EVALUATING THE VALIDITY OF MEARA AND ROGERS' VOCABULARY APTITUDE TEST: A RASCH MEASUREMENT MODEL ANALYSIS

Reza NEJATI[1]*
Mohsen JABBARI[2]
[1]English Department, Faculty of Humanities, Shahid Rajaee Teacher Training University, Tehran, Iran
[2]Iran University of Science and Technology

[1]reza.nejati@sru.ac.ir
[2]mohsenjabbari97@yahoo.com

Manuscript received 17 January 2023
Manuscript accepted 2 December 2023
**Corresponding author*
https://doi.org/10.33736/ils.5290.2023

ABSTRACT

Language aptitude tests, such as the LLAMA_ B3, developed by Meara and Rogers (2022), may play an important role in vocabulary learning research. This paper reports the evaluation of the construct validity of Meara and Rogers' (2022) Vocabulary Aptitude Test. Data collected from 314 participants were analysed using Rasch analysis. All the eigenvalues fell below 2, supporting the assumption of unidimensionality. Q3 findings showed that the assumption of local independence was met. This study found a strong reliability evidence for the items and a convergent estimate of the results. Accordingly, the test results displayed validity for the present sample of students. As far as the analysis shows, the items represent a single underlying construct, meaning that the LLAMA_ B3 is fundamentally coherent. Psychometrically, a single trait seems to have been identified. It can be concluded that both the items and participants behaved predictably, indicating that the test is certainly worth further investigation and refinement.

Keywords: vocabulary aptitude test; construct validity; foreign language aptitude; Rasch analysis

Introduction

The intriguing and slippery notion of language aptitude has captured researchers' attention for many years. Several concepts are typically associated with foreign language aptitude, including talent, giftedness, language acquisition ability, or language acquisition expertise (Ameringer et al., 2018). There are various tests claiming to assess these concepts, which makes it necessary to determine the type of interpretation that will be used before selecting any of the tests. This necessity shifts the focus from the theoretical underpinnings of foreign language aptitude to the instruments used to measure this trait (Bokander & Bylund, 2020).

As emphasis is placed on different learning conditions, researchers start to shift their attention from viewing aptitude as the only determining factor of second language (L2) success to treating it as a dynamic construct (Robinson, 2005). As a result, giant steps have been taken to study what might influence language aptitude as a mental construct, with working memory being the most frequently studied (Doughty & Mackey, 2021; Huang et al., 2020; Wen, 2019). One glaring gap, however, is that these primary studies only reported correlations between aptitude and other variables as holistic constructs rather than investigating the construct validity of aptitude (Li, 2015).

Numerous test batteries, including the Modern Language Aptitude Test (MLAT; Carroll & Sapon, 1959), Pimsleur Aptitude Battery (Pimsleur, 1966), Defence Language Aptitude Battery (DLAB; Petersen & Al-Haik, 1976), Cognitive Ability for Novelty in Acquisition of Language-Foreign (CANAL-FT; Grigornko et al., 2000), Language Learning and Meaning Acquisition (LLAMA; Meara, 2005), and the High-Level Language Aptitude Battery (HiLAB; Linck et al., 2013) have been created to measure the construct of aptitude. MLAT is the most common test to quantify language aptitude (Sparks et al., 2005). However, the LLAMA, which is a freely available and language-neutral test, has also gained a lot of attention (Mikawa & De Jong, 2021). These two unique features set LLAMA apart from the other aptitude measures.

The LLAMA_ B3 tests the ability to learn words. Language learning is believed to be significantly affected by word learning (Harmon et al., 2009). Although this test is a widely accepted tool to test language aptitude, its cultural adaptivity and validity have not been fully investigated. As LLAMA _B3 is language neutral test, it has the benefits of equality for every language. Hence, evaluating the usefulness of this test battery will benefit language researchers. Even though some studies have attempted to assess the validity of language aptitude tests scores, no study so far has examined the construct validity of the LLAMA_B3 test scores within the Iranian EFL context.

This study examines whether the latest version of LLAMA_B3 is a reliable and useful vocabulary aptitude test for Iranian EFL students. The results can provide good insight into students' foreign language aptitude for vocabulary acquisition and serve as a valid starting point for interpreting LLAMA_B3 items. The following research questions are addressed.

1. To what extent are the results of the LLAMA _B3 valid for the present sample of participants?
2. How reliable are LLAMA _B3 scores for the items and the persons in the current study?

According to Messick (1985), validity refers to the "appropriateness, meaningfulness, and usefulness of the specific inferences made from the test scores" (p. 9), and "validation" is the "process of gathering evidence to support the specific inferences made from the test scores" (p. 9). The task of accumulating evidence to support all these inferences can be quite challenging to researchers, as it would not be possible to collect and analyse all types of data for all types of evidence that may be relevant. The present study thus only focuses on the assessment of LLAMA_B3 construct-related validity to determine if it fits the data provided by the test scores. The Rasch model is applied to further understand the data and to determine if the test items measure the same construct.

