 18; 8.37%), respectively.  $\chi^2$  tests revealed significant imbalances between papers that reported PS ( $\chi^2(1, n=215)=22.144, p < 0.001$ ) and IS ( $\chi^2(1, n=215)=43.763, p < 0.001$ ), but no significant imbalances between papers that reported PR ( $\chi^2(1, n=215)=0.042, p = 0.838$ ) and IR ( $\chi^2(1, n=215)= 0.042, p = 0.838$ ).

Table 4  
*Investigation of Rasch-based Reliability*

Language skill or component	# of papers reporting PR	%	# of papers reporting PS	%	# of papers reporting IR	%	# of papers reporting IS	%
Grammar	4	1.86	2	0.93	3	1.40	2	0.93
Vocabulary	17	7.91	5	2.33	15	6.98	9	4.19
Integrated	14	6.51	9	4.19	24	11.16	13	6.05
Listening	9	4.19	7	3.26	16	7.44	6	2.79
Speaking	31	14.42	24	11.16	23	10.70	18	8.37
Reading	14	6.51	6	2.79	20	9.30	3	1.40
Writing	39	18.14	30	13.95	23	10.70	16	7.44
Others	15	6.98	14	6.51	11	5.12	7	3.26
Total	109/215	50.70	73/215	33.95	109/215	50.70	59/215	27.44

*Note:* A paper may appear multiple times if it investigated multiple skills and presented the statistics. The total counts each paper only once so as to accurately represent the percentage of papers reporting these statistics. PR = Person reliability, PS = Person separation, IR = Item reliability, IS = Item separation.

### Research Question 5

Supplemental Table B provides a summary of the proportions of Rasch measurement papers that reported fit values within the recommended infit and outfit mean square range (MnSq, 0.5 – 1.5) and standardized range (ZStd, -1.96 – 1.96) for both items and persons. (Please see the online supplement on the *Language Testing* website). The table also includes other fit criteria ranges adopted by the papers that do not fall within the generally acceptable range and the corresponding proportions of papers that reported them. Percentage proportions were calculated based on a total number of 215 Rasch measurement papers. Papers that investigated integrated skills accounted for the largest proportion of papers reporting infit MnSq ranges (n=15; 6.94%) and outfit MnSq ranges (n=11; 5.09%) for items falling within the generally accepted range. Most papers reporting person infit and outfit MnSq ranges outside the generally accepted range investigated speaking/oral (n=20; 9.26%) and writing (n=9; 4.17%) skills, respectively. By contrast, few papers reported infit and outfit ZStd ranges between -1.96 and +1.96 for persons and items. Notably, only papers that investigated integrated skills and vocabulary reported these values. However, there was significant diversity of other MnSq and ZStd fit criteria ranges for both persons and items falling significantly outside the generally acceptable ranges.

### Research Question 6

Figure 2 presents a visualization of clusters detected using the network modularity maximization method. Each node (circle) represents a paper and the edges (connections) among nodes represent the shared properties of the methods used in the two papers they connect. The colors of the nodes (red and yellow) indicate different clusters. The grayscale coloring of the edges shows the number of shared properties of the methods. The edges connecting nodes between the 1st and

2nd clusters appear to be darker than the within-cluster connections in this figure, but this is simply because all edges originating from one node to all other nodes in the opposite cluster are superposed up to the bottleneck in the center. Within-cluster connections are actually denser than between-cluster connections.

As demonstrated in Table 5, the modularity maximization method detected two clusters (communities), which is more parsimonious than the four-cluster model generated by the spectral partitioning method. Closer inspection of the content of the clusters that emerged in both methods indicated that the two clusters resulting from modularity maximization were theoretically more sensible. Thus, the modularity maximization solution was chosen.

Significant differences were found between clusters 1 and 2 for all variables in the modularity maximization output. Cluster 1 mostly included the measurement of writing (n=49, cluster 2: n=14) and speaking (n=47, cluster 2: n=12) skills using the MFRM (n=95, cluster 2: n=5) and *Facets* (n=102, cluster 2: n=13) (Due to space constraints, Table 5 only presents the results of the chi-square tests for language skills, the Rasch models, and the software used). In comparison, cluster 2 mainly included the measurement of integrated (n=30, cluster 1: n=5), reading (n=26, cluster 1: n=6), and vocabulary (n=25, cluster 1: n=6) skills using the Rasch model (n=69, cluster 1: n=10) and the Rasch-Andrich rating scale model (n=12, cluster 1: n=1). Cluster 2 further included papers that used the widest variety of software, including all listed software except *Bigscale* and *PCRasch*. *Winsteps* (n=38, cluster 1: n=1) was used most often in cluster 2. Cluster 1 had more papers reporting PR (n=71, cluster 2: n=38) and PS (n=62, cluster 2: n=10), whereas cluster 2 included more papers reporting IR (n=70, cluster 1: n=39), unidimensionality check (UD) (n=45, cluster 1: n=12), and local independence check (LI) (n=15, cluster 1: n=4). Therefore, cluster 1 was labelled the production and performance cluster, whereas cluster 2 was labelled the perception and language elements cluster.

