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\(^1\); 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 measurement

---

\(^1\) 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 unspeeded 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; Salzbeger, 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 |.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: infit MnSq=$l+\frac{2}{\sqrt{x}}$ and outfit MnSq=$l+\frac{6}{\sqrt{x}}$, 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 (Drasgow, Levine, & Williams, 1985; Hulin, Drasgow, & 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, Steigninga, & 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*

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

*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**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Description</th>
<th>Example reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive feature</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>Country or region and language skill</td>
<td>NA</td>
</tr>
<tr>
<td>Features of the sample</td>
<td>Sample size</td>
<td>NA</td>
</tr>
<tr>
<td>Features of the instrument</td>
<td>Test used</td>
<td></td>
</tr>
<tr>
<td>The Rasch model used</td>
<td>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)</td>
<td>Ackerman (1994); Andrich (1978); Rasch (1982); Linacre (1960/1980); Linacre (1994)</td>
</tr>
<tr>
<td>Software</td>
<td>Facets, Winsteps, Winmira, RUMM, ConQuest, etc.</td>
<td>Andrich et al. (2009); Linacre (2019a); Linacre (2019b); von Davier (2001)</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item and person reliability</td>
<td>• An index of precision and reproducibility of items / person measures</td>
<td>Fisher (1992); Linacre (1995); Mallinson &amp; Stelmack (2001); Wright (1996b);</td>
</tr>
</tbody>
</table>
| Item and person separation        | • Another representation of reliability to classify items and persons into separable clusters  
|                                   | • The ratio of the true standard deviation over the error of standard deviation | As above         |
| **Psychometric validity**         |                                                                              |                   |
| Infit mean square (MnSq) and standardized (ZStd) | • MnSq: An inlier-sensitive fit index to capture anomalous response patterns on items targeted on persons and vice versa  
|                                   | • ZStd provides a t-test to investigate whether the data have a perfect fit to the model | Linacre (2002a, b); Wright (1994a); Wright & Masters (1990) |
| Outfit mean square (MnSq) and standardized (ZStd) | • MnSq: An outlier-sensitive fit index to capture anomalous responses to items far away from persons and vice versa  
• ZStd provides a t-test to investigate whether the data have a perfect fit to the model | As above  
| Unidimensionality | 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:  
• Principal component analysis of residuals  
• Factor or principal component analysis  
• Infit and outfit statistics | Ackerman et al. (2003); Raïche (2005); Tennant & Pallant (2006); Wright (1994b);  
| Local independence | 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. | 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

<table>
<thead>
<tr>
<th>Language skill or component</th>
<th># of papers using the unidimensionality analysis method</th>
<th>%</th>
<th># of papers using the local independence method</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>4</td>
<td>1.86</td>
<td>2</td>
<td>0.93</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>16</td>
<td>7.44</td>
<td>5</td>
<td>2.33</td>
</tr>
<tr>
<td>Integrated</td>
<td>15</td>
<td>6.98</td>
<td>6</td>
<td>2.79</td>
</tr>
<tr>
<td>Listening</td>
<td>7</td>
<td>3.26</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>Speaking/Oral</td>
<td>9</td>
<td>4.19</td>
<td>2</td>
<td>0.95</td>
</tr>
<tr>
<td>Reading</td>
<td>9</td>
<td>4.19</td>
<td>5</td>
<td>2.33</td>
</tr>
<tr>
<td>Writing</td>
<td>14</td>
<td>6.51</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>Others</td>
<td>5</td>
<td>2.33</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>57/215</strong></td>
<td><strong>36.74</strong></td>
<td><strong>19/215</strong></td>
<td><strong>10.23</strong></td>
</tr>
</tbody>
</table>

*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
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**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>No.</th>
<th>Label</th>
<th>N</th>
<th>No. of LS</th>
<th>Chi-Square Statistics</th>
<th>φ_C</th>
<th>No. of RM</th>
<th>Chi-Square Statistics</th>
<th>φ_C</th>
<th>Software Used</th>
<th>Chi-Square Statistics</th>
<th>φ_C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>df</td>
<td>N</td>
<td>χ²</td>
<td></td>
<td>df</td>
<td>n</td>
<td>χ²</td>
<td></td>
</tr>
<tr>
<td>Modularity 1</td>
<td>1</td>
<td>Production and performance</td>
<td>110</td>
<td>8</td>
<td>7</td>
<td>283</td>
<td>89.18****</td>
<td>4</td>
<td>7</td>
<td>221</td>
<td>146.23****</td>
<td>4</td>
</tr>
<tr>
<td>Modularity 2</td>
<td>2</td>
<td>Perception and language elements</td>
<td>105</td>
<td>8</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral 1</td>
<td>-</td>
<td>-</td>
<td>65</td>
<td>8</td>
<td>21</td>
<td>283</td>
<td>108.22****</td>
<td>2</td>
<td>21</td>
<td>221</td>
<td>150.76****</td>
<td>3</td>
</tr>
<tr>
<td>Spectral 2</td>
<td>-</td>
<td>-</td>
<td>56</td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral 3</td>
<td>-</td>
<td>-</td>
<td>55</td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral 4</td>
<td>-</td>
<td>-</td>
<td>39</td>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 (φ_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/DFF) 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 (θ) (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 θ 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 62

<table>
<thead>
<tr>
<th>General Guidelines for Research Involving Rasch Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion</strong></td>
</tr>
<tr>
<td><strong>Unidimensionality</strong></td>
</tr>
<tr>
<td><strong>Local independence (LI)</strong></td>
</tr>
<tr>
<td><strong>Fit</strong></td>
</tr>
</tbody>
</table>

---

2 Authors would like to thank Mike Linacre for his comments on this table.
<table>
<thead>
<tr>
<th>Criterion</th>
<th>Recommendation</th>
<th>Relevant sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>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).</td>
<td>Aryadoust, Ng, Foo, &amp; Esposito (2020); Linacre (2019a); Saal, Downey, and Lahey (1980).</td>
</tr>
<tr>
<td>Reliability and separation:</td>
<td>Reliability and separation: in ‘objective’ assessments (where raters do not play a role, such as MCQs), Rasch reliability coefficients ≥.80 indicate two or more separable levels of performance in the data (separation ≥ 2). As reliability coefficients become smaller (&lt; .80), 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.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>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 large, sparse judging plans. A useful reference addressing this issue is Saal, Downey, and Lahey (1980).</td>
<td></td>
</tr>
</tbody>
</table>

*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


