On the Construct Validity of the Iranian Ministry of Health Language Exam (MHLE)

Research Article

S. Susan Marandi[1]
Leila Tajik*[2]
Leila Zohali[3]

Received: 2019-09-02 | Revised (1): 2020-02-28 | Accepted: 2020-03-16

Abstract

Considering validity as a unitary concept, this study investigated the construct validity of the Iranian Ministry of Health Language Exam (MHLE). To meet this objective, we first conducted item analysis and reliability analysis, and verified KR20-if-item-deleted indices on the scores of 987 MHLE test takers before running factor analysis. Though the test was found to enjoy a high level of reliability, it suffered from 28 problematic items flagged through item analysis and KR20-if-item-deleted indices. Next, we ran factor analysis on the data, screened through item analysis, by implementing Horn's parallel analysis and Velicer's minimum average partial (MAP) tests. Parallel analysis resulted in overfactoring. The MAP

[1] Associate Professor, Department of English Language and Literature, Faculty of Literature, Alzahra University, Tehran, Iran; susanmarandi@alzahra.ac.ir
[2] Assistant Professor, Department of English Language and Literature, Faculty of Literature, Alzahra University, Tehran, Iran (corresponding author); tajik_l@alzahra.ac.ir
[3] M.A. graduate, Department of English Language and Literature, Faculty of Literature, Alzahra University, Tehran, Iran; lzohali69@gmail.com

DOI: **10.22051/lghor.2020.28036.1180**

test, however, produced results with two to seven factors. Though the 4-factor result of the MAP test seemed to be more logical at first glance, the overall results were rather disappointing. Nineteen items did not load significantly on any factor and a clear pattern of item loading was not found for many items. These findings can be viewed as evidence detracting from the validity of MHLE.

Keywords: language assessment, language testing, proficiency test, item analysis, reliability analysis, factor analysis, MHLE.

Introduction

Early validity theory emerged during the 1930s and 1940s (Fulcher & Davidson, 2007) when the American Psychological Association (APA) recognized the necessity of preparing codes of ethics for testing. Efforts to codify validity standards resulted in introducing four approaches to validation, i.e., content, predictive, concurrent, and construct validity in 1954 (Stapleton, 1997). Later, in 1966, predictive and concurrent validity were reduced to a single category, namely criterion-related validity (Anastasi, 1986; Crocker & Algina, 1986; Stapleton, 1997). Soon, psychometricians felt dissatisfied with treating different types of validity as distinct pieces of evidence for supporting score interpretations (Messick, 1980). It was then that validity was considered a single, unitary concept (Bachman, 1990). In this new orientation, validity is regarded within a unified framework with content and criterion-related evidence in support of the construct validity in testing applications (Messick, 1989). Nowadays, the term construct validity is employed as an umbrella term embracing various types of evidence in favor of validity (Anastasi, 1986; Shepard, 1993) and is defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, p. 13).

The significance of determining validity of test interpretations, especially for high-stakes tests, is evident to test developers and practitioners (Brown, 2005). High-stakes tests are defined as tests whose scores have a significant influence on learners' life options and opportunities (Moses & Nanna, 2007; Spolsky, 1995). Roever (2001) considers admission tests for professional programs, like universities, citizenship tests, and certification exams as high-stakes test assessment situations. An example of high-stakes tests is university entrance examinations which might be administered in different countries. In cases when the test is administered annually on a nationwide basis, the stakes are very high. In Iran, the National Organization for Educational Testing (NOET) organizes a few nationwide high-stakes tests for students wishing to enter bachelor's, master's, and doctoral university degree programs each year. In addition to these examinations, a small number of high-stakes language proficiency tests are administered by the NOET, the Ministry of Health, Treatment, and Medical Education, and a few state-run universities under the supervision of the Ministry of Science, Research, and Technology to those interested in pursuing their higher education at doctoral level.

The Evaluation Center of the Ministry of Health, Treatment, and Medical Education is one of the organizations which develops and administers language proficiency tests to master graduates of medical sciences wishing to pursue their studies at the doctoral (MD) level. These tests are among the prerequisites to taking part in MD entrance examination in fields related to medical sciences. The tests are considered to be of high-stakes nature due to the great impact they have on the future of a considerable number of master-graduate test takers from all over the country. Notwithstanding the significance of the decisions made on the basis of the Ministry of Health Language Exam (MHLE), no valid documents, to our knowledge, has been reported on the test effectiveness as well as the crucial characteristics of the test, namely the reliability and validity of the test interpretations. To address these gaps, we conducted the present study to assess item analysis of the test and to perform an in-depth examination of the reliability and construct validity of the test uses in light of appropriate statistical computational methods. Results of the analysis have the potential to assist the test construction team in revising poor items, if any. Furthermore, the study can give insights to ESP practitioners and students on the constructs which underlie the test and, accordingly, help them to prepare their students and themselves for the exam.

Review of Literature

This section divides the related literature into two sections with theoretical and practical orientations. The initial section provides information on the history of validity, elucidating various orientations towards the concept in addition to defining the concept of validity and detailing various approaches utilized for assessing the construct validity of a test. The second section reviews empirical studies investigating construct validity of various language proficiency tests.

Theoretical Underpinnings

To begin with, one needs to know the history of validity. According to Brown (2010), in the traditional view of validity, it is divided into three sub-types: content, criterion-related, and construct; however, more recently, it has been rethought as a unitary factor known as construct validity. Related literature has yielded a variety of interconnected definitions for the term construct. Anastasi (1986), for instance, defines constructs as "theoretical concepts of varying degrees of abstraction and generalizability which facilitate the understanding of empirical data" (p. 5). Bachman (1990) views constructs as "definitions of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behavior" (p.255). Finally, Fulcher and Davidson (2007) maintain that for a general term to be considered a construct, it must have two features: first, it should be defined in a way so that it is measureable; second, it should be defined in such a way that it can have relationship with other constructs that are different.

As stated above, in more recent years, researchers in the field of measurement have reconceptualized validity as a unitary factor named as construct validity. As an example, Messick (1980) maintained that "construct validity is the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships" (p. 1015). Content validity as one type of evidence for test validity has been defined as "any attempt to show that the content of the test is a representative sample from the domain that is to be tested" (Fulcher & Davidson, 2007, p. 6). It involves two important concepts of content relevance and content coverage (Bachman, 1990). The exploration of content relevance requires the specification of the behavioral domain in question and specification of the task or test domain (Bachman, 1990; Messick, 1980). Content coverage, however, refers to the extent to which the test tasks adequately represent the behavioral domain in question (Bachman, 1990). Bachman (1990) considers examining the content as one of the first facets of a test that need to be taken into consideration by prospective test users. In fact, in designing a test, scholars start with a definition of the content or ability domain, or at the very least, with a list of content perspectives, from which they generate items, or test tasks.

Criterion-related evidence for validity includes concurrent and predictive types of validity. To measure criterion-related validity, there should be a comparison between the test scores with one or more external variables (called criteria) which offer a direct measure of the characteristic or behavior in question (Messick, 1990). In other words, to assure of the criterion-related validity, the tester searches for a relationship between a particular test and a criterion to which he/she wishes to make prediction. Concurrent validity indicates the extent to which an individuals' level on the criterion is related to their performance on a concurrent test (Fulcher & Davidson, 2007). Predictive validity, however, indicates "the extent to which an individual's future level on the criterion is predicted from prior test performance" (Messick, 1990, p. 11).