Review of the Literature

The development of foreign language aptitude (FLA) started in 1958 with John Bissell Carroll. Carroll and Sapon (2002) defined aptitude as "the rate at which people can learn an unknown language, and there are no definite differences between languages in terms of language learning abilities" (as cited in Stansfield & Reed, 2004, p. 54). Simply put, those who can learn a new language quickly have a certain degree of aptitude for learning a foreign language. Nevertheless, Carroll's assumption was not based on a broad theoretical framework (Smith & Stansfield, 2016).

Carroll and Sapon (1959) developed the Modern Language Aptitude Test (MLAT). Since then, language researchers utilising MLAT have modified the construct of language aptitude (O'Malley et al., 1993) although its main features emphasising the interdependency of external and personal factors in FLA are still in practice today (Reiterer, 2018).

A serious concern in specifying the characteristics of an individual to learn a foreign language is that such measurements are sometimes against the principle that learners deserve equal rights and opportunities in education (Skehan, 2016). Foreign language aptitude is sometimes used interchangeably with several other concepts, including talent, natural language-learning ability, or learning expertise (Ameringer et al., 2018). Therefore, testing experts often find it difficult to differentiate between aptitude and talent (Vinkhuyzen el al., 2009).

Aptitude tests seem to be a myth for many language researchers because the tests are inaccessible for research purposes. For instance, some protected tests are only administered to people who work for the US government (Robinson, 2002). Meara developed the Language Learning and Meaning Acquisition test (LLAMA) in 2005. A free version of this aptitude test is easily available (Meara & Rogers, 2019). This test battery is based on what was established by Carroll and Sapon (1959) and Carroll (1962), and measures language aptitude using four different subtests: LLAMA_B3 relates to learning vocabulary, LLAMA_D relates to listening for new words, LLAMA_E relates to sounds and symbols, and LLAMA_F relates to grammar rules (Meara & Rogers, 2019).

LLAMA, a computer-based test battery, is special in that it does not limit its usage and accessibility (Granena, 2013). Although it is not as distinctive as MLAT, LLAMA has been used by many researchers (Artieda & Mu˜noz, 2016; Bokander & Bylund, 2020). The LLAMA manual describes the test as language-neutral (Meara, 2005). Such a feature may have given it an advantage over other language aptitude tests, which depend on specific languages. A participant's native language may inappropriately influence the outcome of an aptitude measure (Granena, 2013).

Researchers in this field have provided some evidence to support the broad argument about the distinctiveness of language aptitude tests. Abrahamsson and Hyltenstam (2008) explained that adopting the LLAMA as a research instrument supports the previous research that personality, attitude, motivation and test difficulty and cognitive loads of the items influence the way students perceive the items. However, very few researchers questioned the validity of this measuring tool (e.g., Bokander & Bylund, 2019; Mikawa & De Jong, 2021; Sachs et al., 2019). In the same vein, there seems to be no valid argument supporting LLAMA_B3. While some grammarians may ignore the role of word learning in second language learning, most educators believe that language cannot be learned without words (Alqahtani, 2015).

Some linguists equate language learning with learning words as words are the most basic and meaningful elements of any language (Walters, 2004). Others, like Li (2015), viewed aptitude as a predictor of different aspects of learning vocabulary and writing.

The differences in the nature of aptitude testing have encouraged experts to agree that more research is required to offer insight into the instrument to measure language aptitude (Singleton, 2017). Due to the variety of constraints for aptitude, the LLAMA test presents the best option for analysis. To understand the investigational and low-stakes purposes of LLAMA_B3 (Meara, 2005), it is essential to conduct a validation study on results from the first subtest of this testing battery named LLAMA_B3: Learning new words.

Bokander and Bylund (2020) were the most recent researchers who assessed the validity of the LLAMA test results (Meara, 2021). Except for LLAMA_B3, they found that the LLAMA battery displayed flaws at all three levels of measurement, including single items, components, and the entire test. It should not be used when the outcome could have serious consequences for test takers (Bokander & Bylund, 2020). To ascertain this position, we need to examine the potential of the LLAMA_B3 in a logical validation process.

The latest research does not provide enough support for all subtests of the LLAMA test battery; for example, Bokander and Bylund (2020) found that the trial response times for LLAMA_D were shorter than other subtests indicating that LLAMA_D was a different subtest from the rest. Further, LLAMA_E can test an individual's ability to connect familiar sounds with an unfamiliar writing system. The logic behind this test is that many students find it difficult to accept that letters do not always represent sounds used in their mother tongue.

Fundamental studies of linguistics and psychology have questioned the validity of data collected using aptitude tests to be interpreted in different teaching or research contexts (Stansfield & Reed, 2004). Similarly, they sought to inform the decision makers to select individuals with better performance in numerous fields. From a developmental viewpoint, it can be observed that second language learning aptitude is determined by heredity and the first language is not as influential as heredity (Dale et al., 2010).

However, the theoretical issues which fall within the scope of language aptitude instruments are far from solved. For example, Planchon and Ellis (2012) argued that on the aptitude test (DLAB), bilinguals outperformed monolinguals. Similarly, Sáfár and Kormos (2008), adopting Hungarian Language Aptitude Test, found that language aptitude has limited power in predicting language learning success. Dörnyei and Skehan (2003) argued that individual learners vary in their natural talent to learn foreign languages. Language aptitude is complicatedly affected by many other factors in the educational setting (Birdsong, 2018). That said, a test like LLAMA does not claim to measure knowledge or intelligence; it only measures untrained perceptual abilities. As a result, preparation before taking the test is pointless (Kagan, 2022).

