Rasch Calibration and Differential Item Functioning (DIF) Analysis of the Indonesian National Assessment Program-Language (INAP-L)

Bahrul Hayat[1], Muhammad Dwirifqi Kharisma Putra[2], Rahmawati[3], & Khairunesa Isa[4]

[1]Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, Indonesia

[2]Faculty of Psychology, Universitas Gadjah Mada, Indonesia

[3]Pusat Asesmen Pendidikan, Indonesia

[4]Center for General Studies and Co-curricular, Universiti Tun Hussein Onn, Malaysia

Correspondence: Bahrul Hayat, Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, Kertamukti St. No. 5, South Tangerang City 12220, Indonesia. E-mail: bahrulhayat@uinjkt.ac.id

Received: April 25, 2023 Accepted: May 29, 2023 Online Published: June 14, 2023

doi:10.5430/wjel.v13n6p402 URL: https://doi.org/10.5430/wjel.v13n6p402

Abstract

This study aimed to evaluate the psychometric properties of the INAP-L instrument by applying the Rasch model. The psychometric evaluation includes item calibration and differential item functioning (DIF) assessment. Participants in this study were 6153 high school students (2326 boys and 3827 girls) aged 14-19 years (mean age = 15.76, SD = 0.78) from 280 schools spread across 34 provinces in Indonesia. The results of the Rasch model analysis show that the assumptions of unidimensionality and local independence are met. The internal consistency analysis of the INAP-L instrument (PSR = 0.80; Ordinal α = 0.81) and item-person targeting showed quite good results. At the item level, it was discovered that four of the forty items did not fit the Rasch model. Meanwhile, the gender-based DIF analysis revealed that all items were free from gender DIF and that two out of forty items showed DIF based on school type. It can be concluded that the INAP-L instrument has good psychometric properties. Furthermore, multilevel analysis was performed to determine the effect of clustering in the data, and it was discovered that INAP-L has a multilevel data structure with ICC = 0.185 and a design effect > 2.00. This means that future research on the relationship among variables on the INAP-L score must consider a multilevel approach. Even though the developer has already published an interim report on INAP-L, the findings of this study can be used as a reference in improving the instrument before conducting INAP-L in the future.

Keywords: INAP, factor analysis, language assessment, Rasch model, validation

1. Introduction

Large-scale assessment (LSA) is a type of assessment that is used to determine whether or not a system is in good shape. In the context of education, LSA monitors the quality of learning outcomes through regular assessments of learning achievement that can be compared from year to year (Beaton & Barone, 2017; Lockheed et al., 2015). Actually, Indonesia already has a National Examination (UN) system that serves as an accurate and fair measurement tool to assess the overall quality of education in Indonesia (Setiadi, 2016). However, in December 2019, the Minister of Education and Culture of the Republic of Indonesia issued a policy to replace the National Examination (UN) with the Minimum Competency Assessment (AKM) and Character Survey (Rahayu et al., 2021). The AKM had been developed three years before the decision as part of the Indonesia National Assessment Program (INAP).

INAP has been a long-standing desire of the Ministry of Education and Culture since the late 1990s when INAP was designed. Furthermore, large-scale national assessments have a long history in Indonesia, dating back to the mid-1970s (for example, Moegiadi et al., 1979). INAP is a program to monitor and map educational achievements at the national and regional levels. INAP is a sample survey-based assessment that has nothing to do with individual student certification decisions. The INAP results will provide diagnostic information to help improve the education system. INAP measures the following competencies: (1) mathematical literacy, (2) reading literacy, (3) scientific literacy, and (4) Indonesian language (Center for Educational Assessment, 2016). These essential literacies promote learning in the twenty-first century (Geisinger, 2016; McFarlane, 2013). These essential literacies are also measured in international LSAs such as the PISA and Trends in International Mathematics and Science Study (TIMSS), in which Indonesia participates. INAP, which measures language literacy through INAP-L, is a new history and is the first time language literacy has been assessed on a large scale in Indonesia.

According to various studies on language assessment, INAP-L is classified as a large-scale language assessment (LLA) (e.g., Davies, 2013; Erickson & Berg-Bengtsson, 2012; Kunnan, 2017), where the assessment is more specific than other LSAs measuring mathematical, reading, and scientific literacy. In general, LLA is used for a variety of purposes in a variety of contexts (Kunnan & Grabowski, 2013). The multiple purposes include monitoring student progress and diagnosing student strengths and weaknesses, while the multiple contexts

include school, college, university, and workplace (Kunnan, 2017). Furthermore, the uniformity of tests and testing practice (including test administration, scoring, reporting, and score interpretation) across geographical regions and administration time is a main feature of large-scale language assessments (Kunnan, 2008). At the school level, LLA could be used to provide diagnostic information to all stakeholders (e.g., teachers, students, parents, and school administrators) (Kunnan, 2017). With the development INAP-L, an LLA, will benefit Indonesian policymakers in terms of language education. This benefit is consistent with recent research on an empirical model of language assessment literacy (Kremmel & Harding, 2020), which measures the same construct as INAP-L. As a result, the implementation of INAP-L will be valuable program for improving Indonesian education quality.

Given Indonesia's high diversity of languages, language literacy is a unique construct that characterizes the INAP program. Furthermore, language competence is a special feature of INAP, since language literacy is not tested in large-scale international assessments [for example, the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS)]. Based on our literature review, there are several studies of large-scale language assessment of language proficiency. Alvarez (2013) measured language proficiency in the form multiple choice examination and tested its construct validity. Another study examines the principles and guidelines for developing a large-scale assessment of English language learners for disabilities (Liu et al., 2017). Additionally, several studies have found that large-scale language assessment has both positive and negative effects on students' language learning strategies (Abbasi et al., 2020). Therefore, INAP-L expected can be valuable resources for the developers and practicioners of the curriculum and instruction.

From methodological perspective, one of the most difficult challenges for INAP developers was applying methodologies that make the results of INAP comparable across study cycles. Using cutting-edge methodologies, such as that used in standard international assessments, is critical to achieving INAP's goal of measuring student competence over time. International large-scale assessments have proven to be very good at producing data comparable across study cycles, allowing the trend of an increasing or decreasing competency of population group (e.g., country) to be known (e.g., Adams, 2003). Such information is made possible by applying modern psychometric theory in the analysis of large-scale assessment data (e.g., Blomeke & Gustafsson, 2017; Maehler & Rammstedt, 2020). Modern psychometric theory, such as the IRT and the Rasch model, has been used in large-scale assessments such as PISA and TIMSS. Therefore, INAP-L must use data analysis standards similar to those used in international studies (i.e., the Rasch model).

However, the methodology used in large-scale national and international assessments is not without criticism and debate (e.g., Feninger & Lefstein, 2014; Goldstein, 2004; Wang, 2001; Zhao, 2020), because all of these studies face well-known and fundamental problems of defining and then attempting to ensure 'comparability of meaning' for their instruments across diverse educational systems and cultures (Goldstein, 2017). Given Indonesia's vast geographical area and economic and social disparities across provinces, INAP-L is undoubtedly not free from the same issues. These issues, however, can be diagnosed using the Rasch measurement model, which has an invariance or sample-free property (Rasch, 1966; Wright, 1968).