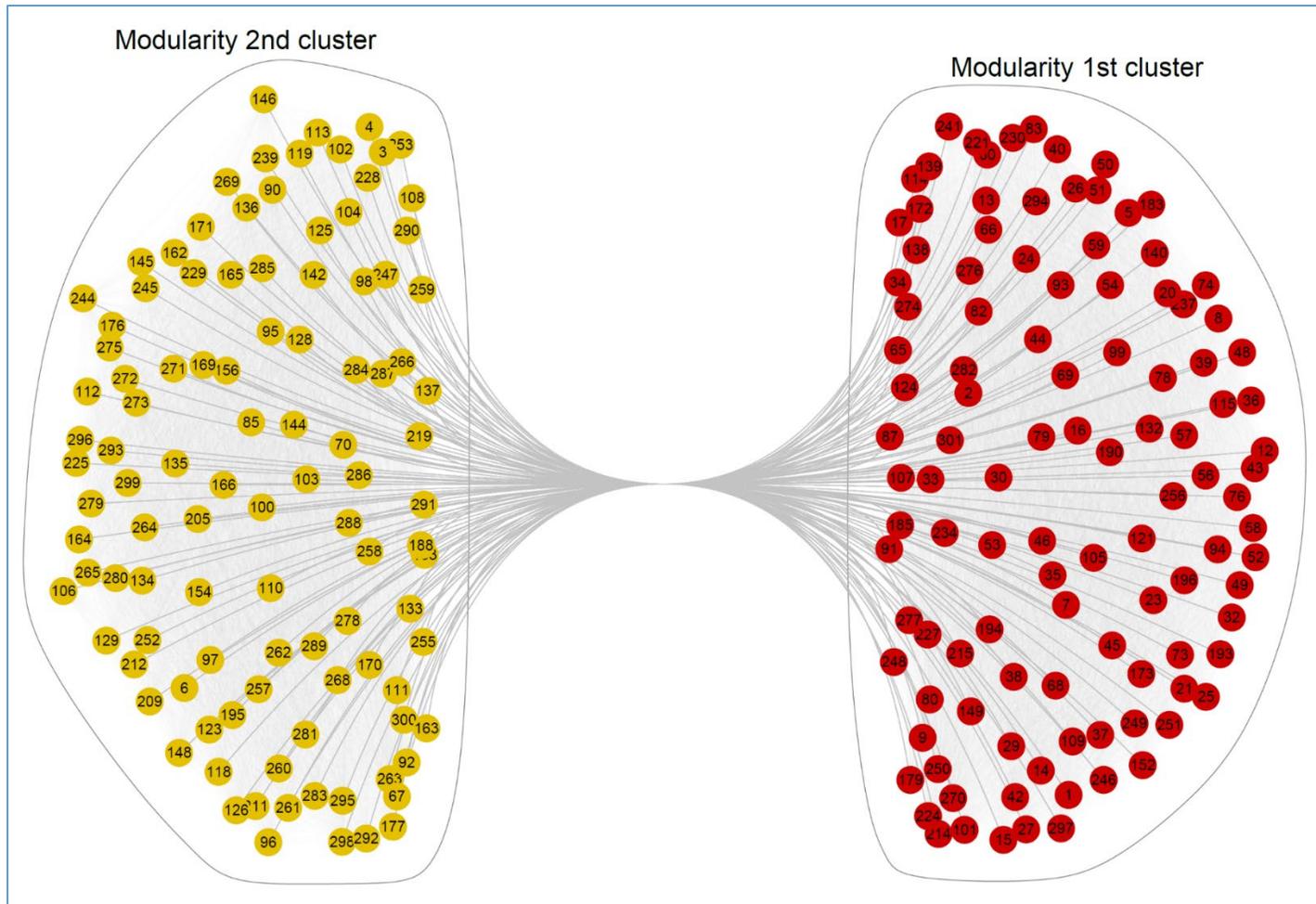
Table 5

*Statistical Tests between Clusters for Language Skills, Rasch Models, and Software*

Cluster	No.	Label	N	No. of LS	Chi-Square Statistics			$\phi_c$	No. of RM	Chi-Square Statistics			$\phi_c$	Soft- ware Used	Chi-Square Statistics			$\phi_c$
					df	N	$\chi^2$			df	n	$\chi^2$			df	n	$\chi^2$	
Modularity	1	Production and performance	110	8	7	283	89.18****	0.813	4	7	221	^146.23****	0.813	4	23	222	^151.23****	0.825
Modularity	2	Perception and language elements	105	8					7					21				
Spectral	1	-	65	8	21	283	^108.22****	0.357	2	21	221	^150.76****	0.477	3	69	222	^198.00****	0.535
Spectral	2	-	56	8					6					14				
Spectral	3	-	55	8					6					14				
Spectral	4	-	39	8					3					2				

Note: df=degrees of freedom; LS=language skill; RM= Rasch model.

^Where there were expected counts less than 5, Fischer's Exact tests were attempted but were not successful. Cramer's V ( $\phi_c$ ) statistics were provided to supplement the results. Cramer's V ranges from 0 to 1, with values closer to 1 indicating a larger effect size (Sun, Pan, & Wang, 2010). +Language Skill, ++Rasch Model, \*\*\* $p < 0.001$ , \*\*\*\* $p < 0.0001$



*Figure 2.* Visual representation of the modularity clusters. Cluster 1 was labelled the production and performance cluster, whereas cluster 2 was labelled perception and language elements. Numbers represent individual papers and lines represent the presence of shared properties between the papers. The connection density within clusters is stronger than between clusters. Analysis and visualization were performed using Wolfram Research Mathematica.