Method

Participants

The study sample comprised 343 male and female Iranian university students (aged 18 to 22). All participants were native speakers of Persian with similar cultural backgrounds. By means of convenient sampling, some high school diploma holders who were doing their BSc were also employed for the study. Twenty-nine participants were later excluded as they did not complete the online test in university learning management system. The responses of the remaining 314 students were analysed.

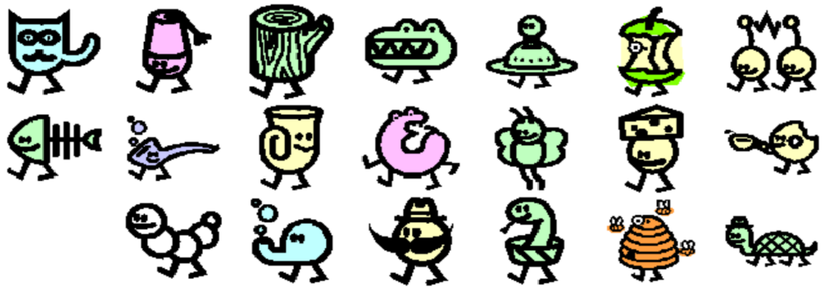
Instruments

For the purpose of the study, the LLAMA_B3, designed by Meara and Rogers, was used.
Developed in the form of a computer programme, the LLAMA _B3 tests the ability to learn the names of unusual objects. The programme shows the participants a series of 20 unknown objects and asks them to learn the names of the objects. The participants have two minutes to complete the task. The program then tests them by showing the objects one by one and asking them to determine the correct name from a list of 20 names. After

two minutes, a new screen appears on a new page. This screen includes the same 20 items that the participants have studied so far but are arranged differently. The participants then need to follow the directions in the bar at the bottom of the page. For example, if the instruction tells them to *Click on the taa*, they must click on the object with that name. If they fail to find the item they are looking for, they are advised to guess it by randomly clicking on an item.

Figure 1

The Set of the Pictures Used in LLAMA_B3



Procedure

As it was the first time that the participants took a language aptitude test, prior to the exposure, they were provided with a Persian translation of the LLAMA_B3 manual and some instructions on how it works. In case of ambiguity, the test takers were allowed to ask for clarification. In addition, they were taught how to enter personal ID codes that would be used to identify them to the computer application.

Out of the 314 responses, 17 respondents experienced online glitches. Each respondent was contacted via email or Telegram and given time to retake the test. Their first attempt to answer the questions was not recorded. Paul Meara was contacted for the raw data and the keystroke for coding the elicited responses via email. All data were graciously provided by Meara for academic research purposes. The entire dataset was collected and codified within 10 months, from October 2021 to July 2022.

Data Analysis

Rasch analysis was used to assess the construct validity of the LLAMA_B3 test results. The researchers tested the assumptions of unidimensionality and local independence. The principle of unidimensionality requires that each human attribute be measured separately (Bond & Fox, 2015). To find out if the LLAMA_B3 was one-dimensional, following Aryadoust et al. (2020), a principal component analysis of residuals (PCAR) was done.

Local independence, similar to detecting multicollinearity in regression, can be examined using several methods, including the G2 and V2 statistics (Chen & Thissen, 1997), the Cramer's V statistic (Baldonado et al., 2015) and Q3 coefficients (Fan & Bond, 2019) which indicate the correlation between the Rasch residuals of two test items. We used Q3 to test the assumption of local independence. If the Q3 coefficient is less than .3, the degree of local independence is satisfactory. When the value is high, the two items measure the primary construct and another construct (Aryadoust et al., 2020).

To provide more evidence of the validity of test scores, CONSTRUCT MAP 4 (Wilson, 2011) was used to analyse variance-covariance structures and item fit statistics. The fit statistics are presented in weighted mean square terms (infit) and unweighted mean square terms (outfit). The mean square (MnSq) index, which is expected to be 1.00, can be used to figure out how unusual the data set is. In MnSq metrics, for example, a value of 1.2 indicates that there is 20 percent noise in the data, while a value of 1.1 indicates less distortion; for standardized (*t*) metrics, a range between 1.96 and -1.96 is recommended (Linacre, 2002).

Besides Rasch analysis indexes (fit and difficulty), some researchers reported item difficulty and discrimination (2-parameter logistic model), difficulty, discrimination and low ability respondent behaviour (3-parameter model), difficulty, discrimination, low ability behaviour, and high ability behaviour (4- parameter model). As such, a four-parameter model (4 PM) provides information about the behaviour of high achievers, along with information from 1 PM, 2 PM, and 3 PM.