In addition, there are other things that have the potential to disturb the psychometric properties of the INAP-L, namely the possibility of item bias (Kreiner, 2013). In the context of INAP-L, the potential for item bias can occur due to gender differences. This opinion is in line with the findings of previous studies which state that language tests tend to be gender biased (Atmawati, 2018). In addition, other study have found that there is a significant difference between the language skills of men and women in the Indonesian sample (Almuzakir & Qamariah, 2019). Apart from gender differences, it should also be considered that there might be differences in language achievement in public and private schools, although research specifically on this matter is still very limited. Therefore, DIF analysis of gender differences and school types will provide important information for INAP-L calibration. The Rasch model provides tools that can be used to examine item bias so that the calibration will produce detailed information about measurement invariance (Smith et al., 2016).

However, until now, no study has been done to assess the psychometric properties and the validity of the INAP-L instrument, as well as to test the functioning of items across different background variables such as gender and school type published in a journal article. Therefore, the purpose of this study was to evaluate the psychometric characteristics of the INAP-L instrument and to assess the DIF based on gender and school type using the Rasch model. This study is the first psychometric assessment of INAP-L formally published in a journal article despite the developer has its own unpublished validation study.

2. Method

2.1 Participants

The sample in this study was senior high school students from all provinces in Indonesia (34 provinces) , with an average number of students per province of 181.76 people totalling 6153 students. Samples were taken from the population using the multistage sampling method from 34 provinces in Indonesia. In the first stage, 280 schools were sampled from 34 provinces. In the second stage, a sample of 10-20 students was drawn from each school. Respondents have an age range of 14 to 19 years, with a mean age of 15.76 and a standard deviation (SD) of 0.78. The sample consisted of 2326 men and 3827 women.

2.2 Instrument: INAP Language Assessment

The INAP-L instrument was developed in 2016 by the Center for Educational Assessment and the Language Development Institute of the Indonesian Ministry of Education and Culture. This instrument consists of 40-item using mixed-format items. 25 items were scored dichotomously and 15 items were scored polytomously. The items with a polytomous score format used three score options (0, 1, and 2)

and four score options (0, 1, 2, and 3). We will limit the explanation of the instrument to prevent leakage of the instrument contents, which allows students to learn the material tested in the next INAP-L.

2.3 Rasch Model for Mixed-Format Test

Rasch measurement (Engelhard & Wind, 2018; Masters & Wright, 1984; Rasch, 1960; Wright, 1968) is a family of mathematical models which is part of modern psychometric theory developed to overcome the limitations of classical test theory (CTT). In modern psychometric theories, a single proficiency variable, θ, is often known as a latent ability that underlies a person's performance on a test. The latent ability, although not directly observable, can be used to predict how well a person will perform on items designed to measure that ability (Wu, 2013). In contrast, CTT does not make any assumption about a latent ability that determines performance of a person on items or test. The Rasch model was originally developed for the analysis of dichotomous data. Along with its development at the University of Chicago, several models are availble and can be used to analyze polytomous data, such as the Rating Scale Model and Partial Credit Model.

Since it was introduced in Indonesia by the late Bruce H. Choppin in 1975 (Nasoetion et al., 1976), this model has been used by Indonesian researchers studied in the United States (Hayat, 1992; Umar, 1987). In the Indonesian context, the Rasch model was introduced for use in large-scale assessment data such as national survey of achievement, and national examination. The Rasch model was also used in TIMSS 1995 and PISA 2000-2015 (von Davier, 2020) and the 2023. The reason for choosing this model is because Rasch measurement models have a feature that distinguishes them from other modern test theory models (e.g., IRT), namely specific objectivity, referring to the principle that comparisons between two objects must be free from the conditions under which the comparisons are made; sufficiency, referring to the statistical property that students with the same raw score will be given the same ability estimate in logits, irrespective of which items they answer correctly on the test (Wainer et al., 1979; Wu et al., 2016).

Technically, in the Rasch model, the parameters of persons, items, and threshold structures are expressed on a log-odds unit (logit). That is, the calibration of items and estimates of person abilities can be compared with each other on the continuum line of the same scale. Consequently, predictive modeling of one's response to an item can be done for dichotomous, polytomous or mixed-format data (Andrich & Marais, 2019; Boone, 2020; Suryadi et al., 2020). The suitable model that can be used for the mixed-format nature of INAP-L is PCM. The basic equation of the PCM is as follows (de Ayala, 2022):

$$\mathrm{P}\left(X_j|\theta,\delta_{\mathrm{jh}}\right)=\frac{\exp\left[\sum_{\mathrm{h}=0}^{\mathrm{x_j}}(\theta-\delta_{\mathrm{jh}})\right]}{\sum_{k=1}^{m_j}exp\left[\sum_{h=0}^{k}(\theta-\delta_{jh})\right]} \quad (1)$$

Where θ is a person ability parameter, δ_{jh} is the h-th threshold parameter for the j-th item. The latter describes the level of relative difficulty in a category h to be selected compared to category (h-1). The use of subscripts on m (that is, mj) indicates that the number of categories can vary for item to item. However, to have the unique characteristics of the Rasch model, certain requirements must be met, namely (Hayat et al., 2023; Yu, 2020): (1) unidimensionality, the instrument measures only one trait. The assumption of unidimensionality was tested in this study using a parallel analysis method based on minimum rank factor analysis (PA-MRFA; Timmerman & Lorenzo-Seva, 2011); and (2) local independence, which means that the test taker's response to an item must be independent of other items or persons. The Q3 method (Yen, 1984) was used in this study to test the requirement of local independence.

In addition to testing assumptions, in implementing the Rasch model, at the item level, the fit indices used are Infit and Outfit. The Infit and Outfit Mean Square (MNSQ) values are used to identify the misfit of items to the model. The expected value of Infit or Outfit for each item is 1.0, with an acceptable range of values between 0.7 to 1.3 (Smith et al., 2008). In addition, the discriminating power index which is similar to the CTT, namely the PTMEA (point-to-measure) correlation, is used as an indicator of item discrimination where the negative value indicates that the item does not fit the model.

2.4 *Differential Item Functioning (DIF) analysis based on the Rasch model*

Furthermore, in addition to fulfilling the requirement of unidimensionality and local independence, the Rasch analysis must fulfill another requirement, namely the postulate that the test is free of differential item functioning (DIF) (Kreiner, 2013). DIF can be problematic for measuring instruments because it identifies items that perform differently across different sample characteristics (Smith et al., 2016). Furthermore, bias can also occur at the test level, namely differential test functioning (DTF), or at both the item and test levels concurrently, called differential functioning of items and test (DFIT) (Temel et al., 2022). Nonetheless, in this study, we focused solely on DIF analysis.