## Discussion

The primary aim of the present study was to review publications (n = 215) involving Rasch measurement in mainstream applied linguistics journals. The coding scheme used here built upon Wright and Stone's (1999) publication that discusses Rasch-based psychometric measurement in detail. We employed descriptive statistics, chi-square tests, and multivariate networks analysis to answer the research questions (RQs) of the study.

The coding scheme comprised three main sections: descriptive features, reliability analysis, and psychometric validity. Substantial variation was observed in the different facets investigated, such as the choice of Rasch models, software packages employed, and the reporting of fit metrics, unidimensionality, and local independence. It was found that Rasch measurement permeated language assessment research in all continents except Africa. The results partially resonate with McNamara and Knoch's (2012) opinion about the regional spread of Rasch measurement research in the US and Australia, with further evidence showing that the model is also extensively used in Japan (more than Australia), China, and the UK. The predominance of writing, speaking, and integrated language skills in the application of Rasch measurement was also documented. The findings related to the individual RQs of this study are discussed below.

### RQ1 and 2

The most frequently investigated language skill was writing, followed by speaking, integrated skills, and vocabulary. Research shows that the assessment of academic writing (as well as speaking) is affected by rater bias (Eckes, 2019; Engelhard, 2013; Wind & Peterson, 2018). To identify and mitigate sources of bias in writing and speaking assessments, researchers have used a variety of methods—most prominently MFRM (Linacre, 1994). This model was found to be the most frequently used method of checking for bias, which is consistent with the results of Wind and Peterson's (2018) recent review.

The most frequently used software was *Facets*, a flexible package that can accommodate a variety of unidimensional Rasch models for dichotomous and polytomous data. *Facets* is also suitable for conducting differential facet/item functioning (DIF/DFI) to investigate bias in measurement and rater effects (Eckes, 2019). Similarly, *Winsteps*, the second most frequently used software, is a package for unidimensional Rasch measurement that provides researchers with a wide range of analyses, inter alia, rating scale modeling, partial credit modeling, DIF, and dimensionality analysis via PCAR (Raquel, 2019). The dominance of these packages is attributed

to factors including their regular maintenance and updating, low cost, and developer assistance. This finding suggests that user-friendly computer applications may have had a significant role in the increasing adoption and gradual spread of unidimensional models.

Although *ConQuest* was not employed as frequently as *Facets* or *Winsteps*, interest in investigating psychometric multidimensionality in language assessment research also appears to be increasing. *ConQuest* is well-maintained and, in addition to unidimensional and multifaceted analyses, is capable of fitting several classes of multidimensional Rasch models such as bifactor models and higher-order models. These configurations lend themselves to the investigation of communicative competence, which posits that language has a multilayer and interwoven structure (Spoden & Fleischer, 2019). Unlike structural equation modeling, which can be affected by high correlations of components (multicollinearity), the multidimensional parametrizations of Rasch measurement are robust against multicollinearity. Another advantage of *ConQuest* is its ability to measure both general language ability and its subcomponents and provide correlation coefficients between latent variables (Wu et al., 1998).

### **RQ3**

Unidimensional Rasch models assume that test taker performance is caused by one underlying latent ability ( $\theta$ ) (Fan & Bond, 2019). The estimation of item difficulty and person ability in these models is reliable and accurate to the extent that evidence supports that the test engages the hypothesized  $\theta$  and not irrelevant dimensions (Linacre, 2019a, b). Therefore, it is important to provide evidence that the test measures one latent construct and that there are no substantive dimensions in the residuals of the Rasch models constructing a substantive secondary dimension. Despite widespread applications of unidimensional Rasch measurement, a large proportion of publications did not report the verification of unidimensionality. In addition, there was evidence for misconceptualization of unidimensionality and, among the reported methods, some would be impertinent to this analysis. For example, some authors frequently alluded to fit statistics as evidence for unidimensionality; however, these metrics are not appropriate for identifying secondary dimensions in data—they can only flag erratic patterns (Linacre, 2019a). Rasch reliability coefficients, similarities of item difficulty parameters, *t*-values analysis, and confidence interval analysis were similarly used, although they are not appropriate methods of unidimensionality analysis. The most suitable methods reported in the publications were the

DIMTEST (see Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001), principal component analysis (of residuals) (Fan & Bond, 2019), and factor analysis.

Another requirement of unidimensional Rasch measurement is the analysis of local dependency, which was reported by six (2.7%) of the publications reviewed. Local independence is intimately related to unidimensionality, as it is estimated by investigating the correlation of Rasch model residuals (Fan & Bond, 2019). While unidimensionality analysis is used to detect subsidiary dimensions in the residuals, local independence analysis only captures correlations between items' residuals. The reliability of studies that did not report local independence is questionable, since the presence of local independence can cause bias in estimating item and person parameters. However, the chi-square tests applied to address research question 3 have a caveat that should be noted in interpreting the results: investigating unidimensionality and local independence is rather challenging in contexts where there are multiple facets affecting measured attributes such as rater-mediated assessments (e.g., writing and speaking). This is mainly due to the presence of missing data in such assessments.

#### **RQ4 and 5**

This study also investigated the use and reporting of reliability and separation indices (RQ4), as well as fit statistics (RQ5), for persons and items. Nearly half of the papers reported person reliability and separation coefficients, whereas item separation was reported by 59 papers (27.44%). This indicates that there was no evidence of the reproducibility of item and person measures for a large number of the publications involving Rasch measurement. Given this lack of information, it is unclear whether the test items or tasks in these studies targeted the ability level of the test takers. In addition, the level of precision in estimating item and person parameters was uncertain (Linacre, 2019a).