The above paragraphs indicate that in the past decades, content and criterion validities have been regarded as evidences or stages for the construct validity as a unified construct. Within the last two decades, however, the concept of validity has enjoyed further explorations and scrutiny. Detailed analysis of the concept resulted in the introduction of other types of evidence, in addition to content and criterion types, which can form validity as a unitary concept. Messick (2005) divides the validity into six sub-components, relevant to all educational and psychological measurements, including performance assessments. In this model, the unified validity includes content validity, structural validity, substantive validity, generalizability validity, external validity, and consequential validity. According to him, "evidence of content relevance, representativeness, and technical quality" consist the content facet of construct validity (p. 6). The substantive aspect includes "theoretical rationales for the observed consistencies in test responses, including process models of task performance" (p. 6), in addition to "empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (p. 6). The structural aspect examines "the extent to which the internal structure of the assessment reflected in the scores, including scoring rubrics as well as the underlying dimensional structure of the assessment tasks, is consistent with the structure of the construct domain at issue" (p. 6). The generalizability aspect appraises "the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks" (p. 6) and includes "validity generalization of test-criterion relationships" (p. 6). The external aspect of validity includes "convergent and discriminant evidence from multitrait-multimethod comparisons" (p. 6), along with "evidence of criterion relevance and applied utility" (p. 6). Finally, the consequential aspect evaluates "the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice" (p. 6).

In his elaboration of the nature of test validity, Weir (2005) provides elaborate discussion on theory-based validity and context validity as a priori validity evidence and scoring validity, criterion-related validity and consequential validity as a posterior validity evidence. As he maintains, theory-based validity necessitates that we describe, fully, the construct we are attempting to measure at the a priori stage. The fuller the description, "the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test" (p.18). Context validity appraises "the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample" (p.19). Weir employs scoring validity as "the superordinate for all the aspects of reliability" (p. 22). It concerns "the extent to which test results are stable over time, consistent in terms of the content sampling and free from bias". Criterion-related validity is concerned with "the extent to which test scores correlate with a suitable external criterion of performance" (p. 35). Lastly, consequential validity examines " whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values" (p. 37). More recently, Shaw and Weir (2007) introduce a framework of validity components which adds cognitive validity to the other components of validity introduced earlier by Weir (2005). As stated by them, cognitive validity requires "collecting both *a priori* evidence on the cognitive processing activated by the test task through piloting and trialling before the test event... and also *a posterior* evidence on constructs measured involving statistical analysis of scores following test administration" (p. 6).

However named, it seems that any sort of related information of a test has a significant contribution to its construct validity (Colliver et al., 2012; Cox & Malone, 2018). Messick (1989) believes that this contribution becomes stronger when there is an explicit measurement of the goodness-of-fit between the information and the theoretical logic which underlies the score interpretation.

In addition to the detailed analysis of the concept of construct validity and its various types of evidence, several approaches have been proposed for assessing the validity of test interpretations. The approaches introduced to examine the validity of criterion-referenced (CR) and norm-referenced (NR) tests seem to overlap. Among others, Hambelton (1982) provides a list of methods that can be used to ensure of the construct validity of CR tests. It includes content analysis, item-objective congruence analysis, Guttman scalogram analysis, exploratory and confirmatory factor analysis, experimental studies and the multitrait-multimethod approach, each of which is best suited for specific purposes. Regarding the construct validation of NR tests, Alderson et al. (1995) introduce test's correspondence with the theory, internal correlation addressed through factor analysis, and multitrait-multimethod matrix.

Though the literature on the validation processes of various types of tests is vast, the most well-known procedure for test validation is factor analysis (Hatch & Farhady, 1982; Kerlinger, 1979). Factor analysis, as a multivariate technique (Alavi & Ghaemi, 2011; Field, 2009; In'nami & Koizumi, 2011; Khine, 2013; Ockey & Choi, 2015; Sawaki 2012; Schmitt, 2011), refers to "an analytic method for determining the number and nature of the variables that underlie larger numbers of variables or measures" (Kerlinger, 1979, p. 180), "techniques for analyzing test scores in terms of some number of underlying factors" (Hatch & Farhady, 1982, p. 255), and "a number of related statistical techniques which help us to determine the characteristics which go together" (Bryman & Cramer, 1990, p. 253). Reyment and Joreskog (1993) define factor analysis as:

> A generic term that we use to describe a number of methods designed to analyze interrelationships within a set of variables or objects [resulting in] the construction of a few hypothetical variables (or objects), called factors that are supposed to contain the essential information in a larger set of observed variables or objects…that reduces the overall complexity of the data by taking advantage of inherent interdependencies [and so] a small number of factors will usually account for approximately the same amount of information as do the much larger set of original observations (p. 71).

There are two basic types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Meyers et al., 2006). In the exploratory analysis, the researcher attempts to identify the few themes, abilities, dimensions, or traits that underlie a relatively larger set of variables by examining the relationships among a set of measures (Bachman, 1990; Meyers et al., 2006). In the confirmatory mode, however, the researcher begins with "hypotheses about traits and how they are related to each other and attempts to either confirm or reject these hypotheses by examining the observed correlations" (Bachman, 1990, p. 260).

Empirical Investigations

In addition to the theoretical studies which detail various approaches utilized for estimating the construct validity of a test, there are extensive empirical studies investigating the construct validity of proficiency tests. Among other proficiency tests, the Test of English as a Foreign Language (TOEFL), in its paper based and Internet-based (iBT) format, has been scrutinized by many researchers interested in validation studies. Hale et al. (1988), Hale et al. (1989), and Kyle et al. (2016), for instance, studied the factor structure of the paper-based TOEFL, consisting of three sections: listening comprehension, structure and written expression, and vocabulary and reading comprehension. All three studies identified a distinct listening comprehension factor and multiple other correlated factors. Freedle and Kostin (1999) investigated the construct validity of minitalks of TOEFL. Though they found evidence supporting that reading and listening items load on two separate factors, their results showed many underlying similarities in the skills measured by TOEFL's listening and reading (minitalk) items.

Regarding the TOEFL internet-based test, Stricker et al. (2005) employed a multiple-group confirmatory factor analysis to examine the factor structure of a prototype of the TOEFL iBT called LanguEdge for Arabic, Chinese and Spanish native language groups. This prototype consisted of four sections of reading, listening, speaking and writing. The authors identified a correlated two-factor model - one for speaking and the other for a combination of reading, listening and writing - for the three language groups. Parallel to this line of research, Sawaki et al. (2009) investigated the factor structure of the TOEFL iBT. They conducted an item-level confirmatory factor analysis for a test completed by participants and could identify a higher-order factor model, with a higher-order general factor (ESL/EFL ability) and four first-order factors for reading, listening, speaking and writing. They found that integrated speaking and writing tasks, requiring language processing in multiple modalities, define the target modalities (speaking and writing). Their results supported the practice of reporting a total score and four scores corresponding to the modalities for the test.

In addition to the TOEFL, other proficiency tests have also been the subject of scrutiny in validation studies. Beglar and Hunt (1999), for instance, examined the construct validity of the revised versions of the University Word Level of Nation's Vocabulary Levels Test and the 2000 Word Level by employing Rasch and classical item analyses. They found that the new forms had statistically significant correlations with the TOEFL. The new versions were also found to be reliable with only three misfitting items. Kim and Kim (2017) validated an English placement test (EPT) developed for a General English Language Program (GELP), the goal of which was to improve reading, speaking and writing skills. The findings showed that the EPT was highly reliable. Additionally, item difficulty and item discrimination indices illustrated that the EPT was appropriately developed. In a recent article, Saito (2019) examined the factors underlying the nine vocabulary measures that were hypothesized to tap into appropriateness (global, semantic, and morphosyntactic accuracy) and sophistication (frequency, range, concreteness, meaningfulness, imageability, and hypernymy) aspects of L2 lexical proficiency. He submitted all participants' performance scores to a factor analysis with oblique rotation method. A three-factor solution was identified, accounting for 78.5% of the total variance in the nine lexical variables. Whereas all the appropriateness measures were clearly clustered into the one single factor, the sophistication measures were divided into two subcomponent factors. The results suggested that the corpus-based frequency and range measures were methodologically and thematically different from all of the abstractness-related measures (i.e., concreteness, meaningfulness, imageability, and pernymy).