In this study, given the fact that various studies have found gender differences in the language abilities of Indonesian students (e.g., Syahputra et al., 2022; Wahyuningsih, 2018), as well as our suspicion that there are differences in the functioning of items based on the type of school, we conducted a Rasch-based DIF evaluation for two types of subgroups (female vs. male and public school vs. private school). In the modern psychometric approach, especially the Rasch model, DIF analysis can be performed using various methods. One of the most commonly used is the Rasch-Welch t-test. In this procedure, item difficulty parameters were estimated individually for a reference group and focal groups through logistic regression. Subsequently, the differences in item difficulties across the groups could be tested for statistical significance. The formula for the Rasch-Welch t-test is shown in Equation 2 (Smith et al., 2016):

$$t=\frac{\mathrm{d}_{i2}-\mathrm{d}_{i1}}{\sqrt{s_{i2}^2-s_{i1}^2}} \quad (2)$$

where d_{i2} is the level of difficulty of the i-th item for the j-th group and s_{i2}^2 is the standard error of estimate of the i-th item for the j-th group. In using this statistic, the magnitude of logit differences (DIF effect size) is the main indicator of whether an item is experiencing DIF or not. Significant t values and DIF effect sizes greater than 0.40 are other indicators that can be used to detect DIF (Choi et al., 2006; Smith et al., 2016).

In this study, two software packages were used for analyzing the INAP-L. The Winsteps 3.73 program (Linacre, 2018) using the joint maximum likelihood (JMLE) estimation method was used for the Rasch analysis including item calibration and DIF analysis. The 'EFA.MRFA' (Timmerman & Lorenzo-Seva, 2011) package in the Rstudio program was used to test the unidimensionality requirement of the mixed-format test.

2.5 Multilevel Modeling

Although the focus of this study is to calibrate the INAP-L items and generate interval-scale person abilities using the Rasch model, we have to take into account the nature of INAP-L data that were collected using the multistage sampling method, which empirically multilevel consisting of provincial, school and student level. To determine whether there is an effect of clustering on the data, such as whether students from the same school or province tend to be similar to one another compared to students from different schools or provinces, multilevel modeling is required even though the person ability is generated from a single level Rasch model without considering the clustering of the data (Raudenbush & Bryk, 2002). We decided that provinces were treated as clusters or the level 2 unit of analysis for this study.

To determine whether multilevel modeling is appropriate for use in the INAP-L data analysis, the intercept only model, the simplest model of multilevel regression was used. The basic equation for the intercept model is (Heck & Thomas, 2015; Raudenbush & Bryk, 2002):

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad (3)$$

Where Y_{ij} is the INAP-L scores for each respondent, β_{0j} is the mean of INAP-L score for the j-th province, and ε_{ij} is the residual component for the i-th student in the j-th province. All errors in level-1 (student level) are assumed to be 0 with a variance that does not vary between provinces (each province has the same error variance). Meanwhile, subscript j indicates that the intercept is the mean INAP-L score for each province. Whereas in the next equation, at level 2 (between provinces), the variance of the average INAP-L score (β_{0j}) is assumed to vary between units of analysis (there is a variance at the student level and a variance at the provincial level):

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (4)$$

Where γ_{00} is the fixed-effect coefficient at the provincial level, namely the mean of INAP-L scores at student levels across all provinces (unit level 2), while u_{0j} is the deviation of the j-th provincial mean of the grand mean of INAP-L. u_{0j} is also known as the random intercept effect (Raudenbush & Bryk, 2002). By combining the two previous equations into one equation, one equation is produced that is used in this study, namely:

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \quad (5)$$

Where ε_{ij} is the residual variance of the student level, u_{0j} can be called a random effect at the provincial level and γ_{00} is the grand mean (Heck & Thomas, 2015). This model will produce an average INAP-L score for all provinces, with two components of variance, namely the variance at the student level (level-1) and the variance at the provincial level (level 2). The information from the variance component that has been "separated" between the two levels of analysis is used to calculate the intraclass correlation coefficient (ICC). In simple terms, ICC is a quantification measure that shows the amount of mutual similarity of scores between individuals in the same cluster (Kreft & de Leeuw, 1998). The ICC formula is:

$$\rho = \frac{\sigma_{\eta_B}^2}{\sigma_{\eta_B}^2 + \sigma_{\eta_W}^2} \quad (6)$$

Where ρ is the ICC, $\sigma_{\eta_B}^2$ is the component of the provincial level variance and $\sigma_{\eta_W}^2$ is the student level variance within the province (Stapleton et al., 2016). After the ICC is generated, the ICC value is used to calculate the design effect (deff) (Muthen & Satorra, 1995). Deff describes quantifies the extent to which the sampling error present in sampling individuals in a sampling design departs from the sampling error that would be expected under simple random sampling (i.e., where each individual had the same chance of being selected). Where clustering is present within level-2 units, individuals will no longer be independent of others selected in the same cluster. This lack of independence can lead to more findings of statistical significance than would be expected under conditions of simple random sampling (Heck & Thomas, 2015). The formula for deff is:

$$deff = [1 + (\text{average cluster size} - 1 \times \rho)] \quad (7)$$

If the deff value is more than 2.0, then the multilevel analysis must be carried out to analyze INAP-L data, although it should be understood that ICC can also be used to make decisions on whether a multilevel model should be used to examine relationships among variables (for example, Putra et al., 2017). All stages used in the multilevel analysis in this study were carried out using the Mplus 8.3 program with the robust maximum likelihood estimation (ESTIMATOR = MLR) method.

3. Results and Discussion

3.1 Unidimensionality and Local Independence

Because INAP-L is theorized as a unidimensional construct, a single score will be generated. However, because the INAP-L is a mixed format test, there are challenges in testing the requirement of unidimensionality that necessitates specialized procedures (e.g., Zhang, 2016). The results of PA-MRFA revealed that there was one main factor of INAP-L with a variance of 41.84%, which is significantly higher than the variances of other three minor dimensions, with a variance proportion of 7.46, 7.36, and 6.84%, respectively. These findings support the Rasch model's unidimensionality requirement of the INAP-L.

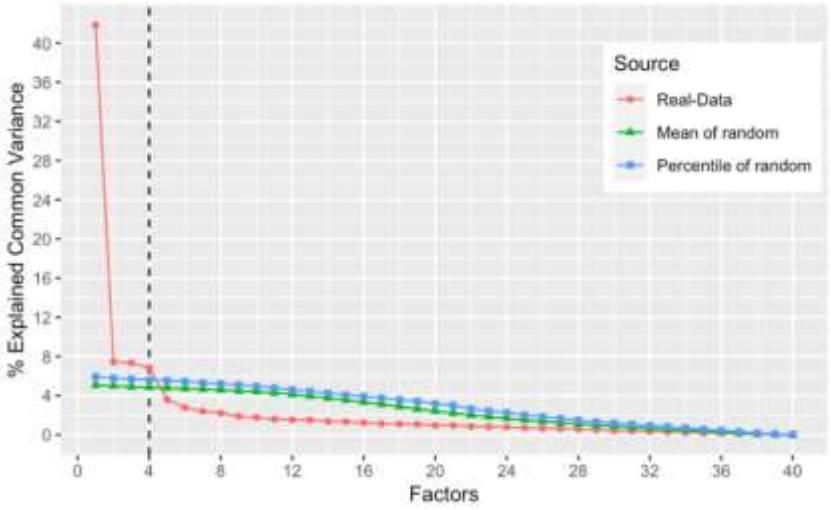

Figure 1. Scree plot from dimensionality assessment

After the unidimensionality requirement was met, the local independence of the INAP-L was tested. These two requirements are related to each other; once unidimensionality is proven, responses to each item will have distinct characteristics from other items measuring the same factor (Yu, 2020). Table 1 shows the results of testing the local independence requirement for this study.