The proportion of publications reporting fit statistics was similarly low, suggesting a lack of evidence for the conformity of the data to the predictions of the Rasch measurement in these publications. Because fit statistics are important requirements of Rasch measurement analysis, there can be little confidence in the psychometric validity of the results of papers that did not report fit measures (Linacre, 2019a). Among the different ranges proposed to evaluate item and person fit MnSq, the liberal range between 0.5 and 1.5 was most commonly used, likely due to the large number of writing and speaking studies where MFRM is applicable. Whereas this range is useful for MFRM (Linacre, 2019b), studies have shown that a more stringent range is appropriate for

multiple choice questions (0.8–1.2 for high-stakes decisions and 0.7–1.3 for mid- or low-stakes decisions) and polytomous data (0.6–1.4) (Boone, Staver, & Yale, 2014; Wright & Linacre, 1994). ZStd coefficients, on the other hand, have been shown to be prone to inflation in large samples (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008).

Some guidelines for choosing fit ranges are derived from previous experimental and simulation studies here. Although the conventional productive fit ranges described earlier may be convenient for Rasch-based research, we recommend that researchers establish the upper bounds of fit MnSq values based on the formulas proposed in Smith et al. (1998). The lower bound of fit MnSq has fewer practical consequence than the upper bound, since small fit MnSq values indicate increasing resemblance to a Guttman scale. Therefore, we recommend setting the lower bound of fit MnSq at 0.6, 0.7, or 0.8 for Likert scale questionnaires, low-stakes dichotomous tests, and high-stakes dichotomous tests, respectively (Bond & Fox, 2015; Linacre, 2002). For polytomous data, Smith et al. (2008) showed that outfit MnSq values were more stable in the rating scale model analyses, whereas infit MnSq values were more stable for the partial credit model. Recent studies have proposed bootstrapped CIs for polytomous fit indices as an alternative to conventional approaches (e.g., Seol, 2016). Wolfe's (2008, 2013) RBF method is recommended for estimating the CIs of fit indices generated by Winsteps.

A useful criterion for fit MnSq in multifaceted data has been 0.5–1.4 / 1.5 (Linacre, 2002). Similar fit ranges have been adopted in studies of the central tendency effect, limited range effect, and rater accuracy by Engelhard (2002) and Myford and Wolfe (2004). Nevertheless, Wolfe (2004, p. 48) expressed doubt as to whether fit indices are sensitive enough to capture rater effects. He called for “a series of simulation studies designed to document the sampling distributions of these [fit] indices and the rates with which these indices accurately and inaccurately nominate (or fail to nominate) ratings that are simulated to exhibit each of these rater effects.” It should be noted that MnSq and Zstd values are sensitive to the count of observations rather than the sample size; therefore, missing data may cause an inflated fit, especially in multifaceted data. Linacre (2002b) recommended at least 10 observation per category in polytomous and multifaceted data. Specifically, there should be at least  $25*(m+1)$  test takers for a stable analysis ( $m$  is the step or threshold, i.e., the number of scoring categories minus 1) and  $100*(m+1)$  test takers if there is inconsistency in the use of scoring categories.

Our review shows that Zstd fit indices are more sensitive to sample size. Smith et al. (2008) showed that Type I error rates are significantly higher for Zstd fit indices than fit MnSq indices. Specifically, Smith et al. (2008, p. 8) observed two limitations of Zstd: “for cases where mean square statistics fell within the range 0.7–1.3, the t-statistics [Zstd] increased in magnitude as sample size increased [...] Similarly, where mean square statistics identified misfit outside the 0.7–1.3 range, t-statistics only identified misfit as the sample size increased to beyond 200.” The instability of Zstd indices is attributed to their derivation from standard deviations (SDs) ( $Zstd = [(MnSq^{1/3} - 1) (3/SD)] + SD/3$ ), which tend to increase with larger and heterogeneous samples, and MnSq (Karabatsos, 2000). Therefore, to evaluate fit for samples larger than 250, it is advisable to rely more on MnSq indices than Zstd.

Finally, Smith and Su (2003) and empirical studies by Aryadoust, Goh, and Lee (2011) showed that fit MnSq indices are not sensitive to parameter invariance across subpopulations. Invariance is conventionally measured using DIF analysis in Rasch measurement. It is recommended that DIF analysis be investigated as a follow-up step to fit analysis. Guidelines for DIF analysis in language assessment have been presented by Ferne and Rupp (2007) as well as Raquel (2019).

## **RQ 6**

Network analysis detected several discrete communities of practice characterized by the amount of details concerning the analysis, software, language skills, and Rasch model used. The results of modularity and spectral cluster analyses consistently confirmed the same pattern in the communities, although the former analysis returned a more meaningful solution. It was found that Rasch-based research involving production language skills (writing and speaking) formed a discrete community of practice. Modularity cluster 1 focused on production skills using MFRM and *Facets*, *Bigscale*, and *PCRasch* software. This cluster also reported PR and PS more often. By contrast, cluster 2 focused on language components and perception skills (vocabulary & reading), and integrated skills and used the Rasch model or the Rasch-Andrich rating scale model most often with *Winsteps*. Papers in this cluster frequently reported item reliability, unidimensionality, and local independence. This is likely because the techniques used to investigate unidimensionality and local independence are applicable to ‘objective’ tests, whereas establishing these requirements is challenging for rater-mediated data (like those in cluster 1) due primarily to missing data. Overall, two discrete research streams involving Rasch measurement in language assessment

emerged from the networks system analysis: the production and performance cluster that comprised of research on rater-mediated assessments and the perception and language elements cluster that entailed research on objective assessments.