The JMETRIK software was used to analyse the 4-PM Rasch. This software estimates parameters using joint maximum likelihood (JML). Software packages that perform Rasch analyses generally use one of three estimation methods: conditional maximum likelihood estimation (CMLE), joint maximum likelihood estimation (JMLE), or marginal maximum likelihood estimation (MMLE). According to Nicklin and Vitta (2022), these methods produce similar results. Nevertheless, the JMETRIK manual (Meyer, 2014) does not provide hard-and-fast instructions on how discrimination indices should be interpreted. Instead, it maintains that when discrimination indexes are high, it is more probable that high-scoring examinees will get the item correct, while low-scoring examinees will tend to miss it.

As an extension of Rasch analysis, Wright maps facilitate the analysis of test items and participants' abilities. Several Rasch analysis software packages, including CONSTRUCT MAP 4, which was used in this study, provide a map of the person-item distribution, also known as the Wright map. The Wright Map indicates the level of readiness of the respondent by assessing the difficulty of the task and the respondent's ability (Linacre, 2002; Wilson, 2011), allowing a better understanding of how prepared the respondents are.

To examine the quality of the test scores, the separation index for items and persons was used. Further, this index measures the number of levels of item difficulty or person ability in the data (Linacre, 2019). A test that shows high separation (>2) can distinguish between items/people of different difficulty/ability levels.

Results and Discussion

The mean of the LLAMA_B3 scores of the participants was 8.5 (SD = 5.19), with scores ranging from 1 to 19 (maximum score of 20). According to Meara (2005), scores between 25% and 45% are considered average. Table 1 shows that the average score obtained in this study was 42%. It is reasonable to assume that the distribution of scores is fairly normal since the skewness index is less than 0.5 and the kurtosis value is less than 1 (Pallant, 2011).

Table 1
Descriptive Statistics

N	Mean	Std. Deviation	Minimum	Maximum	Skewness	Kurtosis
314	8.50	5.19	1.00	19.00	.46	-.81

As mentioned earlier, Rasch analysis was used to assess the construct validity of the LLAMA_B3 test results. The principal component analysis of residuals (PCAR) results helped establish the assumption of test unidimensionality. The associations among the item responses can be explained by a single underlying latent variable, which represents the target construct that is being measured (Bond et al., 2021). The key findings of the PCAR are presented in Table 2.

Table 2
Principal Component Analysis of Residuals

	Function 1	Function 2	Function3	Function4	Function5
Eigen value	1.85	1.51	1.45	1.30	1.26
Proportion Variance	0.09	0.08	0.07	0.07	0.06
Proportion Explained	0.25	0.20	0.20	0.18	0.17

Following Linacre (2006), if the first value of the correlation matrix of the residuals is less than 2.00, the residuals are treated as random noise. However, if the eigenvalue exceeds 2.00, there may be a second dimension besides the primary Rasch dimension. Table 2 indicates that the eigenvalues of the subfunctions were all below 2.00. Thus, the assumption of unidimensionality was considered met. Hambleton et al. (1991, p. 9) stated that the assumption of unidimensionality cannot be fully met, since "... several cognitive, personality, and test-taking factors always affect test performance, at least to some extent." Additionally, unidimensionality does not remain the same across different samples of participants (Linacre, 2019).

For the assumption of local independence, we used Q3 index. The highest correlation was between items 5 and 16 (-0.25). Hence, local independence was met, meaning that the items in the test are independent of each other and effectively measure only language aptitude. Glas (2016) suggests that tests that focus on item local independence are excellent indicators of unidimensionality. Such tests are useful in identifying potential problems with the scale or items.

For the fit indexes, CONSTRUCT MAP 4 was used to analyse the responses of the students. A convergent estimate of the results was obtained (Variance-Covariance Matrix: +2.002 and -2 log likelihood=+11.458.300). Following Wilson (2011), we therefore posit that the LLAMA_B3 is valid for the sample of students examined in this study.

Table 3
Scale Quality Statistics

Statistic	Items	Persons
Separation Index	5.65	2.29
Reliability	0.96	0.84

Table 3 provides details on the quality of the LLAMA_B3. There was a reliability of .96 for the items and .84 reliability for the persons in the current study. According to Duncan et al. (2003), reliability between .70 and .79 is considered acceptable, between .80 and .89 is considered good, and between .90 and .99 is considered excellent. The separation index for items and persons were 5.65 and 2.29 respectively, meaning the quality of the test was satisfactory. As such, the LLAMA_B3 can distinguish between items/people of different difficulty/ability levels.

Table 4
Fit Statistics for Test Items

Item* step	Outfit Unweighted MnSq	*t*	Infit weighted MnSq	*t*
Item 1	1.32	3.6	1.19	2.5
Item 2	1.11	1.3	1.05	0.7
Item 3	1.23	2.7	1.21	2.8
Item 4	1.03	0.3	1.07	1.0
Item 5	1.17	2.0	1.16	2.1
Item 6	1.16	1.9	1.18	2.4
Item 7	1.10	1.2	1.12	1.6
Item 8	1.05	0.6	1.01	0.2
Item 9	1.04	0.5	0.95	-0.6