Table 1. Local independence testing of the INAP-L

Q3	Item	Code	Item	Code
0.20	Item 24	Perbandingan_Musim_02	Item 26	Perbandingan_Musim_04
0.20	Item 3	Batik_03	Item 15	Penyakit_Vektor_02
0.19	Item 35	BIN_2017_BINT03_03	Item 36	BIN_2017_BINT03_04
0.18	Item 17	Penyakit_Vektor_04	Item 18	Penyakit_Vektor_05
0.15	Item 3	Batik_03	Item 37	BIN_2017_BINT05_01
-0.15	Item 17	Penyakit_Vektor_04	Item 24	Perbandingan_Musim_02
-0.14	Item 16	Penyakit_Vektor_03	Item 24	Perbandingan_Musim_02
-0.13	Item 37	BIN_2017_BINT05_01	Item 38	BIN_2017_BINT05_02
-0.13	Item 6	Batik_06	Item 25	Perbandingan_Musim_03
-0.13	Item 3	Batik_03	Item 38	BIN_2017_BINT05_02

With the criterion that the Q3 between pairs of items is not > 0.25 (DeMars, 2010), as can be seen in Table 4 there is no items that experience local dependence. Items that have the highest residual correlation (Q3=0.20) are item 24 and item 26 pair, item 3 and item 15 pair, which is below the threshold of 0.25. In other words, the requirement of local independence in this study was fulfilled. These findings indicate that the developers of the INAP-L have succeeded in developing items that do not have very high similarity in item construction and wording (Bandalos, 2021) measuring the same construct.

3.2 Item Measure, Fit Statistics, and Threshold Parameter

Table 2 contains an overview of the psychometric characteristics of the INAP-L instrument, including fit statistical tests, item difficulty level, and step parameters for all items. As can be seen in the table, there are four items whose values are unacceptable, namely item 37 because the PTMEA value was found to be negative, item 1 found the MNSQ outfit value of 1.39 > 1.30, item 3 found the MNSQ outfit value of 1.44 > 1.30, and item 15 found an MNSQ outfit value of 1.34 > 1.30. In addition, all items show acceptable MNSQ infit and outfit (0.70-1.30). The difficulty level of the items is in a symmetrical range of values (-2.29 to 2.33) and it is found that item 37 (BIN_2017_BINT05_01) with a location at 2.33 logit is an item that is difficult for the respondent to answer correctly, and item 35 (BIN_2017_BINT03_03) with a location at -.2.29 logit is the easiest item to answer correctly.

Table 2. Item calibration results of INAP-L

Item	Item ID	δ	Infit	Outfit	PTMEA	Threshold (τ)		
						τ_1	τ_2	τ_3
37	BIN_2017_BINT05_01	2.33	1.13	2.00*	-0.14*			
14	Penyakit_Vektor_01	2.17	0.98	0.88	0.24			
20	Penyakit_Vektor_07	1.86	0.90	0.72	0.40	1.31	2.37	
27	Perbandingan_Musim_05	1.76	0.94	0.82	0.33			
30	BIN_2017_BINT01_02	1.55	0.93	0.81	0.36			
2	BATIK_02	1.55	0.97	0.95	0.28			
31	BIN_2017_BINT01_03	1.18	0.97	0.93	0.31			
18	Penyakit_Vektor_05	1.10	0.92	0.87	0.45	0.33	1.88	
1	BATIK_01	1.02	1.11	1.39*	0.34	0.39	0.86	1.75
13	Laskar_Pelangi_06	1.02	0.87	0.83	0.51	0.09	1.94	
22	Penyakit_Vektor_09	0.82	0.92	0.88	0.40			
3	BATIK_03	0.76	1.26	1.44*	-0.12*			
17	Penyakit_Vektor_04	0.76	0.92	0.90	0.46	0.14	1.38	
10	Laskar_Pelangi_03	0.69	1.10	1.15	0.14			
19	Penyakit_Vektor_06	0.67	1.12	1.20	0.10			
15	Penyakit_Vektor_02	0.44	1.24	1.34*	-0.05*			
25	Perbandingan_Musim_03	0.29	0.95	0.95	0.52	-0.71	0.45	1.16
23	Perbandingan_Musim_01	0.04	1.17	1.20	0.06			
16	Penyakit_Vektor_03	-0.05	0.88	0.88	0.52	-1.41	1.30	
40	BIN_2017_BINT05_04	-0.11	0.93	0.91	0.41			
6	BATIK_06	-0.13	1.17	1.22	0.43	-0.85	-0.17	0.61
32	BIN_2017_BINT01_04	-0.29	0.97	0.97	0.35			
28	Perbandingan_Musim_06	-0.36	0.88	0.86	0.55	-0.99	0.27	
26	Perbandingan_Musim_04	-0.47	1.19	1.23	0.27	-2.18	-0.76	1.53
33	BIN_2017_BINT03_01	-0.52	0.94	0.93	0.40			
4	BATIK_04	-0.61	0.92	0.90	0.42			
36	BIN_2017_BINT03_04	-0.85	0.90	0.88	0.44			
38	BIN_2017_BINT05_02	-0.87	0.94	0.92	0.39			
9	Laskar_Pelangi_02	-0.89	0.91	0.89	0.51	-1.27	-0.51	
21	Penyakit_Vektor_08	-0.91	1.02	1.01	0.28			
24	Perbandingan_Musim_02	-0.98	1.21	1.24	0.15	-2.18	0.22	
12	Laskar_Pelangi_05	-1.07	0.93	0.94	0.44	-2.43	0.29	
7	BATIK_07	-1.08	1.10	1.23	0.31	-1.84	-0.32	
34	BIN_2017_BINT03_02	-1.17	0.96	0.94	0.36			
29	BIN_2017_BINT01_01	-1.30	0.98	0.98	0.31			
11	Laskar_Pelangi_04	-1.34	0.92	0.87	0.41			
8	Laskar_Pelangi_01	-1.39	0.93	0.89	0.38			
5	BATIK_05	-1.45	0.91	0.85	0.41			
39	BIN_2017_BINT05_03	-1.88	0.96	0.92	0.28	-5.84	2.09	
35	BIN_2017_BINT03_03	-2.29	0.94	0.90	0.32			

Furthermore, for 15 polytomous items, the results of PCM analysis show that all step parameter values increased from low to high, as expected by the model. Furthermore, none of the step parameters are reversed. These findings indicate that in the INAP items, which have a polytomous scoring format, the response options function as expected.