### **Limitations**

The current study is not without its limitations. First, the data were extracted from 21 journals. Future research may extend the scope of the data and perhaps include more papers and even dissertations and research reports that apply Rasch measurement. Second, the included papers were not coded for “targeting of persons and items” and “the threshold ordering of polytomous items” (Tennant & Conaghan, 2007, pp. 1360-1361). Tennant and Conaghan (2007, p. 1361) argued that comparing the item mean score with the person mean score would “provide an indication of how well targeted the items are for people in the sample.” If item and person mean scores are roughly similar, the measurement is said to be well targeted. For polytomous data, the monotonicity of steps should also be investigated. Monotonicity means “that the probability of more extreme or greater responses on an item corresponds with a greater amount of the latent trait being measured” (Kean, Brodke, Biber, & Gross, 2018, p.97). For example, there should be a consistency between higher levels of language ability and higher scoring categories on rating scales. Bond and Fox (2015) advised that, for monotonic steps, step difficulty should increase by 1.2 to 5 logits.

### **Conclusion and Guidelines**

This study showed that the use of Rasch measurement has been gradually increasing in language assessment as more major journals published studies that used this method. Rasch measurement was frequently used in writing, speaking, and integrated skills research; *Facets* and *Winsteps* were the most frequently used Rasch software packages; and an array of unidimensional Rasch models were adopted in publications involving Rasch measurement. Despite this respectable spread, a number of the publications investigated did not present evidence of reliability, fit, unidimensionality, and local independence. This finding suggests a lack of evidence for reliability and psychometric validity in these studies (Wright & Stone, 1999). This shortcoming was especially evident in studies that used Rasch measurement as a preliminary validation instrument. We suggest that applied linguistics journals should require authors to present rigorous evidence of reliability and psychometric validity in manuscripts submitted for publication. Some general guidelines for a study involving Rasch measurement are presented in Table 6 which consists of (i) item and person reliability and separation indices (in MFRM analysis, evidence for the reliability

of other facets should be presented); (ii) item and person infit and outfit MnSq indices; (iii) evidence of unidimensionality and local independence (in multidimensional Rasch model analysis, evidence of unidimensionality per dimension should be presented). In addition, when data lend themselves to a Wright Map, journals should require the inclusion of a Wright Map in which the ordering and spacing of items is compared to theory as a way of investigating construct validity (see Boone et al., 2014, pp. 111-158).

Table 6<sup>2</sup>  
*General Guidelines for Research Involving Rasch Measurement*

Criterion	Recommendation	Relevant sources
Unidimensionality	<p>Principal component analysis of Rasch residuals (PCAR): PCAR detects components in Rasch residuals, which are the differences between predicted and actual data. If the detected components explain more than two eigenvalues (&gt;2), they are respectable and warrant further investigation. In this situation, the component identified explains more than random noise and likely partitions the test items into meaningful structures not explained by the model used. This indicates the attenuation of unidimensionality.</p> <p>Each component divides items into two subsets representing two ends of the components. To determine the impact of the identified dimensions, Linacre (1998, p. 639) suggested that analysts should (i) measure the ability of each test taker on each subset; (ii) cross-plot the measures; and (iii) plot the difficulty of each test item in the subset against the original item difficulty. By examining these plots, analysts can identify traces of perturbation; for example, if the difficulty of items change significantly, they will fall off-diagonal, suggesting potential perturbation in the item.</p>	Fan & Bond (2019); Linacre (1998)
Local independence (LI)	Q3 coefficients or Rasch residual correlations: Like unidimensionality, LI can be investigated using Rasch model residuals, e.g., via Q3 coefficients. Small values of Q3 coefficients do not affect item difficulty, but Q3 coefficients exceeding  0.30  could point to violations of LI.	Chen & Thissen (1997); Christensen, Makransky, & Horton, (2017); Liu & Maydeu-Olivares (2013)
Fit	Infit MnSq and outfit MnSq: As previously discussed, more stringent criteria for evaluating fit should be set in high-stakes tests. Establish the upper bounds of fit MnSq indices using Smith et al.'s (1998) recommendation when the sample is normally distributed (Infit $MnSq = 1 + \frac{2}{\sqrt{x}}$ ; and Outfit $MnSq = 1 + \frac{6}{\sqrt{x}}$ , where x=sample size). Although overfit is viewed as a lesser concern, it could indicate sizeable LI in the data (Ponocny & Ponocny-	Bond & Fox (2015); Linacre (2019a); Ponocny & Ponocny-Seliger, (1999); Smith et al. (1998)

<sup>2</sup> Authors would like to thank Mike Linacre for his comments on this table.