The factor structure of proficiency tests has also been explored in the Iranian context. Salehi (2011), for instance, employed exploratory factor analysis to investigate construct validity of a reading comprehension section of University of Tehran English Proficiency Test (UTEPT). Principal component analysis extracted 11 factors out of the 35 items. Due to the unexpected number of factors, Salehi ran another method of extraction, principal axis factoring, to corroborate the findings. Surprisingly, the second method also yielded 11 factors. In a more recent study, Alavi et al. (2018) examined the construct validity of IELTS listening comprehension test (LCT), implementing structural equation modelling (SEM) and assessed differential item functioning (DIF) through cognitive diagnostic modelling (CDM) and Mantel Haenszel (MH). Initially, they administered a proficiency test designed by the university of Cambridge to 480 participants; next, they administered a 40-item IELTS LCT developed by the University of Cambridge to 463 participants, out of 480 participants. Data was analyzed with use of LISREL to explore the construct validity of the test. Additionally, for detecting the potential DIF items, MH and CDM were used to make the results of DIF related findings more reliable. The results of the first study confirmed an appropriate model fit, so that all four constructs, i.e., gap filling, diagram labelling, multiple choice and short answer on IELTS LCT, had a statistically significant contribution to IELTS LCT. The second study examined the DIF items to argue the validity of IELTS LCT. MH detected 15 DIF items and CDM detected at least 6 DIF items and at most 12 DIF items.

Though the above section may elucidate that there is a vast literature assessing the construct validity of various types of language proficiency tests, to our knowledge, there are, still, tests of proficiency type, which have been unnoticed in terms of their validation. An example would be the Iranian Ministry of Health Language Exam for which we could find no validation reports. Due to the significance of ensuring of validity of this high-stakes test, we addressed this neglected aspect of the exam in the present research.

Method

Data

Our data consisted of the scores of 987 MHLE test takers, including 518 female and 469 male master graduates, who took the test on October 8th, 2010. They had graduated from different universities across the country and had majored in different fields of study related to medical sciences, namely biostatistics, health economics, medical parasitology, medical ethics, medical immunology, epidemiology, health education, medical informatics, artificial limbs, medical bacteriology, reproduction biology, environmental health, clinical biochemistry, pregnancy health, professional health, nursing, molecular medicine, functional proteomics, medical entomology and vector control, clinical psychology, military psychology, medical biotechnology, medical genetics, health policy, health and social welfare, disasters and emergencies health, audiology, anatomical sciences, nutrition science, neuroscience, food science and technology, pharmacology, physiotherapy, physiology, medical physics, sport physiology, medical mycology, occupational therapy, speech therapy, health sciences management, social work, addiction studies, biomedical engineering, tissue engineering, medical nanotechnology, medical virology, hematology, medical education, and bacterial toxins.

It has to be noted that the prime reason we researched the 2010 test data was that the Evaluation Center officials did not permit us to have access to the more recent test scores for keeping privacy and confidentiality of the fairly recent test results, as they put it. To ensure that analyzing this version of the test has much relevancy to the present context, we had no option but to make sure that the test content, format and characteristics had remained almost untouched through the years. Our analysis of the MHLE test items administered thereafter by the Ministry along with informal talks with the Evaluation Center officials and managers of the language institutes which held preparation classes for the MHLE candidates, provided us with strong indications that the tests administered through the years have been almost constant in terms of their content and characteristics. Equally importantly, we found that, roughly, the same examination board, following the same policies for test development, has been involved in the test development and administration through the years. These indications prompted us to feel confident in analyzing the test and be hopeful that the findings can offer benefits to different communities involved.

Instrumentation

The instrument was the Ministry of Health Language Exam which is an English proficiency test administered by the Evaluation Center of the Ministry of Health, Treatment and Medical Education. The test includes 100 multiple-choice items with equal weighting and with no negative scoring. The rationale for no negative marking is not announced in the exam webpage, nor is any information provided on who the examination board for test development and administration are. Test takers who meet the cutoff point of 55 on the test are granted the language certificate which is a prerequisite for the doctoral entrance examination of the Ministry of Health, Treatment and Medical Education. The MHLE is administered five or six times a year on a regular basis. The exact date of the exam is announced a month in advance. The exam centers are primarily located in Tehran, with a few centers in other large and populated cities, including Mashhad, Tabriz, Esfahan, and Shiraz.

As also announced in various websites for the preparation of the exam, the MHLE consists of three sections. The first section is named Listening Comprehension, while the second and third sections are merely distinguished as Part 2 and Part 3. The listening comprehension section consists of 30 items (i.e., items 1 to 30) in three parts which measure understanding main ideas, listening for specific information, and inferring the speaker's meaning. Part 2 consists of items 31 to 70. While it is not explicitly mentioned in the test, Part 2 is devoted to assessing test takers' grammar and vocabulary skills. Though it is not always easy to distinguish between the two as some items seem to be addressing both, the first 16 items (i.e., items 31 to 46) are more aligned with learners' grammar knowledge. More specifically, these items, mainly, measure the examinees' mastery of verb tenses, modals, adverb of transition, passive voice, parts of speech, etc. The next 24 items of Part 2, however, (i.e., items 47 to 70) are devoted to vocabulary knowledge, primarily including sentence completion and synonym questions. The last part of the exam, named as Part three, assesses reading comprehension skill. This section consists of six passages with 30 items, measuring test takers' understanding of main ideas of the texts, their vocabulary knowledge, their making inferences, etc.

Data Analysis

In order to examine the construct validity of the MHLE, descriptive statistics, item analysis, reliability analysis, and KR20-if-item-deleted statistics were calculated before conducting factor analysis. To obtain these information, we, first, inserted our data, namely the item and total scores of 987 MHLE test takers, into the Test Analysis Program (TAP) (version 14.7.4). According to Brooks and Johanson (2003), this program reports test analysis information, including raw scores, percentage scores, summary statistics, reliability, standard error of measurement, item difficulty, item discrimination, and distractor analyses. It, also, details several item-total correlation indices, namely biserial, point-biserial, and adjusted point-biserial correlations (Crocker & Algina, 1986). Additionally, as Allen and Yen (1979) maintain, TAP provides some relatively unique features, such as providing confidence intervals for examinee scores; allowing the creation of a table of specifications and analyzing those subsets of items; creating individual grade reports for examinees; sorting item analysis and examinee results; allowing input of a grading scale so that letter grades can be assigned automatically to percentage scores; allowing user choice of proportions used for calculating discrimination indices; and calculating the number of items needed to attain a desired level of reliability, using the Spearman- Brown prophecy formula. In our study, employing the TAP, initially, we obtained descriptive statistics of the sample test performance data including the mean, median and standard deviation of the set of scores. Next, item analysis of the test and its reliability estimate were calculated. Item analysis consisted of analyzing item difficulty, item discrimination (ID) and item-total correlation (ITC) (Downing & Haladyna, 2006). In classical test theory, item analysis is considered as a source of evidence for construct validity (Van der Walt & Steyn, 2008), and due to the fact that factor analysis is based upon correlation matrices, determining problematic items using item analysis has been regarded as being a crucial preliminary step (O'Connor, 2000).