Item 10	1.14	1.7	1.09	1.2
Item 11	0.99	-0.1	0.98	-0.3
Item 12	1.16	1.9	1.18	2.3
Item 13	1.06	0.8	1.03	0.4
Item 14	0.97	-0.3	0.92	-1.0
Item 15	1.03	0.5	0.88	-1.5
Item 16	0.99	-0.0	0.95	-0.6
Item 17	1.00	0.1	0.94	-0.7
Item 18	1.06	0.7	0.98	-0.2
Item 19	0.86	-1.9	0.80	-2.7
Item 20	1.14	2.01	1.40	2.80
Average	1.08	0.9	1.04	0.5

Table 4 illustrates that the largest infit index belonged to item 3 (1.21), while the smallest value belonged to item 19 (0.8); the remaining items were between 0.8 and 1.21, and the average infit was reported as 1.04. According to Wright at el. (1994), the mean square range of Infit and Outfit is 0.8-1.2 for high stakes and 0.7-1.3 for ordinary tests. Thus, almost all of the items in the LLAMA_B3 fit the Rasch model. However, the t values of infit indexes for items 1,3, 5,6,12,19, and 20 were outside the acceptable range of 1.96 and -1.96 (Yan et al., 2020), indicating that these items were less compatible with the model. The problem of misfit items can be seen from two perspectives: One, the item may discriminate poorly, or two, it may work well but not fit the measurement trait defined by other items (McNamara, 1996).

In this study, we examined a four-parameter model allowing the upper asymptote of each item to be fewer than 1 (Linacre, 2004), taking into account the possibility that even a high ability respondent might occasionally answer an easy question incorrectly. In Table 5, we present the results of this analysis.

Table 5

4-Parameter Estimates and Standard Errors

Item	A (SE)	B (SE)	C (SE)	U (SE)
	discrimination	*difficulty*	*chance(guess)*	*upper asymptote (high ability learner's behaviour)*
Item1	2.02 (0.47)	-0.65 (0.20)	0.20 (0.07)	0.83 (0.03)
Item2	2.45 (0.29)	-0.76 (0.12)	0.13 (0.05)	0.93 (0.02)
Item3	2.00 (0.42)	-1.03 (0.19)	0.18 (0.08)	0.87 (0.03)
Item4	2.27 (0.32)	-0.38 (0.11)	0.11 (0.05)	0.95 (0.02)
Item5	1.85 (0.35)	0.05 (0.15)	0.12 (0.05)	0.91 (0.04)
Item6	1.53 (0.29)	0.09 (0.19)	0.13 (0.05)	0.93 (0.04)
Item7	1.91 (0.38)	-0.06 (0.16)	0.20 (0.06)	0.95 (0.03)
Item8	2.16 (0.35)	0.45 (0.12)	0.12 (0.04)	0.95 (0.03)

Item9	2.46 (0.28)	0.76 (0.10)	0.06 (0.02)	0.95 (0.03)
Item10	1.67 (0.35)	0.31 (0.17)	0.11 (0.04)	0.90 (0.05)
Item11	2.11 (0.34)	0.51 (0.13)	0.14 (0.04)	0.96 (0.03)
Item12	2.05 (0.35)	0.26 (0.13)	0.09 (0.04)	0.92 (0.04)
Item13	1.94 (0.34)	0.64 (0.17)	0.12 (0.04)	0.70 (NaN)
Item14	2.52 (0.25)	0.82 (0.10)	0.04 (0.02)	0.95 (0.03)
Item15	2.29 (0.33)	0.85 (0.14)	0.10 (0.03)	0.70 (NaN)
Item16	1.54 (0.32)	0.96 (0.25)	0.16 (0.04)	0.70 (NaN)
Item17	2.12 (0.35)	1.00 (0.16)	0.07 (0.03)	0.70 (NaN)
Item18	1.66 (0.39)	0.58 (0.26)	0.27 (0.05)	0.70 (NaN)
Item19	2.39 (0.34)	0.87 (0.10)	0.10 (0.02)	1.00 (0.00)
Item20	2.11 (0.41)	0.97 (0.16)	0.21 (0.04)	0.95 (0.03)

The second column of Table 5 illustrates the estimate of the item discrimination parameter. Item 14 had the highest value of 2.52 and item 6 had the lowest value of 1.53. The interpretation of discrimination indices is not prescriptive. It could be said that when discrimination indexes are high, high-scoring examinees are more likely to get the item right, while low-scoring examinees are more likely to miss it. It should be emphasised that discrimination, like other test characteristics, needs to be viewed in light of its purpose. For example, if LLAMA_B3 is to be used for selection purposes, it would be advisable to select highly discriminatory items.

The third column shows the estimate of the item difficulty parameter. The Rasch analyses calculate item difficulty as *Z* scores, which range from -3 (easiest) to 3 (most difficult). Based on this, from the present data set, item 3, with a *Z* score of -1.03, was the easiest item, and item 17, with a *Z* score of 1.00, was the most difficult. As with the previous index, this index needs to be weighed following the test's purpose. When using LLAMA_B3 for placement purposes, choosing items that consider the examinee's location is advisable. It is recommended that the items are presented in order of ease, that is, from the easiest to the most difficult. Further explanation is provided in the Wright Map presented in Figure 2.