Criterion	Recommendation	Relevant sources
	<p>Seliger, 1999). Therefore, establishing the lower bound of fit indices is also important. In this case and when a normal distribution is violated, the lower bounds of the conventional fit ranges are useful: 0.8 – 1.2 (high-stakes multiple choice questions (MCQs)); 0.7 – 1.3 (mid-/low-stakes MCQs); 0.6 – 1.4 (polytomous data, e.g., surveys); and 0.5 – 1.4 /1.5 (multifaceted data).</p> <p>Zstd infit and outfits seem to be most useful when datasets consist of &lt; 250, beyond which they can become inflated. If the Zstd indices fall outside of the range -1.96 to +1.96 in samples &lt; 250, the data do not fit the model <i>perfectly</i>. In this case, if the fit MnSq values also indicate erratic patterns in the data, misfitting items/persons (or other facets) should be closely examined. In high-stakes tests, achieving perfection is desirable; therefore, erratic Zstd indices could indicate significant issues in the data, although with sparse data (e.g., in multifaceted Rasch analysis) or heterogeneous samples with high SDs, Zstd can easily become inflated.</p>	
Reliability	<p>Reliability and separation: in ‘objective’ assessments (where raters do not play a role, such as MCQs), Rasch reliability coefficients <math>\geq .80</math> indicate two or more separable levels of performance in the data (separation <math>\geq 2</math>). As reliability coefficients become smaller (<math>&lt; .80</math>), it becomes increasingly unlikely to identify distinct groups of items and persons. High reliability indicates a high likelihood that high-ability test takers actually had high ability measures and low-ability test takers had low measures.</p> <p>Reliability and separation are measured using true and observed variance; therefore, low reliability may simply indicate little variance in the data and a homogenous sample. Low reliability in objective tests indicates that, on average, high standard error of measurement (SEM) for some items or persons. While reliability and separation indicate sample-level precision, SEM indicates item- or person-level precision. For example, if SEM = 0.2 and test item difficulty = 1.00 logits, there is 68% probability that the difficulty measure falls between 0.8 and 1.2 logits.</p> <p>The criteria suggested above apply to Spearman-type reliabilities, which focus on differences (variance), rather than inter-rater reliabilities, which focus on similarities. There are currently no generally-accepted inter-rater reliability coefficients suitable for <i>large, sparse</i> judging plans. A useful reference addressing this issue is Saal, Downey, and Lahey (1980).</p>	Aryadoust, Ng, Foo, & Esposito (2020); Linacre (2019a); Saal, Downey, and Lahey (1980)

*Note:* MnSq = Mean Square; Zstd = Standardized.

Finally, the significance of replication research in applied linguistics has recently been discussed (e.g., Marsden, Morgan-Short, Thompson, & Abugaber, 2018; Morgan-Short et al., 2018). Guidelines provided in these studies are applicable to replication in Rasch measurement research. The replicability of item and person measures, fit, unidimensionality, local

independence, and reliability/separation estimates in studies involving Rasch measurement should be investigated in future research. With the availability of data repositories such as IRIS (Marsden, Mackey, & Plonsky, 2016), researchers can use readily available measurement tools along with pertinent datasets. Rasch measurement allows for reproducibility analysis by offering researchers the opportunity to conduct differential item/facet/distractor/rater functioning. Such analyses would reveal whether measurement tools maintain their properties under different conditions.

The field of language testing has several challenges to address in the future, including the need to address validation problems that can arise when integrated tasks are used in tests. Rasch measurement will provide a useful tool to model the dimensionality of these assessments and ascertain the reliability of measurements. It is anticipated that Rasch measurement will be further extended to address scaling issues while maintaining accessibility and ease of use of its tools and software. For this reason, it is imperative that best practices for Rasch measurement are established now.

### **Acknowledgements**

We wish thank to Ang Bee Hoon, Sim Yue Ling, Jamie Loo Fang Ting, and Stella Ferdinand for their assistance in coding the selected papers. We would also like to thank three anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. The study was supported by two grants from The National Institute of Education (NIE) of Nanyang Technological University (NTU), Singapore (grant numbers = RI 2/16 VSA and RI 1/18 VSA). The opinions expressed in this article are the authors' own and do not necessarily reflect the view of NIE and NTU.

### **References**

- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using Multidimensional Item Response Theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51. <https://doi.org/10.1111/j.1745-3992.2003.tb00136.x>
- Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised item response modelling software (Version 4). *Australian Council for Educational Research*.
- Andersen, E. B. (1977, March 01). Sufficient statistics and latent trait models [journal article]. *Psychometrika*, 42(1), 69-81. <https://doi.org/10.1007/bf02293746>
- Andrich, D. (1978, December 01). A rating formulation for ordered response categories [journal article]. *Psychometrika*, 43(4), 561-573. <https://doi.org/10.1007/bf02293814>
- Andrich, D., Sheridan, B., and Luo, G. (2009). *RUMM2030: Rasch unidimensional models for measurement (computer program)*. RUMM Laboratory.
- Armstrong, A. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the lz person-fit statistic. *Practical Assessment Research & Evaluation*, 12(16). <http://pareonline.net/getvn.asp?v=12&n=16>

- Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, 8(4), 361-385. <https://doi.org/10.1080/15434303.2011.628632>
- Aryadoust, V., Ng, L. Y., Foo, S., & Esposito, G. (2020). A neurocognitive investigation of test methods and gender effects in a computerized listening comprehension test. *Computer Assisted Language Learning—First Online*.
- Aryadoust, V., Tan, H.A.H., & Ng, L.Y. (2019). A Scientometric review of Rasch measurement: The rise and progress of a specialty. *Frontiers in Psychology*, 10, 2197. doi: 10.3389/fpsyg.2019.02197
- Baghaei P., & Aryadoust, V. (2015). Modeling test method effect with a multidimensional Rasch model. *International Journal of Testing*, 15, 71-87. <https://doi.org/10.1080/15305058.2014.941108>
- Baghaei, P., Kemper, C. J., Reichert, M., & Greiff, S. (2019). Applying the mixed Rasch model in assessing reading comprehension. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume II: Advanced Methods*, (pp. 15-32). Routledge. <https://doi.org/10.4324/9781315187808>
- Baldonado, A. A., Svetina, D., & Gorin, J. (2015). Using necessary information to identify item dependence in passage-based reading comprehension tests. *Applied Measurement in Education* 28, 202-218. <https://doi.org/10.1080/08957347.2015.1042154>
- Barabasi, A. (2016). *Network science*. Cambridge University Press.
- Barabási, A. (2003). *Linked*. Perseus Books.
- Batty, A. O. (2014). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, 32(1), 3-20. <https://doi.org/10.1177/0265532214531254>
- Bejar, I. I. (1980). *A procedure of investigating the unidimensionality of achievement tests based on item parameter estimates*. *Journal of Educational Measurement*, 17(4), 283-296. <https://doi.org/10.1111/j.1745-3984.1980.tb00832.x>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading MA: Addison-Wesley.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. <https://doi.org/10.1088/1742-5468/2008/10/p10008>
- Bond, T., & Fox, C. (2015). *Applying the Rasch model* (3rd ed.). Lawrence Erlbaum Associates.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Science & Business Media. <https://doi.org/10.1007/978-94-007-6857-4>
- Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. <https://doi.org/10.3102/10769986022003265>
- Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, 41(3), 178-194. <https://doi.org/10.1177/0146621616677520>
- Chung, F. R. K. (1997). *Spectral graph theory*. American Mathematical Society.

- DeMars, C. E. (2017). Infit and outfit: Interpreting statistical significance and magnitude of misfit in conjunction. *Journal of Applied Measurement*, 18(2), 163-177.
- Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang. <https://doi.org/10.3726/978-3-653-04844-5>
- Eckes, T. (2019). Many-Facet Rasch Measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*, (pp. 153-176). Routledge. <https://doi.org/10.4324/9781315187815>
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika* 56, 495–515. <https://doi.org/10.1007/BF02294487>
- Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Erlbaum.
- Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge/Taylor & Francis Group.
- Fan, J., & Bond T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*, (pp. 83-102). Routledge. <https://doi.org/10.4324/9781315187815>
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4(2), 113–148. <https://doi.org/10.1080/15434300701375923>
- Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 15-38). Springer New York.
- Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.
- Fisher, W. P. (1992). Reliability statistics. *Rasch measurement transactions*, 6(3), 238.
- Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659, 1–44.
- Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Empirical Press.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement* 9, 139-164. <https://doi.org/10.1177/014662168500900204>
- Hodge, K. J., & Morgan, G. B. (2017). Stability of INFIT and OUTFIT como simulated estimates in applied setting. *Journal of Applied Measurement*, 18(4), 383-392.
- Holster, T. A., & Lake, J. (2016). Guessing and the Rasch model. *Language Assessment Quarterly*, 13(2), 124-141.
- Hulin, C. L., Drasgow, F., & Parsons, C. (1983). *Item response theory: Applications to psychological measurement*. Dow & Jones Irwin.
- Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1, 152-176.
- Kean, J., Brodke, D. S., Biber, J., & Gross, P. (2018). An introduction to Item Response Theory and Rasch analysis of the eating assessment tool (EAT-10). *Brain impairment: A*

- Multidisciplinary Journal of the Australian Society for the Study of Brain Impairment*, 19(Spec Iss 1), 91–102. <https://doi.org/10.1017/BrImp.2017.31>
- Lee, Y. W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, 21(1), 74-100. <https://doi.org/10.1191/0265532204lt260oa>
- Linacre, J. M. (2019a). *A user's guide to WINSTEPS® MINISTEP Rasch-model computer programs. Program Manual 4.4.7.* <https://www.winsteps.com/winman/copyright.htm>
- Linacre, J. M. (2019b). *FACETS: Computer program for many faceted Rasch measurement (Version 3.82.1).* Mesa Press.
- Linacre J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions*, 19(3), 1032.
- Linacre, M. (2002b). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 31(1), 85-106.
- Linacre, J. M. (2002a). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J. M. (1999). Explorations into local independence with T-Rasch Ponocny, I., Ponocny-Seliger, E. *Rasch Measurement Transactions*, 13, 710.
- Linacre, J. M. (1998). Structure in Rasch residuals: why principal components analysis. *Rasch measurement transactions*, 12(2), 636.
- Linacre, J. M. (1997). An all-purpose person fit statistic? *Rasch Measurement Transactions*, 11(3), 582-583.
- Linacre, J. M. (1995). Reliability and separation nomograms. *Rasch Measurement Transactions*, 9(2), 421.
- Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd Ed.). Chicago: MESA.
- Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, 73(2), 254-274. <https://doi.org/10.1177/0013164412453841>
- Lord, F., & Wright, B. D. (2010). Fred Lord and Ben Wright discuss Rasch and IRT Models. *Rasch Measurement Transactions*, 24(3), 1289-1290.
- Mallinson, T., & Stelmack, J. (2001). Going beyond unreliable reliabilities. *Rasch Measurement Transactions*, 14(4), 787-788.
- Marsden, E., Mackey, A., & Plonsky, L. (2016). *The IRIS repository: Advancing research practice and methodology.* In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Routledge. <https://doi.org/10.4324/9780203489666>
- Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, 68. <https://doi.org/10.1111/lang.12286>
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174. <https://doi.org/10.1007/bf02296272>
- McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test1. *Language Testing*, 8(2), 139-159. <https://doi.org/10.1177/026553229100800204>
- McNamara, T. F. (1996). *Measuring second language performance.* Longman.
- McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, 29(4), 555-576. <https://doi.org/10.1177/0265532211430367>

- Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, 31(4), 453-477. <https://doi.org/10.1177/0265532214527277>
- Morgan-Short, K., Marsden, E., Heil, J., Issa, B., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multi-site replication in SLA research: Attention to form during listening and reading comprehension in L2 Spanish. *Language Learning*, 68.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, 18(1), 41-68. <https://doi.org/10.3102/10769986018001041>
- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582. <https://doi.org/10.1073/pnas.0601602103>
- Raîche, G. (2005). Critical Eigenvalue sizes (variances) in standardized residual principal components analysis (PCA). *Rasch Measurement Transactions*, 19(1), 1012.
- Raquel, M. (2019). The Rasch measurement approach to differential item functioning (DIF) analysis in language assessment research. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques*, (pp. 103-131). Routledge. <https://doi.org/10.4324/9781315187815>
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.
- Rasch, G. (1963). "The poisson process as a model for a diversity of behavioral phenomena," in *International Congress of Psychology, Vol. 2*, p. 2
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207-230. <https://doi.org/10.3102/10769986004003207>
- Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44(1), 75-92. <https://doi.org/10.1111/j.2044-8317.1991.tb00951.x>
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413.
- Salzberger, T. (2012). Reporting a Rasch analysis. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models in Health* (pp. 347-361): John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118574454.ch19>
- Sayama, H. (2015). *Intoduction to the modeling and analysis of complex systems*. Open SUNY Textbooks.
- Schotten, M., Aisati, M., Meester, W. J. N., Steinginga, S., & Ross, C.A. (2018). *A brief history of Scopus: The world's largest abstract and citation database of scientific literature*. In F. J. Cantu-Ortiz (Ed). *Research analytics: Boosting university productivity and competitiveness through Scientometrics* (pp. 33-57). Taylor & Francis.

- Smith, A.B., Rush, R., Fallowfield, L.J., Velikova, G., and Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8(33), <https://doi.org/10.1186/1471-2288-8-33>.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1(2), 199-218.
- Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of outcome measurement*, 2(1), 66-78.
- Smith, R. M., & Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement*, 4(2), 153-163.
- Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8(1), 33. <https://doi.org/10.1186/1471-2288-8-33>
- Smith, R.M., Linacre, J.M., and Smith, Jr., E.V. (2003). Guidelines for manuscripts. *Journal of Applied Measurement*, 4, 198-204.
- Spoden, C., & Fleischer, J. (2019). Multidimensional Rasch models in first language listening tests. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume II: Advanced Methods*, (pp. 33-55). Routledge. <https://doi.org/10.4324/9781315187808>
- Stewart, J., McLean, S., & Kramer, B. (2017). A response to Holster and Lake regarding guessing and the Rasch model. *Language Assessment Quarterly*, 14(1), 69-74.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589-617. <https://doi.org/10.1007/bf02294821>
- Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 357-375). Springer New York.
- Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of educational psychology*, 102(4), 989-1004. <https://doi.org/10.1037/a0019507>
- Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, 57(8), 1358-1362. <https://doi.org/10.1002/art.23108>
- Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch measurement transactions*, 20(1), 1048-1051.
- von Davier, M. (1996). Mixtures of polytomous Rasch models and latent class models for ordinal variables. *Softstat*, 95.
- von Davier, M. (2001). *WINMIRA [Computer Program]*. ASCAssessment Systems Corporation, USA and Science Plus Group.
- Von Luxburg, U. (2006). *A tutorial on spectral clustering: Technical Report, 149*. Berlin: Max Planck Institute for Biological Cybernetics.
- Wang, W. C., Wilson, M., and Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. *Objective measurement: Theory Pract.* 4, 139–155.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.
- Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 35(2), 161-192. <https://doi.org/10.1177/0265532216686999>

- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35–51.
- Wolfe, E. W. (2008). RBF.sas (Rasch bootstrap Fit): A SAS macro for estimating critical values for Rasch model fit statistics. *Applied Psychological Measurement*, 32, 585–586.
- Wolfe, E. W. (2013). A bootstrap approach to evaluating person and item fit to the Rasch model. *Journal of Applied Measurement*, 14(1), 1–9.
- Wright, B. D., & Stone, M. A. (1979). *Best test design*. MESA press.
- Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. MESA press.
- Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright, B. D. (1994a). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright, B. D. (1994b). Unidimensionality coefficient. *Rasch Measurement Transactions*, 8(3), 385.
- Wright, B. D. (1996a). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10(3), 509-511.
- Wright, B. D. (1996b). Reliability and separation. *Rasch measurement transactions*, 9(4), 472.
- Wright, B. D., & Masters, G. N. (1990). Computation of outfit and infit Statistics. *Rasch Measurement Transactions*, 3(4), 84-85.
- Wright, B., & Stone, M. (1999). Validity. *Measurement Essentials (2<sup>nd</sup> ed.)* (pp. 167-171). WIDE RANGE, INC.
- Wu, M. L., Adams, R., and Wilson, M. (1998). *ACER ConQuest (Version 1.0) [Computer Package]*. Council for Educational Research.
- Yamashita, J., & Shiotsu, T. (2015). Comprehension and knowledge components that predict L2 reading: A latent-trait approach. *Applied Linguistics*, 38(1), 43-67. <https://doi.org/10.1093/applin/amu079>