Item difficulty, also called Item Facility (IF), examines the percentage of test takers correctly answering a given item. According to Brown (2005), item facility ranges from .00 to 1.00 and ideal items in a norm-referenced test (NRT) have an IF of 0.50. In the present study, based on Brown (2005), items with IFs between 0.15 and 0.85 are considered acceptable. Item discrimination indicates "the degree to which an item separates the students who performed well from those who did poorly on the test as a whole" (Brown, 2005, p. 68). Theoretically, item discrimination ranges from -1.00 to 1.00 and the higher the ID index, the better the item discriminates. According to Brown (2005), items with ID index higher than 0.40 are considered good items; items with ID indices between 0.20 and 0.29 are considered acceptable items; and, items with ID indices lower than 0.19 are poor items which need to be revised or discarded from the test battery. Falvey et al. (1994) consider items with the discrimination value of below 0.20 as unacceptable items. In the present study, the cutoff point for item discrimination was considered to be below 0.15; otherwise, far too many items would have been deleted. This decision is justified based on Hughes (2003, p. 226) who maintained "there is no absolute value that one can give for a satisfactory discriminating index. The important thing is the relative size of the indices". In addition to identifying items with below 0.15 ID index in the present study, items with negative ID values were located and removed. According to Downing and Haladyna (2006), the negative ID index of an item indicates that the items test something different because the students in the low group have outperformed the students in the high group. Also, too easy items may result in negative ID index (Downing & Haladyna, 2006).

In addition to item difficulty and item discrimination, item-total correlations were calculated through adjusted point-biserial correlation; that is, the correlation of any single item with the total test. According to Falvey et al. (1994), items with correlations below 0.20 with the total test and items with negative values of ITC are considered unacceptable items. Negative values of ITC might indicate a different construct of the item (Alavi, 1997). Removal of items with low and negative ITCs increases the reliability of the test (Downing & Haladyna, 2006).

Following item analysis, the test reliability was examined. Reliability has been defined as "the extent to which the results can be considered consistent or stable" (Brown, 2005, p. 175). Different strategies, including test-retest, equivalent forms, and internal consistency are employed to estimate the test reliability (Hughes, 2003). In the present study, the internal consistency of the test was examined through KR20 coefficient due to the fact that this measure calls for one test and a single administration of the test (Stoker & Impara, 1995). To make sure that all items contribute to the test reliability, KR20-if-item-deleted statistics was also checked. According to Jackson et al., (2002), when the KR20 value is higher than the current index with the item deleted, one should consider deleting this item to improve the overall reliability of the test. Please note that in order to meet the requirements of the KR20 coefficient, the scores of the multiple-choice items of the test were transformed into categorical data.

Next, the items were coded in the *R* software *paramap* package, version 1.0., and an exploratory factor analysis (EFA) was conducted on the test. To determine the optimal number of factors to retain in the EFA, we applied two modern validation procedures widely recommended by statisticians (e.g. O'Connor, 2000), namely Horn's parallel analysis and Velicer's minimum average partial (MAP) test. Due to the fact that we assumed the existence of correlation between factors (Tabachnick & Fidell, 2008), we employed Principal Axis Factoring (PAF) method with Oblimin rotation to extract factors. Before conducting factor analysis, data adequacy and sphericity were examined by KMO and Bartlett's test.

Results and Discussion

Descriptive Statistics

Table 1 summarizes the descriptive statistics of the total test and its subsections. As the table indicates, the mean of the examinees' performance on the total test was 41.73 which is not a high value considering the total score for the whole test (which was 100). The highest mean score for the items of the test subsections belonged to the vocabulary section (which was 13.18 out of 24.00). The median of the total test was 40.00. The low values of the mean and median indicated that the overall performance of the students was not satisfactory. This might be explained by the overall difficulty of the test as well as the item difficulty and item discrimination indices. The standard deviation of the total test was 11.99 and the largest standard deviations in the subsets are for listening and reading comprehension sections (4.92 and 4.02 respectively).

Table 1.
Descriptive Statistics of the Total Test and the Main Sections of the MHLE

Sections	No. of Items	Mean	Median	Std. Deviation	No. of Subjects
Listening Comprehension	30	10.36	9.00	4.92	987
Grammar	16	6.23	6.00	2.41	987
Vocabulary	24	13.18	13.00	3.97	987
Reading Comprehension	30	11.94	11.00	4.02	987
Total	100	41.73	40.00	11.99	987

Item Analysis

Table A1 in the Appendix reveals the item difficulty, item discrimination, and adjusted point-biserial correlation indices. As the table shows, items 14, 84, and 100 have IFs lower than 0.15 and are regarded as difficult items. On the other hand, item 49 with an IF greater than 0.85 is considered an easy item. Similar to difficult questions, easy items are not considered ideal types of questions and need to be revised or removed accordingly.

Also, as the table indicates, items 1, 5, 6, 9, 15, 19, 23, 24, 25, 26, 48, 51, 52, 53, 63, 65, 68, 77, 86, and 88 have item discrimination indices higher than 0.40 and are considered to be good items in this regard. Items 2, 3, 11, 17, 27, 29, 33, 34, 35, 40, 41, 42, 45, 47, 54, 59, 60, 70, 73, 78, 79, 83, 93, 97, 98, and 99 have ID indices between 0.20 and 0.29 and are acceptable items. Items 12, 14, 18, 22, 37, 38, 43, 44, 58, 61, 71, 74, 84, 90 and 100 have item discrimination indices of less than 0.15 and are regarded as poor items. There are also two items, namely items 20 and 67, which have negative item discrimination indices. Several reasons might explain an item's negative ID index. At times, the item measures something irrelevant to what it was supposed to measure which makes the test takers too confused to respond appropriately. Additionally, low or high item facility or difficulty indices might result in a negative ID index. Equally important is the ambiguous and unspecific test instructions.

The table also presents adjusted point-biserial correlations of every single item with the total test. According to the criterion proposed by Falvey et al. (1994) and as the table reveals, items 2, 7, 10, 12, 14, 18, 27, 30, 35, 37, 38, 39, 41, 42, 43, 44, 50, 54, 55, 60, 61, 72, 73, 74, 75, 80, 81, 83, 84, 85, 89, 90, 96, 97, and 100 have very low correlations (lower than .2) with the total test. In addition to these items, questions 20, 58, 67 and 71 are considered poor items due to their negative values of ITC. As explicated by Alavi (1997), the negative value of an item's ITC might indicate a different construct of the item. In this study, the low or negative ITC of the listening comprehension questions can further be explained on other grounds. Due to the fact that the test' audio recordings are played in the test setting, and only once, most probably, the administration condition, including the technical faults, noise or poor quality of the tapes, plays a crucial role in testees' performance which, in consequence, results in the items' poor ICT.

Reliability Analysis

The KR20 reliability statistics revealed that the MHLE enjoys a reliability of 0.862, which is considered an acceptable reliability index (Hughes, 2003). Table A2 in the Appendix presents KR20-if-item-deleted of the total test. As the table indicates, the removal of items 10, 14, 18, 20, 38, 41, 42, 43, 44, 50, 55, 58, 61, 67, 71, 73, 74, 80, 81, 84, 90, 97, and 100 results in a reliability higher than 0.862.

Factor Analysis on the Total Test

To determine the component abilities underlying performance on the MHLE, a factor analytic study was undertaken. Before conducting factor analysis, data adequacy and sphericity were examined through KMO and Bartlett's test. Table 2 presents the related findings.

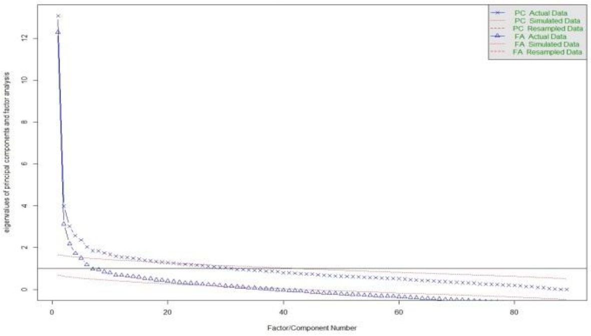
Table 2.
KMO and Bartlett's Test on the Total Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.826
Bartlett's Test of Sphericity	Approx. Chi-Square	13565.995
	df	4950
	Sig	0 .000

According to Hinton et al. (2004), when the KMO test result is 0.5 or higher, the data are suitable for factor analysis. According to our findings, the KMO test statistics for the present study was equal to 0.82 which was far higher than the critical value of 0.5. Hinton et al. (2004) also believe that a significance level of $p < 0.05$ for Bartlett's test of sphericity indicates that it is safe to continue with factor analysis. Our Bartlett's test digit (i.e. $p < 0.001$) confirmed that the assumptions of performing factor analysis were met.