3.3 Separation Reliability Indices

The Rasch model does not use the same concept of reliability as the classical approach. The Rasch model estimates the reliability of both test items and persons (Wright & Masters, 1982). Person separation reliability (PSR) is an estimate of how well the instrument can categorize respondents based on the trait measured. In other words, PSR can describe the INAP-L instrument's internal consistency. Item separation reliability (ISR), on the other hand, indicates how well respondents can categorize items based on item difficulty hierarchy. The results of the analysis show that the INAP-L PSR is 0.80, and the ISR is 1.00. As additional information, the reliability coefficient from the CTT perspective in the form of Ordinal Cronbach's alpha (Ordinal α) is also reported at 0.81. These findings indicate that the internal consistency of the INAP-L instrument is very good. Furthermore, the high ISR findings also indicate that the sample size in this study is sufficient to confirm the distribution of item difficulty levels, as well as the high PSR indicates that this instrument is sensitive enough to distinguish respondents with high literacy and respondents with low literacy (Linacre, 2018).

3.4 Wright Map: Person-Item Targeting

Having previously presented information regarding the results of item parameter estimation, the relationship between the level of latent trait of test takers and the level of item difficulty can be compared simultaneously using the Wright Map (Wilson & Draney, 2002). The Wright Map results of the analysis of the INAP-L instrument can be seen in Figure 2 below:

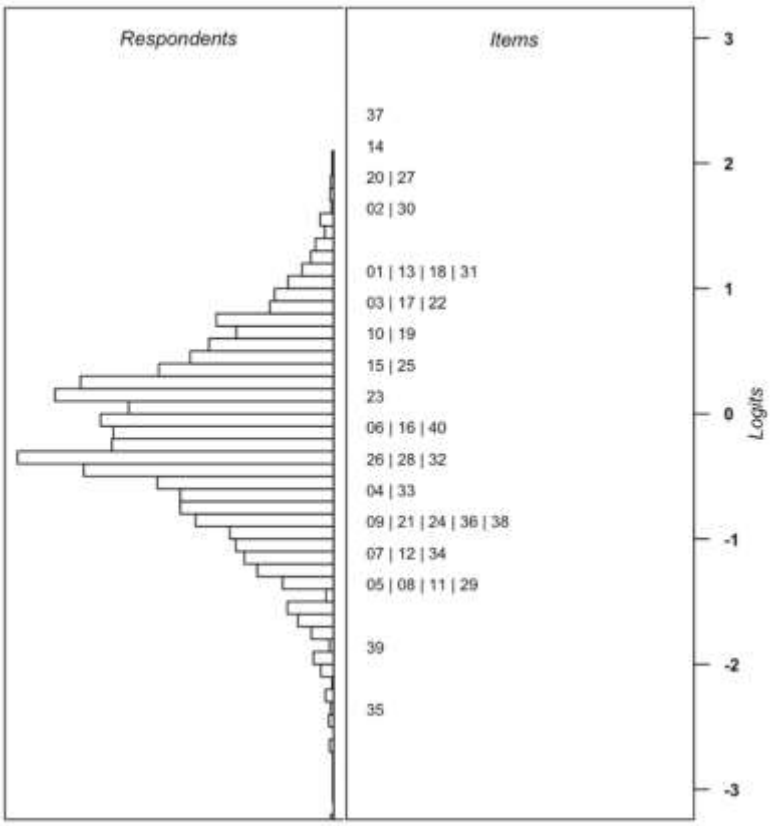

Figure 2. Wright Map of the INAP-L

As can be seen in Figure 2, item 37 is the most difficult item, and item 35 is the easiest item. Next, it can be seen that the average respondent's trait level is -0.1862 (SD = 0.6948) is slightly lower than the average item difficulty level of 0 (SD = 1.17) in the original logit scale. The ability range of respondents is -3.230 logit to 2.100 logit. After the transformation of the original logit to the mean of 500 and SD 100 was performed, the range of the score has a minimum value of 177 and a maximum value of 710. The mean of student ability is 481.3775 with an SD of 69.482. With a small mean difference, it was found that the INAP-L instrument was on-target for the sample in this study. In addition, it can be seen that of the 40 items, only two items, namely items 37 and 14, whose positions are outside the range of the respondent's abilities. This shows that the targeting person-item of the INAP-L instrument is good and can function optimally in measuring language literacy.

3.5 DIF Analysis

Table 3 contains the results of the DIF analysis on the INAP-L instrument based on two background variables, namely gender (female vs. male) and school type (public vs. private). Based on gender differences, we found that none of the items had DIF using predetermined criteria (significant t and DIF contrast > 0.40). This means that all INAP-L items are gender DIF-free. Meanwhile, based on differences in school type, we found that two items exhibited DIF, namely item 30 (BIN_2017_BINT01_02) with DIF contrast = -0.41 and item 39 (BIN_2017_BINT05_03) with DIF contrast = -0.43 (See Table 3).

Table 3. DIF analysis results of the INAP-L

Item	Item ID	Gender (Female vs male)			School type (Public vs private)		
		DIF contrast	SE	t	DIF contrast	SE	t
1	BATIK_01	0.17	0.03	4.96	0.23	0.03	6.95
2	BATIK_02	0.15	0.07	2.07	0.31	0.07	4.33
3	BATIK_03	0.21	0.06	3.51	0.14	0.06	2.38
4	BATIK_04	0.00	0.06	0.00	-0.18	0.06	-3.19
5	BATIK_05	-0.14	0.06	-2.16	0.00	0.06	0.00
6	BATIK_06	0.10	0.03	3.66	-0.07	0.03	-2.48
7	BATIK_07	-0.08	0.04	-1.87	-0.17	0.04	-3.97
8	Laskar_Pelangi_01	-0.31	0.06	-4.89	-0.02	0.06	-0.32
9	Laskar_Pelangi_02	-0.09	0.04	-2.46	0.15	0.04	4.12
10	Laskar_Pelangi_03	0.00	0.06	0.00	-0.13	0.06	-2.21