The fourth column displays the estimation of the lower asymptote parameter (pseudo-guess). Our data set showed that item 18 had the highest index (0.27). IRT literature (Rulison & Loken, 2009) suggested that C indices over .4 should be considered problematic. Guessing occurs especially when students feel incapable of solving the question using their existing abilities.

The last column shows the estimate of the upper asymptote parameter (careless errors), denoted as U. It is evident that the lowest index was .7 (items 13,15,16,17,18), which is significantly higher than the 0.00 value. Loken and Rulison (2010) used the upper asymptote of 1 to calculate the probability of a high-ability student failing to answer an easy item correctly. Careless errors can produce more serious estimation biases than guesses, particularly when these errors occur early in a test. When students are anxious, careless, unfamiliar with computer techniques, distracted by poor test conditions, or misinterpret the question, gifted students can sometimes miss items that they should have answered correctly (Rulison & Loken, 2009).

As mentioned in the data analysis section, CONSTRUCT MAP 4 was used in this study to provide a map of the person-item distribution known as the Wright map. Using the Wright Map, we determined how prepared the respondents were for the task at hand (Figure 2).

Figure 2
Map of Person and Response Model Estimates

```
===============================================================
Z   raw    N of students     Items
----------------------------------------------------------------
3  |                 |
   |                |
   |                |
   |                |
   |                |
   |19         XXXXXXXX|
   |                |
   |                |
2  |18              XX|
   |17              X|
   |                |
   |17           XXXXXX|
   |                |
   |16          XXXXXXX|
   |                | Item 17.1
   |15             XXX| Item 15.1
1  |                 | Item 14.1
   |14           XXXXX| Item 16.1
   |                |Item 9.1 Item 13.1 Item 19.1
   |13          XXXXXX|
   |12         XXXXXXXX|
   |                |
   |11       XXXXXXXXXX| Item 20.1
0  |                 | Item 8.1 Item 10.1 Item 11.1 Item 18.1
   |10          XXXXXXX| Item 12.1
   |9          XXXXXXX|
   |8               X| Item 5.1
   |8  -------XXXXXXXXXXXXX| Item 6.1
   |                |
   |7      XXXXXXXXXXXXX|
   |                 | Item 7.1
-1 |6       XXXXXXXXX| Item 1.1 Item 4.1
   |                 |
   |5  XXXXXXXXXXXXXX|
   |                 |
   |4     XXXXXXXXXXX| Item 2.1 Item 3.1
   |                 |
   |                 |
   |3      XXXXXXXXXX|
-2 |                 |
   |2     XXXXXXXXXXX|
   |                 |
   |                 |
   |                 |
   |1      XXXXXXXXXX|
   |                 |
   |                 |
-3 |                 |
================================================================
Each X represents 2 students, each row is 0.128 logits
```

Figure 2 illustrates the Wright map showing respondents' performance on LLAMA_B3. In the two vertical broken line, the left side represents the students, whereas the right side represents the items. On the left side of the map, participants' abilities are arranged from the most to the least able in *Z* scores and Raw scores, starting from the top. On the right side of the map, the most difficult items are at the top and the easiest items are at the bottom. At the point where the mean item is 0.00 logit, we have items 8, 10, 11 and 18, followed by easier items, such as items 12, 5, 6, 7, 1, 4, 2 and 3. Items 20, 9, 13, 19, 16, 14, 15, and 17 are located on the top side of the graph, indicating a higher degree of difficulty.

In the sample of 48 respondents, 15% who scored 16+ (maximum 20), could not be located by any specific test item, meaning no specific test item could be used to evaluate their ability level. Throughout Figure 2, empty spaces are used as a way of demonstrating this point. This can be taken to mean that the test designer needs to develop some items that will be appropriate for learners of higher abilities. It is possible that some individual factors have affected the performance of the test takers. Our findings also showed that 60 students, almost 20% of the sample, scored below the easiest items, that is, items 2 and 3. In other words, the students found the test to be too challenging for them. Since there is a difficulty gap between -1.5 logit and -1.00 logit, 28 students (nearly 9% of the sample) were left without appropriate items. This difficulty gap is displayed from the fact that no students and items (no x in front of the logits as shown in figure 2) are displayed for these logits. There is also a difficulty gap between 0.6 and 0.7 logits, where 26 students (8%) were not provided with specific items.

Similarly, 26 respondents or 8% of the sample, who scored at 0.5 and 0.6 logits, were not given appropriate items at this level of ability. Overall, 188 students (nearly 60%) did not find items that matched their abilities. It may be due to the fact that the test items, strange figures along with meaningless words, made up a decontextualised test.

It should be noted that there was an overlap of difficulty between the following pairs or sets of items: items 2 & 3; 1 & 4; 8, 10, 11 & 18; 9, 13 & 19; as well as all items with item 1. Accordingly, each set's first or last item performs the same function as the other. This means that some of these items can be omitted from the test. However, it should also be noted that the modifications proposed here are not intended to reject the construct validity of the test; rather, they are intended to inform test users of some potential limitations.