In order to be able to compare the results of the analysis before and after removing problematic items, factor analysis was first conducted on the whole test, including 100 items, using both parallel analysis and MAP test. The results of parallel analysis suggested overfactoring, i.e. 31 factors, which might be explained for two reasons: First, parallel analysis appears to be sensitive to the number of variables (test items) and some emerged factors may be trivial ones (O'Connor, 2000); second, the heterogeneous nature of the test could also lead to the overfactoring. The MAP test produced five factors. However, scrutinizing items loading on the factors extracted, we found no clear pattern. For this reason, we planned to conduct factor analysis for the second time after removing the problematic items, already identified through item analysis and KR20-if-item-deleted statistics. Since many of the items would have been removed based on item analysis and KR20-if-item-deleted indices, we decided to delete only those items which were problematic according to three of the four criteria. In consequence, items 14, 18, 20, 38, 43, 44, 58, 61, 67, 71, 74, 84, 90 and 100 were removed from later analyses. Again, parallel analysis overestimated the number of factors, extracting 25 factors. Figure 1 illustrates the scree plot of the analysis.

Figure 1.
Parallel Analysis Scree Plot



The MAP test, however, produced results with two, three, four, five, six, and seven factors. Due to logical considerations and our knowledge about the contents of the exam, which was a proficiency test with an average difficulty level consisting of three parts but assessing four types of knowledge- namely listening comprehension, grammar, vocabulary, and reading comprehension- we found the four-factor solution more logical. In the following, each factor is introduced and the logic behind item loadings on each factor is discussed. Table A3 in the appendix details the pattern matrix of 86 items on four factors extracted based on the correlation matrix of each single item with each factor.

Factor One. Table 3 provides information on 24 items; namely items 1, 4, 5, 6, 9, 15, 17, 19, 22, 24, 25, 26, 28, 31, 32,33, 34, 39, 41, 45, 48, 52, 54, and 79 which loaded on factor one. From among these items, the majority, i.e. 13 items, had been included in the listening comprehension section of the test and had higher correlations with the factor (from 0.37 to 0.66) compared with other items. Seven items pertained to the grammar section; three items were related to the vocabulary part, and one item aimed at checking the test takers' reading comprehension skill. The loadings of these items on factor one are not high, however, ranging from 0.32 to 0.45.

Table 3.
Items Loading on Factor One

Listening Comprehension				Reading	Comprehension	Grammar	Vocabulary
Items	Loading	Items	Loading	Items	Loading	Items	Loading
1	.49	31	.38	48	.45	79	.32
4	.46	32	.41	52	.33		
5	.66	33	.35	54	.35		
6	.65	34	.34				
9	.60	39	.34				
15	.45	41	.33				
17	.50	45	.38				
19	.50						
22	.43						
24	.64						
25	.57						
26	.45						
28	.37						

It can perhaps be claimed that most listening comprehension items, including items 1, 4, 6, 9, 15, 17, 19, 22, 24, 25 and 26, assess understanding local linguistic meanings. Item 5 measures knowledge of the sound system, while item 28 seems to have been designed to evaluate test takers' inference of an implied meaning and intention. According to Buck (2001), knowledge of the sound system includes relevant aspects of grammatical knowledge- namely phonology, stress and intonation- and understanding local linguistic meanings includes the whole of grammatical knowledge- not only phonology, stress and intonation, but also vocabulary and syntax, as well as the ability to use that knowledge automatically in real time. In addition to listening comprehension items, seven items assessing students' knowledge of grammar loaded on factor one (i.e. items 31, 32, 33, 34, 39, 41 and 45, with loadings ranging from 0.33 to 0.41, which is slightly lower compared with the other items related to factor one). Except items 34 and 45, the rest appear to be inference items in which the students have to obtain and infer some information from the stem of the items to be able to answer the questions. All vocabulary items loading on this factor, i.e. items 48, 52, and 54 are open questions which contain an underlined term, the synonym of which the students should select from among the four responses. Due to the point that the students have to infer some information from the stem of the item to be able to answer the item correctly, items of this type might be labeled inference questions. Likewise, the reading comprehension item which correlated with factor one, i.e. item 79, is an inference question measuring students' general comprehension of the passage, not addressing specific information in the text.

As is evident from the table, items from all subsections of the test loaded on factor one. The highest to the lowest loadings belong to listening comprehension, vocabulary, grammar and reading comprehension items respectively. Due to the disparity of item loadings from different divisions of the test, labeling the extracted factor is not an easy task. However, it appears that 10 items (one from listening comprehension, five from grammar, three from vocabulary and one from reading comprehension sections), all, intend to examine the students' inferencing abilities. Nonetheless, the loading of items from different subsections of the MHLE on the same factor was contrary to expectations.

Factor Two. All items loading on factor two, namely items 2, 3, 7, 8, 11, 12, 13, 16, 21, 23, 27, 29, 30, were listening comprehension items. Table 4 shows item loadings on this factor.

Table 4.
Items Loading on Factor Two

Listening Comprehension	
Items	Loading
2	-0.56
3	0.51
7	0.47
8	0.46
11	-0.42
12	0.60
13	-0.46
16	-0.45
21	0.46
23	0.43
27	-0.47
29	0.48
30	0.51

As tables 3 and 4 indicate, listening comprehension items loaded on two separate factors (13 items on factor one and 13 items on factor two). An explanation for this result can be the underlying abilities these items measure. As mentioned previously, listening comprehension items loading on factor one mostly assess the examinees' knowledge of understanding local linguistic meanings. However, listening items correlating with factor two can be claimed to have aimed at measuring students' understanding of inferred meanings in the audio texts. In addition to these inferencing items correlating with factor two, item 28 was also reported above to have been intended for measuring candidates' inferencing abilities. Surprisingly, however, item 28 correlated with factor one. These results are unexpected in two respects. First, it was expected that all items examining testees' ability in inferencing be loaded on the same factor, whereas the findings revealed that the items which aimed at evaluating candidates' inferencing skills were divided between factors one and two. Even if we consider the inferencing items of the listening comprehension section to be a separate factor, there is no logic behind the loading of item 28 on factor one. Second, there is no justification for why listening comprehension questions which addressed learners' knowledge of understanding local linguistic meanings correlated on the same factor with 9 vocabularies, grammar, and reading comprehension items intended to examine students' inferencing skill. Another equally important point about factor two is that this factor appears to be highly related to the difficulty criterion and not just similar content, since items with average and high IF values (i.e. IF ≥ 0.50) correlate negatively with this factor, while difficult items (i.e. IF < 0.50) correlate positively with it. The only explanation for this finding seems to be the tests' administration conditions, including poor quality of audio recordings and the existence of extra noise.

Factor Three. Eighteen items- i.e. items 42, 46, 47, 49, 51, 53, 56, 57, 62, 63, 64, 65, 66, 68, 69, 70, 76, and 77- loaded on factor three. Fifteen of these items were from what may be considered the vocabulary section of the MHLE, with loadings higher than the items belonging to the grammar and reading comprehension parts of the test. Table 5 depicts items loading on this factor.