11	Laskar_Pelangi_04	-0.06	0.06	-1.01	-0.26	0.06	-4.27
12	Laskar_Pelangi_05	-0.10	0.05	-2.08	-0.06	0.05	-1.30
13	Laskar_Pelangi_06	0.00	0.05	0.00	-0.19	0.05	-4.19
14	Penyakit_Vektor_01	0.11	0.09	1.29	-0.16	0.09	-1.80
15	Penyakit_Vektor_02	0.14	0.06	2.46	0.06	0.06	1.12
16	Penyakit_Vektor_03	-0.23	0.04	-5.03	0.00	0.04	0.00
17	Penyakit_Vektor_04	0.00	0.04	0.00	0.17	0.04	4.27
18	Penyakit_Vektor_05	0.00	0.05	0.00	-0.24	0.05	-5.28
19	Penyakit_Vektor_06	0.12	0.06	2.09	-0.06	0.06	-1.04
20	Penyakit_Vektor_07	0.18	0.06	3.05	0.28	0.06	4.83
21	Penyakit_Vektor_08	-0.19	0.06	-3.36	0.25	0.06	4.25
22	Penyakit_Vektor_09	0.06	0.06	0.96	0.00	0.06	0.00
23	Perbandingan_Musim_01	0.19	0.06	3.42	0.06	0.06	1.03
24	Perbandingan_Musim_02	0.08	0.04	1.79	0.00	0.04	0.00
25	Perbandingan_Musim_03	0.14	0.03	4.51	0.21	0.03	6.82
26	Perbandingan_Musim_04	0.13	0.04	3.33	0.00	0.04	0.00
27	Perbandingan_Musim_05	-0.02	0.08	-0.29	-0.07	0.08	-0.90
28	Perbandingan_Musim_06	-0.08	0.04	-2.03	-0.16	0.04	-4.18
29	BIN_2017_BINT01_01	-0.22	0.06	-3.61	-0.02	0.06	-0.34
30	BIN_2017_BINT01_02	-0.22	0.07	-2.90	-0.41*	0.07	-5.54
31	BIN_2017_BINT01_03	-0.07	0.07	-1.00	-0.09	0.07	-1.29
32	BIN_2017_BINT01_04	-0.13	0.06	-2.37	0.00	0.05	0.00
33	BIN_2017_BINT03_01	-0.09	0.06	-1.51	0.00	0.06	0.00
34	BIN_2017_BINT03_02	0.00	0.06	0.00	0.00	0.06	0.00
35	BIN_2017_BINT03_03	-0.24	0.08	-3.03	0.00	0.08	0.00
36	BIN_2017_BINT03_04	-0.16	0.06	-2.75	-0.02	0.06	-0.41
37	BIN_2017_BINT05_01	0.23	0.09	2.49	0.14	0.09	1.50
38	BIN_2017_BINT05_02	0.00	0.06	0.00	-0.11	0.06	-1.94
39	BIN_2017_BINT05_03	-0.26	0.13	-1.98	-0.43*	0.13	-3.28
40	BIN_2017_BINT05_04	-0.20	0.06	-3.62	0.00	0.06	0.00

Given the direction of the negative effect size, this finding shows that the two items function differently for students from public and private schools, where students from private schools tend to have an advantage since these two items are statistically easier for them. Another fact is that the two items containing the DIF are linking items of INAP-L to an international assessment. We suspect that, unlike students in public schools, students in private schools have been exposed to the content of the international assessment items.

3.6 Comparative Analysis Across Provinces: Descriptive Findings

In addition to the psychometric analysis of INAP-L, descriptive statistics containing the INAP-L scores for each province was reported. For this purpose, we transformed the scores in the original scale (logit) into the new scale with a mean of 500 and the SD of 100. This transformation was performed to avoid negative values and makes the interpretation of the score easier. Table 4 contains each province's mean, minimum and maximum scores, and SD.

Table 4. Mean scores of the INAP-L for all provinces (ordered by highest mean scores)

No	Provinces	Mean	Minimum	Maximum	SD
1	South Sumatera	569.797	336	710	74.650
2	Riau	530.773	414	652	45.553
3	Banten	518.535	255	671	66.731
4	West Kalimantan	515.708	322	658	61.774
5	South Sulawesi	508.630	336	671	55.954
6	East Kalimantan	507.076	232	637	80.261
7	Central Kalimantan	505.979	349	648	59.413
8	East Nusa Tenggara	504.344	177	677	95.545
9	Riau Islands	503.393	274	608	45.131
10	Bengkulu	500.525	308	689	71.385
11	Maluku	493.539	326	665	55.132
12	Bali	492.640	336	591	45.033
13	Central Java	489.399	349	659	57.907
14	North Sumatera	486.604	292	671	73.501
15	West Java	485.280	349	627	50.705
16	East Java	483.682	268	630	77.056
17	Lampung	481.872	361	668	62.782
18	West Nusa Tenggara	481.527	228	659	82.303
19	South Kalimantan	480.577	336	608	55.622
20	Special Capital Region of Jakarta	477.389	216	595	58.089
21	Papua	475.893	320	617	54.991
22	Jambi	471.622	308	601	60.843

23	North Sulawesi	471.004	274	617	54.207
24	Special Region of Yogyakarta	470.541	232	673	76.213
25	Bangka Belitung	459.938	292	589	56.899
26	Southeast Sulawesi	459.623	199	617	60.736
27	Aceh	458.138	274	620	66.473
28	Central Sulawesi	453.961	274	598	55.017
29	West Sulawesi	449.729	274	610	60.779
30	West Sumatera	448.762	232	591	59.504
31	Gorontalo	445.522	295	627	55.989
32	West Papua	444.109	236	586	69.739
33	North Maluku	423.822	206	582	71.198
34	North Kalimantan	423.261	292	551	52.693

As seen in Table 4, the province with the highest average INAP-L score is South Sumatra, which is quite different from the national capital, namely the Special Capital Region of Jakarta. Meanwhile, the province with the lowest average INAP-L score is North Kalimantan, which is the "youngest" province in Indonesia. However, the information we produce in this study is only descriptive, considering that no other variables can be used as predictors to explain why the INAP-L scores vary or what causes them at the student or teacher level. Future research can explore the factors that influence the INAP-L score so that this variation can be explained.

In addition, as the Ministry of Education and Culture officially implemented the INAP-L, many language researchers in Indonesia have researched language learning and assessment models, such as researchers from East Kalimantan (Sudrajat, 2019), East Java (Kocimaheni et al. , 2022), Bali (Gotama & Astini, 2022), Lampung (Yumithasari et al., 2022), West Java (Afipah, 2022). This latest study demonstrates that language assessment is a well-studied topic. We hope that by publishing our research findings in the form of INAP-L scores for all provinces in Indonesia, future studies will use the findings of this study as early indicators for diagnosing learning difficulties and developing appropriate language learning strategies. This is consistent with the findings of studies that emphasize the importance of effective language learning in schools in line with language skills measured by national and international assessments (Meriana & Murniarti, 2021).

3.7 Multilevel modeling

The multilevel intercept model analysis of the INAP-L score (theta) resulting from the Rasch analysis was conducted on two units of analysis, namely 6153 students (level-1) nested in 34 provinces (level-2). The results of the analysis show that the average number of respondents per province (average cluster sizes) is 180,971 people. On the total score scale, the estimated variance at the student level was 4011,404, the variance at the school level was 2391,284, and the mean INAP-L score between provinces (national mean) was 481,666. With this variance, an ICC of 0.179 is obtained, resulting in a deff = 33.215, which is significantly higher than the cutoff of 2.00. These findings suggest that if the INAP-L score is to be tested for interrelationships with other variables from non-cognitive measurements (e.g., opportunity-to-learn, motivation, parental support), a multilevel approach (e.g., multilevel regression) must be used. When the INAP-L score is analyzed using a single-level approach, biased estimation results are produced, and no proper inferences can be drawn.

Aside from that, the Center for Educational Assessment can take one of two approaches when processing the INAP-L data. The first approach is to calibrate using the multilevel Rasch model (e.g. Lamprianou, 2013). This approach, however, has extremely complicated computational procedures and interpretations that necessitate careful mathematical considerations. Although the multilevel Rasch model has been extensively researched in academia (see, DeMars, 2020; Raudenbush et al., 2003), its application in LSA decision-making has yet to be tested. While the second method is similar to the method used in this study, in which the calibration was produced using a single-level Rasch analysis approach, further analysis of the relationship between variables requires a multilevel analysis (Luppescu, 2013; Thompson, 2005). Furthermore, when compared to PISA or TIMSS, which have the same sampling design (multistage sampling) but still use single-level analysis in the item calibration process, the second approach is justified.