Conclusion

Testers and users must understand what the results of the items and people's performance tell them about the theory they are testing and what the theory tells them about the people and items they are testing. The analysis thus far suggests that most of the items fit the model and should represent a single underlying capability. The findings in this paper support the argument that LLAMA_B3 is fundamentally coherent and valid for the sample of students that was studied. Research that focuses on other populations would be able to determine whether the same findings could be replicated. Additionally, it would be beneficial to analyse the predictive validity of LLAMA_B3 to gain better understanding of the usefulness of the instrument.

Based on the findings of this study, we conclude that scores on these items are highly reliable, indicating that the items measure the same underlying concept. The current study, nonetheless, did not include any evidence of criterion-related validity. For this, further research is needed to establish the predictive power of LLAMA_B3 in vocabulary learning. The present study also did not provide evidence for differential validity. It may be helpful to design a factorial study to compare the performance of male and female learners and try the test with other age groups and learners of foreign languages other than English as evidence of differential validity. This would allow researchers to determine differences between the results of the two genders, or if the results are consistent regardless of the language or age group. It would also help to ascertain if there is any bias in the test results.

Acknowledgement

The research was supported by a grant from Shahid Rajaee Teacher Training University, with an agreement number of 4933 and a date of May 27, 2023.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*, 481-509. https://doi.org/10.1017/S027226310808073X

Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International Journal of Teaching and Education*, *III* (3), 21-34. https://doi.org/10.20472/te.2015.3.3.002.

Ameringer, V., Green, L., Leisser, D., & Turker, S. (2018). Introduction: Towards an interdisciplinary understanding of language aptitude. In S. M. Reitererm (Eds.), *Exploring language aptitude: Views from psychology, the language sciences, and cognitive neuroscience* (pp. 19-42). English Language Education, vol 16. Springer. https://doi.org/10.1007/978-3-319-91917-1_1

Artieda, G., & Mu˜noz, C. (2016). The LLAMA tests and the underlying structure of language aptitude at two levels of foreign language proficiency. *Learning and Individual Differences, 50*, 42-48. https://doi.org/10.1016/j.lindif.2016.06.023

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Baldonado, A. A., Svetina, D., & Gorin, J. (2015). Using necessary information to identify item dependence in Passage-Based Reading Comprehension Tests*. Applied Measurement in Education, 28*(3), 202-218. https://doi.org/10.1080/08957347.2015.1042154

Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology, 9*, 81-98.

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA Language Aptitude Tests. *Language Learning, 70*(1), 11-47. https://doi.org/10.1111/lang.12368

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). L. Erlbaum.

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test manual*. Psychological Corporation.

Carroll, J. B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87-136). University of Pittsburgh Press.

Carroll, J. B., & Sapon, S. (2002). *Modern Language Aptitude Test: Manual*. Second Language Testing.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioural Statistics, 22*(3), 265-289. https://doi.org/10.2307/1165285

Dale, P. S., Harlaar, N., Haworth, C., & Plomin, R. (2010). Two by two: A twin study of second language acquisition. *Psychological Science, 21*(5), 635-640. https://doi.org/10.1177/0956797610368060

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (p. 589-630). Oxford, UK: Blackwell.

Doughty, C. J., & Mackey, A. (2021). Language aptitude: Multiple perspectives. *Annual Review of Applied Linguistics, 41*, 1-5. https://doi.org/10.1017/s0267190521000076

Duncan, P. W., Bode, R. K., Min Lai, S., Perera, S., & Glycine (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*(7), 950-963. https://doi.org/10.1016/s0003-9993(03)00035-2

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment. *Quantitative Data Analysis for Language Assessment I*, 83-102 https://doi.org/10.4324/9781315187815-5

Glas, C. A. (2016). Frequentist model-fit tests. *Handbook of item response theory, volume two: statistical tools*, 313-361. https://doi.org/10.1201/b19166-22

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *Modern Language Journal, 84*(3), 390-405. https://doi.org/10.1111/0026-7902.00076

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* SAGE Publications

Harmon, J. M., Wood, K. D., & Kiser, K. (2009). Promoting vocabulary learning with the interactive word wall. *Middle School Journal, 40*(3), 58-63. https://doi.org/10.1080/00940771.2009.11495588

Huang, T., Loerts, H., & Steinkrauss, R. (2020). The impact of second- and third-language learning on language aptitude and working memory. *International Journal of Bilingual Education and Bilingualism, 25*(2), 522-538. https://doi.org/10.1080/13670050.2019.1703894

Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test. *Language Learning & Language Teaching*, 35, 105-130. https://doi.org/10.1075/lllt.35.04gra

Kagan, J. (2022). Aptitude Test: Definition, how it's used, types, and how to pass. *Investopedia*. https://www.investopedia.com/terms/a/aptitude-test.asp

Li, S. (2015). The construct validity of language aptitude. *Studies in Second Language Acquisition, 38*(4), 801-842. https://doi.org/10.1017/s027226311500042x

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*(2), 878. https://www.rasch.org/rmt/rmt162.pdf

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95-110.

Linacre J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7*(1), 129-139.

Linacre, J. M. (2015). What do infit and outfit, mean-square and standardized mean? Rasch Measurement, *Trans 2002*, *16*, 878-882.