Table 5.
Items Loading on Factor Three

Grammar		Vocabulary		Reading Comprehension	
Items	Loadings	Items	Loadings	Items	Loadings
42	.40	47	.53	76	.32
46	.43	49	.42	77	.31
		51	.55		
		53	.53		
		56	.37		
		57	.56		
		62	.33		
		63	.45		
		64	.57		
		65	.73		
		66	.44		
		68	.59		
		69	.77		
		70	.47		

On the other hand, items 42 and 46, which appear to be intended to tap learners' grammatical knowledge (item 42 assesses knowledge about verb sequences and item 46 measures knowledge about prepositions) also load on factor three, which does not appear to be very logical. Scrutinizing the two items from the reading comprehension section which loaded on factor three reveals that item 77 can be considered a vocabulary item inserted in the reading comprehension section, as it requires a synonym for a word in the text. Item 76, however, measures the ability to draw inferences from content; hence, there seems to be no obvious reason why it loaded on this factor. At the same time, there is no justification for why 15 vocabulary items loaded on factor three while three others loaded on factor one.

Factor Four. Items 82, 86, 87, 88, 89, 91, 93, 94, 95, 96, 98 and 99 loaded on factor four. All these items had been included in the reading comprehension section of the MHLE. Table 6 summarizes items loading on this factor.

Table 6.
Items Loading on Factor Four

Reading Comprehension	
Items	Loading
82	.31
86	.56
87	.58
88	.65
89	.56
91	.50
93	.48
94	.54
95	.48
96	.48
98	.38
99	.30

All in all, the reading comprehension questions which loaded on factor 4 seem to have aimed at checking specific information in the text. Items 87, 93, 95, and 96, however, appear to evaluate the test takers' general comprehension skills. Overall, however, it seems that factor four can be considered related to the reading comprehension skill.

Altogether, a detailed account of the whole findings reveals that from among the 86 items included in the factor analysis, 67 had loadings higher than 0.3 with the factors extracted. Nineteen items, including items 10, 35, 36, 37, 40, 50, 55, 59, 60, 72, 73, 75, 78, 80, 81, 83, 85, 92, and 97, however, did not load significantly on any factor. It needs to be mentioned that items 37, 59, 83, and 97 had loadings of 0.14 to 0.25 with factor one; items 35, 36, 50, 55, 60, 72, 73, 75, 80, and 92 had loadings of 0.12 to 0.28 with factor three; and items 10, 40, 78, 81, and 85 had loadings of 0.15 to 0.29 with factor four. Examining the item analysis of these questions confirms that 14 of these 19 items were found to be statistically problematic in terms of one or two of their item characteristics. Items 10, 50, 55, 73, 80, and 81 did not have suitable item-total correlations or KR20-if-item-deleted indices. Items 10, 35, 37, 50, 55, 60, 72, 75, 80, 85, and 97 did not have satisfactory adjusted point-biserial correlations. Recall that prior to conducting the factor analysis, we made a decision to remove only items which were problematic in three of their four item characteristics so as to retain the most possible number of items for the analysis. This decision might partly explain why 19 items failed to significantly correlate with any of the factors. It is more difficult, however, to explain why the other five items (i.e. items 36, 40, 59, 78, and 92) do not load on any of the extracted factors, although we should perhaps point out that their loadings were somewhat higher (i.e. between 0.20 to 0.26) than the obviously faulty items. Items 40 and 59 had correlations of 0.26 with factors four and one respectively. Items 36 and 78 had correlations of 0.20 with factors three and four respectively. And item 92 correlated with factor three at 0.25.

As the preceding section indicates, the item loadings, did not reveal a precise pattern. Simply put, contrary to our expectation, items on listening comprehension, grammar, vocabulary and reading comprehension did not load on separate factors; Several explanations may be tentatively offered for this finding: First of all, there were many items which were found to be poor based on their IF, ID, adjusted point-biserial correlation and KR20-if-item-deleted indices. However, we only removed those items which were seen to be problematic in three of the four criteria; hence, other problematic items were retained which might have affected the results. Another point which might have affected the results is the heterogeneity of the items of the MHLE which appear to have been collected from different available proficiency tests, instead of having been developed for the purpose of the MHLE exam. Another possibility is the specific background knowledge needed on the part of the examinees to respond to some of the questions, particularly in the listening and reading parts. Regarding the listening comprehension section, as mentioned previously, the administration conditions of the test, such as listening to the audio files without headphones, and, in consequence, the presence of background noise, may also have influenced candidates' performance. Brindley (1998) enumerates a range of factors affecting testees' performance on listening comprehension tests, including lack of background knowledge and the noise of the setting. Other factors he lists like the nature of the input (speech rate, length, background, syntax, vocabulary, noise, accent, register, propositional density, amount of redundancy, etc.), the nature of the assessment task (amount of context provided, clarity of instructions, availability of question preview, whether the task calls for recognition only or synthesis, etc.), and individual listener factors (memory, interest, background knowledge, motivation, etc.) might have affected the MHLE test takers' performance on listening comprehension questions as well. All things considered, we come up with the unsatisfactory conclusion that the MHLE lacks a clear factor structure, not distinct in terms of listening comprehension, grammar, vocabulary and reading comprehension.

Conclusion

This study was an attempt to investigate the construct validity of the Ministry of Health Language Exam. To be more specific, it addressed the distinctness of the test in terms of listening comprehension, grammar, vocabulary and reading comprehension. Based on the view of validity as a unitary concept, an attempt was made to collect various types of evidence to check construct validity of the test. For this purpose, descriptive statistics, reliability analysis, item analysis, and factor analysis were conducted. The low values of the mean and median indicated that the overall performance of the students was not satisfactory. The results of the reliability analysis were acceptable; however, the item analyses detected many problematic items. Finally, exploratory factor analysis, applying parallel analysis and Velicer's MAP test, was conducted on the total test. While results of parallel analysis suggested overfactoring, MAP test produced two to seven factors. Scrutinizing items loading on these factors, we could not find any clear pattern. Exploratory factor analysis was done for the second time with the poor items, i.e. 14 items problematic on three of the four item characteristics, removed. Again, parallel analysis resulted in overfactoring and the MAP test extracted two to seven factors. Having knowledge of the test content which composed of four sections and inspecting the item loadings, we selected the four-factor result as being more logical.

Analyzing the results indicated that the majority of items loading on factor one, i.e. 13 items, were listening comprehension items with high loading values. Eleven items from the grammar, vocabulary and reading comprehension sections also loaded on this factor. A detailed analysis of the items loading on this factor indicated that 10 items required the examinees to infer meaning from the audio or the text. The most number of listening questions, however, assessed understanding local linguistic meanings and one item measured knowledge of the sound system. Another 13 items from listening comprehension section loaded on factor two. Since no other item from other sections of the test loaded on this factor, this factor could be safely labeled listening comprehension. Surprisingly, however, like 10 questions of other sections loading on factor one, all listening items correlating with factor two seemed to have aimed at measuring the test takers' inferencing ability. The majority of items loading on factor three, i.e. 14 items, were related to vocabulary section, with a few grammar and reading comprehension questions. Since the grammar items assessed knowledge about verb sequences and prepositions, there is no logic behind their correlating with factor three. Regarding reading comprehension questions, one item was a synonym question, which could be considered as a vocabulary item inserted in reading comprehension section. Another reading question, however, measured the ability to draw inferences from content; hence, there is no explanation for its loading on this factor. Factor four was an exclusively reading comprehension factor. All the items, i.e. 12, were reading comprehension questions. The items, however, did not share the same underlying component abilities. Four of them appeared to evaluate candidates' general comprehension skills. Hence, they seemed to be more apt to have been loaded on factor one. Eight other items seemed to have aimed at checking specific information in the text. Nineteen items did not load significantly on any factor.

As the results revealed, findings were rather disappointing. Fourteen items were found problematic based on three criteria. Nineteen items did not load significantly on nay factor. A clear pattern of item loading was not found for many items. These findings can be viewed as evidences on the necessity of revising the MHLE. The first point to consider is that item characteristics be closely examined in initial steps of test development, since problematic items seriously threaten the validity of the test. Next, questions comprising future tests of MHLE have to be developed particularly for the purpose of the exam, instead of being compiled from various other available proficiency tests. Also, caution needs to be exercised in including audio and text materials which are not biased in favor of students from particular majors. In addition to the test and individual variables, test administration conditions should be improved in order not to contribute adversely to the candidates' performance. Altogether, it is highly recommended that, prior to all its administration, the MLHE be analyzed in depth in term of its item analysis, reliability and validity. Hopefully, these considerations help turn the MHLE into a highly valid high stakes test capable of selecting English proficient students for furthering their studies at the doctorate level.