4. Conclusions

Based on the results of the psychometric analysis, it can be concluded that the INAP-L instrument has good psychometric properties. The objective measurement requirements for the Rasch measurement model are fulfilled. INAP-L developers successfully developed language literacy assessment instruments that fulfilled local independence assumptions and are free from DIF-gender. Although there are items that experience DIF due to school type, however, the number is very small. In addition, the application of the Rasch measurement model in this study confirms the legacy of the Rasch model in terms of its usefulness in large-scale assessments, especially in Indonesia since it was first introduced in 1976 at the Ministry of Education and Culture of the Republic of Indonesia. Further, this is the first study that conducted the construct validity assessment of the INAP-L published in a scientific journal. This study is also very unique in measuring the construct of language assessment literacy, which has still not received adequate attention from Indonesian academia.

Our suggestion for future research is to investigate the relationship between language competency and other non-cognitive variables to explain differences in INAP-L achievement across provinces. INAP-L data can be used for further research by requesting permission to use data officially from the Ministry of Education and Culture. Furthermore, the implications of this INAP-L research are for teachers and schools in diagnosing language learning and designing appropriate learning strategies based on the map of learning competency strengths and weaknesses. Language learning in schools must also adapt to the material and competencies that will be tested in national and international assessments such as INAP-L and PISA Language Assessment.

References

Abbasi, M. G., Ahmad, A., & Khattak, Z. I. (2010). Negative influence of large scale assessment on language learning strategies of the Secondary School Certificate (SSC) students. *Procedia – Social and Behavioral Sciences, 2*(2), 4938-4942. https://doi.org/10.1016/j.sbspro.2010.03.799

Adams, R. J. (2003). Response to 'cautions on OECD's recent educational survey (PISA)'. *Oxford Review of Education, 29*(3), 377-389. https://doi.org/10.1080/03054980307445

Afipah, H. (2022). Perkembangan bahasa anak usia 4 tahun melalui asesmen observasi di TK Sejahtera Kota Bekasi. *Jurnal Cemerlang PAUD, 1*(1), 33-42. https://doi.org/10.46368/.v1i1.471

Almuzakir, & Qamariah, Z. (2019). *Gender differences and students' verbal communication*. In Proceedings of the the 3rd International Conference on English Language Teaching (INACELT).

Álvarez, I. A. (2013). Large-scale assessment of language proficiency: Theoretical and pedagogical reflections on the use of multiple-choice tests. *International Journal of English Studies, 13*(2), 21-38. https://doi.org/10.6018/ijes.13.2.185861

Andrich, D., & Marais, I. (2019). *A course in rasch measurement theory: measuring in the educational, social and health sciences*. Singapore: Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-7496-8

Atmawati, D. (2018). Gender bias in Javanese society: A study in language forms choice to men and women. *Humaniora, 9*(3), 257-264. https://doi.org/10.21512/humaniora.v9i3.4937

Bandalos, D. L. (2021). Item meaning and order as causes of correlated Residuals in Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal.* Advanced online publications. https://doi.org/10.1080/10705511.2021.1916395

Blomeke, S., & Gustafsson, J. E. (2017). *Standard setting in education: The Nordic countries in an international perspective*. Cham, Switzerland: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-50856-6

Boone, W. J. (2020). Rasch basics for the novice. In M. S. Khine (Ed.), *Rasch measurement: Applications in quantitative educational research* (pp. 9-30). Singapore: Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-15-1800-3_2

Davies, A. (2013). Fifty years of language assessment. In A. J. Kunnan (ed.), *The Companion to Language Assessment, First edition, Vol. III* (pp. 3-21), Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118411360.wbcla127

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). New York, NY: Guilford Press.

DeMars, C. (2010). *Item response theory.* Oxford, England: Oxford University Press, Inc. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

DeMars, C. E. (2020). Multilevel Rasch Modeling: Does misfit to the Rasch model impact the regression model? *The Journal of Experimental Education, 88*(4), 605-619. https://doi.org/10.1080/00220973.2019.1610859

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511732911

Erickson, G. & Åberg-Bengtsson, L. (2012). A Collaborative Approach to National Test Development. In D. Tsagari, & I. Csépes (Eds), *Collaboration in Language Testing and Assessment* (pp. 93-108). Bern, Switzerland: Peter Lang.

Erickson, G. (2010). Good practice in language testing and assessment: A matter of responsibility and respect. In T. Kao & Y. Lin (Eds.), *A New Look at Teaching and Testing: English as Subject and Vehicle* (pp. 237–258). Taipei, Taiwan: Bookman Books.

Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: an ironic investigation. *Journal of Education Policy,* 29(6), 845-855. https://doi.org/10.1080/02680939.2014.892156

Geisinger, K. F. (2016). 21st century skills: What are they and how do we assess them? *Applied Measurement in Education,* 29(4), 245-249. https://doi.org/10.1080/08957347.2016.1209207

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice, 11*(3), 319-330. https://doi.org/10.1080/0969594042000304618

Goldstein, H. (2017). Measurement and evaluation issues with PISA. In L. Volante (Ed.), *The PISA Effect on Global Educational Governance (*pp. 49-58). Milton Park, UK: Routledge. https://doi.org/10.4324/9781315440521-4

Gotama, P. A. P., & Astini, N. K. S. (2022). Asesmen otentik dalam pembelajaran Bahasa Indonesia. *Lampuhyang, 13*(2), 43-57. https://doi.org/10.47730/jurnallampuhyang.v13i1.88

Hayat, B. (1992). *A mathematics item bank for Indonesia*. Doctoral dissertation, University of Chicago.

Hayat, B., Rahayu, W., Putra, M. D. K., Sarifah, I., Deviana, T., Puri, V. G. S., & Isa, K. (2023). Metacognitive Skills Assessment in Research-Proposal Writing (MSARPW) in the Indonesian university context: Scale development and validation using multidimensional item response models. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, 12*(1), 31-47. https://doi.org/10.15408/jp3i.v12i1.31679

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques* (3rd ed.). New York, NY: Routledge. https://doi.org/10.4324/9781315746494

Kocimaheni, A. A., Laksono, K., Mintowati., & Nurhadi, D. (2022). Literasi asesmen bahasa calon guru Bahasa Jepang: Persepsi dan praktiknya. *Jurnal Pendidikan Bahasa Jepang*, *8*(1), 19-26.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage. https://doi.org/10.4135/9781849209366

Kreiner, S. (2013). The Rasch model for dichotomous items. In K. B. Christensen, S. Kreiner, & M. Mesbah. (Eds.), *Rasch models in health*. Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118574454.ch1

Kremmel. B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the Language Assessment Literacy Survey. *Language Assessment Quarterly, 17*(1), 100-120. https://doi.org/10.1080/15434303.2019.1674855

Kunnan, A. (2017). Large-scale language assessments: Empirical studies. In E. Hinkel (Ed.), *Handbook of research on second language learning*. New York, NY: Routledge. https://doi.org/10.4324/9781315716893-34

Kunnan, A. J. (2008). Large-scale language assessment. In E. Shohamy & N. Hornberger (Eds.) *Encyclopedia of language and education, 2nd Edition, Volume 7: Language testing and assessment* (pp. 135-155). Berlin, Germany: Springer Science.