Linacre, J. M. (2019). *A user's guide to WINSTEPS® Rasch-Model Computer Programs*: Program Manual 4.4. 6. Mesa-Press, Chicago, IL. https://www.winsteps.com/winman/copyright.htm

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high level language proficiency. *Language Learning, 63*, 530-566. https://doi.org/10.1111/lang.12011

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology, 63*(3), 509-525. https://doi.org/10.1348/000711009x474502

McNamara, T. F. (1996). *Measuring second language performance.* Longman.

Meara, P. (2005). *LLAMA Language aptitude tests*. Lognostics.

Meara, P. M., & Rogers, V. E. (2019). *The LLAMA tests*. Lognostics. https://www.lognostics.co.uk/tools/LLAMA_3/index.htm

Messick, S. (1985). *Standards for educational and psychological testing*. American Psychological Association.

Meyer, P. J. (2014). *Applied measurement with jMetrik* (1st ed.). Routledge.

Mikawa, M., & De Jong, N. H. (2021). Language neutrality of the LLAMA test explored: The case of agglutinative languages and multiple writing systems. *Journal of the European Second Language Association*, *5*(1), 87-100. https://doi.org/10.22599/jesla.71

Nicklin, C., & Vitta, J. P. (2022). Assessing Rasch measurement estimation methods across R packages with yes/no vocabulary test data. *Language Testing*, *39*(4), 513-540. https://doi.org/10.1177/02655322211066822

O'Malley, J. M., Parry, T. S., & Stansfield, C. W. (1993). Language Aptitude Reconsidered. *The Modern Language Journal, 77*(2), 1-24. https://doi.org/10.2307/328950

Pallant, J. F. (2011). *SPSS survival manual: A step by step guide to data analysis using the SPSS program*. Allen & Unwin.

Pimsleur, P. (1966). The Pimsleur language aptitude battery. Harcourt, Brace, Jovanovic.

Planchon, A., & Ellis, E. (2012). A diplomatic advantage? The effects of bilingualism and formal language training on language aptitude amongst Australian diplomatic officers. *Language Awareness, 23*(3), 203-219. https://doi.org/10.1080/09658416.2012.742907

Reiterer, S. M. (2018). Exploring language aptitude: Views from psychology, the language sciences, and cognitive neuroscience. Springer-Nature.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude, and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfield and Hernstad (1991). In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211-265). John Benjamins Publishing.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25*, 46-73. https://doi.org/10.1017/s0267190505000036

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83-101. https://doi.org/10.1177/0146621608324023

Sachs, R., Akiyama, Y., & Nakatsukasa, K. (2019). The value of introspective measures in aptitude-treatment interaction research. *Journal of Second Language Studies, 2*(2), 336-364. https://doi.org/10.1075/jsls.19001.sac

Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *IRAL - International Review of Applied Linguistics in Language Teaching, 46*, 113-136. https://doi.org/10.1515/iral.2008.005

Singleton, D. (2017). Language aptitude: Desirable trait or acquirable attribute? *Studies in Second Language Learning and Teaching, 7*(1), 89-103. https://doi.org/10.14746/ssllt.2017.7.1.5

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in L2 processing and acquisition* (pp. 105-130). John Benjamins.

Sparks, R. L., Javorsky, J., & Ganschow, L. (2005). Should the modern language aptitude test be used to determine course substitutions for and waivers of the foreign language requirement? *Foreign Language Annals*, *38*(2), 201-210. https://doi.org/10.1111/j.1944-9720.2005.tb02485.x

Stansfield, C. W., & Reed, D. J. (2004). The story behind the modern language aptitude test: An interview with John B. Carroll (1916–2003). *Language Assessment Quarterly: An International Journal, 1*(1), 43-56. https://doi.org/10.1207/s15434311laq0101_4

The LLAMA tests. (2020). *The LLAMA Tests*. https://www.lognostics.co.uk/tools/LLAMA_3/index.htm

Vinkhuyzen, A. a. E., Van Der Sluis, S., Posthuma, D., & Boomsma, D. I. (2009). The heritability of aptitude and exceptional talent across different domains in adolescents and young adults. *Behaviour Genetics*, *39*(4), 380-392. https://doi.org/10.1007/s10519-009-9260-5

Walters, J. (2004). Teaching the use of context to infer meaning: a longitudinal survey of L1 and L2 vocabulary research. *Language Teaching, 37*(4), 243-252. https://doi.org/10.1017/s0261444805002491

Wen, Z. E. (2019). Working Memory as Language Aptitude [Review of *Working Memory as Language Aptitude*]. In Z. E. Wen (Ed.), *The Phonological/Executive Model* (pp. 187-214). Routledge. https://doi.org/10.4324/9781315122021-10

Wilson, M. (2011). Some notes on the term: "Wright Map". *Rasch Measurement Transactions*, *25*(3), 1331.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Löf, P. (1994). *Reasonable mean-square fit values. Rasch Measurement Transactions, 8*(3), 370. https://www.sciepub.com/reference/117986

Yan, Z., Heene, M., & Bond, T. (2020). *Applying the Rasch model* (4th ed.). Routledge.