Further studies can be conducted on the this nationwide high-stakes test by adopting a mixed-method approach. Future research can include interviews with stakeholders, that is test-takers and test developers, to provide a more comprehensive view of the validity of the test. Other ways of measuring validity can also be employed by comparing the test to other tests that measure similar qualities to see how highly correlated the two measures are, which is an indication of the validity of the test.

References

Alavi, M. (1997). An investigation of the construct validity of reading comprehension test in an academic context: Using rhetorical structure theory. *Paper presented in BALEAP Conference,* University of Wales, Swansea, UK.

Alavi, S. M., & Ghaemi, H. (2011). Application of structural equation modeling in EFL testing: A report of two Iranian studies. *Language Testing in Asia*, *1*(3), 22-35. https://doi.org/10.1186/2229-0443-1-3-22.

Alavi, S. M., Kaivanpanah, S., & Masjedlou, A.P. (2018) Validity of the listening module of international English language testing system: Multiple sources of evidence. *Language Testing Asia*, *8.* https://doi.org/10.1186/s40468-018-0057-4

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Reviews of Psychology*, *37*(1), 1-15.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary test. *Language Testing, 16*(2), 131-162.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*, 171-191.

Brooks, G. P., & Johanson, G. A. (2003). TAP: Test analysis program*. Applied Psychological Measurement, 27*(4), 303-304.

Brown, J. D. (2005). *Testing in language programs.* McGraw Hill.

Brown, T. (2010). Construct validity: A unitary concept for occupational therapy assessment and measurement. *HKJOT*, *20*(1), 30-42.

Bryman, A., & Cramer, D. (1990). *Quantitative data analysis for social scientists*. Routledge.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity… and back? *Medical Education*, *46*(4), 366-371.

Cox, T. L., & Malone, M. E. (2018). A validity argument to support the ACTFL Assessment of Performance Toward Proficiency in Languages (AAPPL). *Foreign Language Annals*, *51*(3), 548-574.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory.* Holt, Rinehart, and Winston.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Lawrence Erlbaum Associates Publishers.

Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students*. Longman.

Field, A (2009). *Discovering statistics using SPSS*. Sage Publications.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2-32.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language* (TOEFL Research Report No. 32). Educational Testing Service.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, Jr., J. W. (1988). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6*(1), 47-76.

Hatch, E., & Farhady, H. (1982). *Research design and statistics for applied linguistics.* Newbury House.

Hinton, P. R., Brownlow, C., McMurray, I., & Cozens, B. (2004). *SPSS explained*. Taylor & Francis.

Hughes, A. (2003). *Testing for language teachers.* Cambridge University Press.

In'nami, Y, & Koizumi, R. (2011). Structural equation modelling in language testing and learning research: A review. *Language Assessment Quarterly*, *8*(3), 250-273. https://doi.org/10.1080/15434303.2011.582203.

Jackson, T. R., Draugalis, J. R., Slack, M. K., Zachry, W. M., & D'Agpstino, J. (2002). Validation of authentic performance assessment: A process suited for Rasch modeling. *American Journal of Pharmaceutical Education*, *66*(3), 233-243.

Kerlinger, N. F. (1979). *Behavioral research: A conceptual approach.* Holt, Rinehart and Winston.

Khine, M. S. (2013). *Application of structural equation modeling in educational research and practice*. Sense Publishers.

Kim, Y. M., & Kim, M. (2017). Validations of an English placement test for a general English language program at the tertiary level. *JLTA (Japan Language Testing Association) Journal, 20*, 17-34.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319-340.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(10), 12-27.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1990). *Validity of test interpretation and use* (Research Report No. 90.11). Educational Testing Service.

Messick, S. (2005). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation.* Sage Publications.

Moses, M. S., & Nanna, M. J. (2007). The testing culture and the persistence of high stakes testing reforms. *Education and Culture, 23*(1)*,* 55-72.

Ockey, G, & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*(3), 305-319. https://doi.org/10.1080/15434303.2015.1050101.

Reyment, R., & Joreskog, K. G. (1993). *Applied factor analysis in the natural sciences.* Cambridge University Press.

Roever, C. (2001). Web-based language testing. *Language Learning and Technology, 5*(2), 84-94.

Saito, K. (2019). To what extent does long-term foreign language education improve spoken second language lexical proficiency? *TESOL Quarterly,53*(1), 82-101.

Salehi, M. (2011). On the construct validity of the reading section of the University of Tehran English Proficiency Test. *Journal of English Language Teaching and Learning*, *222,* 129-159.

Sawaki, Y (2012). *Factor analysis: The encyclopedia of applied linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0407.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1)*,* 5-30.

Schmitt, T. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psycho-educational Assessment*, *29*(4), 304-321. https://doi.org/10.1177/0734282911406653.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford University Press.

Stapleton. C. D. (1997, January). *Basic concepts in exploratory factor analysis (EFA) as a tool to evaluate score validity: A right-brained approach* [Paper presentation]. The annual meeting of the Southwest Educational Research Association, Austin.

Stoker, H. W., & Impara, J. C. (1995). Basic psychometric issues in licensure testing. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 167-186). Nebraska Series on Measurement and Testing.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the language test across language groups (TOEFL Monograph Series MS-32).* Educational Testing Service.

Tabachnick, B. G., & Fidell, L. S. (2008). *Using multivariate statistics* (5th Ed.). Pearson.

Van der Walt, J. L., & Steyn, F. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics*, *38*, 191-204.

Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan.

Appendix

Tables A1-A3

Table A1.

IF, ID and Point-Biserial Correlation Estimates of the MHLE

Items	Item Difficulty	Discrimination index	Adjusted Point Biserial
1	0.43	0.46	0.37
2	0.61	0.20	0.16
3	0.20	0.23	0.30
4	0.28	0.32	0.30
5	0.35	0.42	0.35
6	0.35	0.48	0.40
7	0.30	0.16	0.18
8	0.28	0.35	0.33
9	0.32	0.43	0.36
10	0.24	0.16	0.14
11	0.45	0.23	0.20
12	0.16	0.11	0.16
13	0.48	0.31	0.23
14	0.09	0.03	0.07
15	0.54	0.57	0.43
16	0.45	0.31	0.24
17	0.24	0.27	0.29
18	0.35	0.08	0.06
19	0.34	0.45	0.37
20	0.32	-0.01	-0.03
21	0.32	0.38	0.36
22	0.16	0.17	0.24
23	0.53	0.47	0.35
24	0.45	0.49	0.39
25	0.25	0.52	0.49
26	0.56	0.41	0.31
27	0.50	0.23	0.18
28	0.24	0.31	0.32
29	0.30	0.25	0.22
30	0.29	0.19	0.19
31	0.75	0.32	0.25
32	0.44	0.31	0.22
33	0.74	0.29	0.25
34	0.20	0.23	0.25
35	0.40	0.25	0.17
36	0.59	0.36	0.25
37	0.26	0.12	0.16
38	0.22	0.15	0.09
39	0.24	0.18	0.17
40	0.30	0.22	0.23
41	0.36	0.20	0.13
42	0.58	0.21	0.13
43	0.37	0.02	0.00
44	0.17	0.07	0.09
45	0.20	0.21	0.24
46	0.42	0.37	0.29
47	0.85	0.21	0.22
48	0.62	0.41	0.31