Kunnan, A. J., & Grabowski, K. (2013). Large scale second language assessment. In M. Celce-Murcia et al. (Eds.), *Teaching English as a second or foreign language*, 4th edition (pp. 304-319). Boston, MA: Heinle/Cengage.

Lamprianou, I. (2013). Application of single-level and multi-level Rasch models using the lme4 package. *Journal of Applied Measurement, 14*(1), 79-90.

Liu, K. K., Ward, J. M., Thurlow, M. L., & Christensen, L. L. (2017). Large-Scale Assessment and English Language Learners With Disabilities. *Educational Policy*, *31*(5), 551-583. https://doi.org/10.1177/0895904815613443

Luppescu, S. (2013). Using Rasch measures in a multi-level context. *Rasch Measurement Transactions,* 27(2), 1413-1414.

Maehler, D. B., & Rammstedt, B. (2020). *Large-scale cognitive assessment: Analyzing PIAAC data*. Cham, Switzerland: Springer International Publishing AG. https://doi.org/10.1007/978-3-030-47515-4

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*(4), 529-544. https://doi.org/10.1007/BF02302590

McFarlane, D. A. (2013). Understanding the challenges of science education in the 21st century: New opportunities for scientific literacy. *International Letters of Social and Humanistic Sciences, 4,* 35-44. https://doi.org/10.18052/www.scipress.com/ILSHS.4.35

Meriana, T., & Murniarti, E. (2021). Analisis pelatihan asesmen kompetensi minimum. *Jurnal Dinamika Pendidikan, 14*(2), 110-116. https://doi.org/10.51212/jdp.v14i2.7

Moegiadi., Mangindaan, C., & Elley, W. B. (1979). Evaluation of achievement in the Indonesian education system. *Evaluation in Education: International Progress, 2*(4), 282-351. https://doi.org/10.1016/0145-9228(79)90001-3

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267-316. https://doi.org/10.2307/271070

Nasoetion, N., Djalil, A., Musa, I., Soelistyo, S., Choppin, B. H., & Postlethwaithe, T. N. (1976). *The development of educational evaluation models in Indonesia*. Jakarta, Indonesia: Office of Educational and Cultural Research and Development (BP3K) of the Ministry of Education and Culture.

Nasrullah, N., Ainol, A., & Waluyo, E. (2022). Analisis kemampuan numerasi siswa kelas VII dalam menyelesaikan soal AKM (Asesmen Kompetensi Minimum) kelas. *Jurnal THEOREMS (The Original Research of Mathematics), 7*(2), 117-124. https://doi.org/10.31949/th.v7i1.4109

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49-57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self-reported criminal behavior. *Sociological Methodology*, *33*(1), 169-211. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00130.x

Smith, A. B., Cocks, K., Parry, D., & Taylor, M. (2016). A differential item functioning analysis of the EQ-5D in cancer. *Value In Health, 19*(8), 1063-1067. https://doi.org/10.1016/j.jval.2016.06.005

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral*

Statistics, *41*(5), 481-520. https://doi.org/10.3102/1076998616646200

Suryadi, B., Hayat, B., & Putra, M. D. K. (2020). Evaluating psychometric properties of the Muslim Daily Religiosity Assessment Scale (MUDRAS) in Indonesian samples using the Rasch model. *Mental Health, Religion & Culture*, *23*(3-4), 331-346. https://doi.org/10.1080/13674676.2020.1795822

Syafitri, R. A., Anggraini, S. A., & Saragi, S. M. (2021). Analisis keefektifan penerapan aplikasi AKSI (Assesmen Kompetensi Siswa Indonesia) di SD Negeri 130002 Kec. Sei Tualang Raso, Kota Tanjung Balai. *Dirasatul Ibtidaiyah, 1*(2), 185-197.

Syahputra, E., Wiranda, A., Hamdany, S., & Pardamean, P. (2022). Analisis perbandingan uji kemampuan Bahasa Indonesia antara pria dan wanita. *Jurnal Multidisiplin Dehasen (MUDE)*, *1*(3), 251-258. https://doi.org/10.37676/mude.v1i3.2511

Temel, G.Y., Rietz, C., Machunsky, M., & Bedersdorfer, R. (2022). Examining and Improving the Gender and Language DIF in the VERA 8 Tests. *Psych*, *4*, 357-374. https://doi.org/10.3390/psych4030030

Thompson, M. (2005). Combining Rasch scaling and multi-level analysis. In R. Maclean (Ed.), *Applied Rasch measurement:A book of exemplars. Papers in honour of John P. Keeves* (pp. 197-206). Cham, Switzerland: Springer.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209-220. https://doi.org/10.1037/a0023353

Umar, J. (1987). *Robustness of the simple linking procedure in item banking using the Rasch model*. Doctoral dissertation, University of California, Los Angeles, United States of America.

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item Response Theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 11.1–11.25). Paris, France: IEA.

Wahyuningsih, S. (2018). Men and women differences in using language: A case study of students at STAIN Kudus. *EduLite: Journal of English Education, Literature and Culture*, *3*(1), 79-90. https://doi.org/10.30659/e.3.1.79-90

Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher*, *30*(6), 17-21. https://doi.org/10.3102/0013189X030006017

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (pp. 325-332). Berlin, Germany: Springer-Verlag. https://doi.org/10.1007/978-4-431-65955-6_35

Wright, B. D. (1968). *Sample-free test calibration and person measurement*. In Proceedings of the 1967 Invitational Conference of Testing Problems. Princeton, NJ: Educational Testing Service.

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

Wu, M. (2013). Using item response theory as a tool in educational measurement. In M. M. C. Mok (Ed.), *Self-directed learning oriented asessments in the Asia-Pacific* (pp. 157-186). Tokyo, Japan: Springer. https://doi.org/10.1007/978-94-007-4507-0_9

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Yu, C. H. (2020). Objective measurement: How Rasch modeling can simplify and enhance your assessment. In M. S. Khine (Ed.), *Rasch measurement: Applications in quantitative educational research* (pp. 47-76). Cham, Switzerland: Springer. https://doi.org/10.1007/978-981-15-1800-3_4

Yumithasari, R., Sunyono., & Munaris. (2022). Pengembangan instrumen asesmen portofolio untuk mengukur kemampuan berbahasa Indonesia tulis peserta didik kelas IV sekolah dasar. *Edukatif: Jurnal Ilmu Pendidikan, 4*(3), 4713-4720. https://doi.org/10.31004/edukatif.v4i3.2796

Zhang, M. (2016). *Exploring dimensionality of scores for mixed-format tests*. Doctoral dissertation, The University of Iowa, United States of America.

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, *21*, 245-266. https://doi.org/10.1007/s10833-019-09367-x