Items	Item Difficulty	Discrimination index	Adjusted Point Biserial
49	0.90	0.18	0.20
50	0.37	0.17	0.10
51	0.46	0.61	0.46
52	0.36	0.44	0.35
53	0.49	0.46	0.35
54	0.71	0.25	0.19
55	0.41	0.19	0.15
56	0.53	0.35	0.21
57	0.74	0.39	0.32
58	0.47	0.00	-0.03
59	0.50	0.29	0.22
60	0.69	0.23	0.17
61	0.30	0.13	0.14
62	0.31	0.32	0.32
63	0.47	0.54	0.40
64	0.71	0.34	0.27
65	0.56	0.41	0.31
66	0.58	0.31	0.25
67	0.16	-0.08	-0.12
68	0.55	0.60	0.43
69	0.74	0.32	0.27
70	0.74	0.23	0.20
71	0.68	0.01	-0.04
72	0.85	0.18	0.16
73	0.53	0.26	0.14
74	0.22	0.09	0.06
75	0.20	0.18	0.18
76	0.77	0.31	0.27
77	0.50	0.42	0.29
78	0.59	0.29	0.22
79	0.23	0.21	0.24
80	0.72	0.17	0.10
81	0.25	0.17	0.12
82	0.42	0.37	0.28
83	0.42	0.22	0.16
84	0.05	0.03	0.10
85	0.29	0.19	0.19
86	A0.42	0.52	0.40
87	0.42	0.33	0.27
88	0.56	0.46	0.35
89	0.56	0.19	0.16
90	0.22	0.03	0.04
91	0.34	0.30	0.27
92	0.47	0.30	0.22
93	0.39	0.29	0.23
94	0.36	0.33	0.28
95	0.16	0.18	0.21
96	0.21	0.18	0.18
97	0.49	0.24	0.13
98	0.36	0.28	0.24
99	0.22	0.23	0.20
100	0.08	0.02	0.00

Table A2.
Item Total Statistics

Item	KR20 if Item Deleted	Item	KR20 if Item Deleted	Item	KR20 if Item Deleted
1	0.859	39	0.861	77	0.860
2	0.861	40	0.861	78	0.861
3	0.860	41	0.862+	79	0.860
4	0.860	42	0.862+	80	0.862+
5	0.859	43	0.863+	81	0.862+
6	0.859	44	0.862+	82	0.860
7	0.861	45	0.860	83	0.861
8	0.859	46	0.860	84	0.862+
9	0.859	47	0.861	85	0.861
10	0.862+	48	0.860	86	0.858
11	0.861	49	0.861	87	0.860
12	0.861	50	0.862+	88	0.859
13	0.861	51	0.858	89	0.861
14	0.862+	52	0.859	90	0.863+
15	0.858	53	0.859	91	0.860
16	0.860	54	0.861	92	0.861
17	0.860	55	0.862+	93	0.860
18	0.863+	56	0.861	94	0.860
19	0.859	57	0.860	95	0.861
20	0.864+	58	0.864+	96	0.861
21	0.859	59	0.861	97	0.862+
22	0.860	60	0.861	98	0.860
23	0.859	61	0.862+	99	0.861
24	0.859	62	0.859	100	0.862+
25	0.858	63	0.858		
26	0.860	64	0.860		
27	0.861	65	0.860		
28	0.860	66	0.860		
29	0.861	67	0.864+		
30	0.861	68	0.858		
31	0.860	69	0.860		
32	0.861	70	0.861		
33	0.860	71	0.864+		
34	0.860	72	0.861		
35	0.861	73	0.862+		
36	0.860	74	0.862+		
37	0.861	75	0.861		
38	0.862+	76	0.860		

Note. + indicates that KR20 (0.862) improves if the item is removed

Table A3.
Standardized Loadings (Pattern Matrix) Based upon Correlation Matrix

Item	MR1	MR3	MR4	MR2	h2	u2
1	**.49**	.09	-.06	.40	.473	.53
2	.27	.10	.04	**-.56**	.376	.62
3	.29	.03	.14	**.51**	.462	.54
4	**.46**	-.05	.14	.09	.279	.72
5	**.66**	-.05	-.01	-.03	.399	.60
6	**.65**	.03	.00	.00	.439	.56

Item	MR1	MR3	MR4	MR2	h2	u2
7	.19	-.02	.04	**.47**	.295	.71
8	.33	.07	.09	**.46**	.436	.56
9	**.60**	.01	.01	.01	.373	.63
10	.13	.02	**.15**	.11	.076	.92
11	.28	.05	.12	**-.42**	.262	.74
12	-.05	.19	.04	**.60**	.404	.60
13	.34	.10	.07	**-.46**	.325	.67
15	**.45**	.14	.09	.33	.475	.53
16	.29	.10	.12	**-.45**	.307	.69
17	**.50**	-.08	.12	.03	.283	.72
19	**.50**	-.04	.22	.06	.379	.62
21	.32	.15	.04	**.46**	.438	.56
22	**.43**	-.09	.12	.10	.234	.77
23	.39	.11	.02	**.43**	.439	.56
24	**.64**	-.08	.11	.02	.440	.56
25	**.57**	.12	.24	.03	.572	.43
26	**.45**	-.02	.11	.01	.251	.75
27	.17	.29	-.03	**-.47**	.318	.68
28	**.37**	-.01	.26	.06	.287	.71
29	.30	.04	-.07	**.48**	.355	.64
30	.10	.13	-.01	**.51**	.311	.69
31	**.38**	.20	-.07	-.12	.214	.79
32	**.41**	.06	-.09	-.05	.159	.84
33	**.35**	.33	-.20	-.13	.263	.74
34	**.34**	.12	.02	-.04	.167	.83
35	.08	**.28**	-.04	-.02	.092	.91
36	.17	**.20**	.10	.01	.133	.87
37	**.14**	.07	.08	.05	.056	.94
39	**.34**	-.08	.04	.00	.110	.89
40	.10	.12	**.26**	-.03	.141	.86
41	**.33**	.00	-.14	-.06	.093	.91
42	.00	**.40**	-.17	-.01	.147	.85
45	**.38**	-.02	.10	-.05	.171	.83
46	.14	**.43**	-.04	-.01	.226	.77
47	-.11	**.53**	.09	.06	.291	.71
48	**.45**	.15	-.06	-.01	.249	.75
49	.13	**.42**	-.06	-.04	.215	.78
50	.11	**.16**	-.10	.02	.042	.96
51	.14	**.55**	.14	.06	.475	.52
52	**.33**	.27	.04	-.07	.260	.74
53	.13	**.53**	-.04	.03	.338	.66
54	**.35**	.12	-.12	-.05	.139	.86
55	.07	**.12**	.07	.10	.053	.95
56	.06	**.37**	-.01	-.01	.151	.85
57	.09	**.56**	.02	-.06	.364	.64
59	**.26**	.03	.13	-.08	.117	.88
60	.20	**.28**	-.12	-.19	.154	.85
62	.23	**.33**	.04	.00	.233	.77
63	.13	**.45**	.18	.02	.360	.64
64	-.06	**.57**	.06	.03	.335	.66
65	-.07	**.73**	-.07	.02	.485	.52
66	.00	**.44**	.02	.00	.204	.80
68	.02	**.59**	.25	.01	.507	.49
69	-.22	**.77**	.03	.06	.541	.46
70	.00	**.47**	-.05	-.06	.215	.79

Item	MR1	MR3	MR4	MR2	h2	u2
72	.20	**.21**	-.02	-.12	.116	.88
73	.09	**.12**	.08	-.06	.047	.95
75	.06	**.20**	.07	.19	.116	.88
76	.03	**.32**	.24	-.08	.222	.78
77	.12	**.31**	.14	-.04	.192	.81
78	.09	.16	**.20**	.00	.115	.89
79	**.32**	.05	.10	-.04	.153	.85
80	-.03	**.17**	.11	-.13	.060	.94
81	-.10	.12	**.29**	.01	.101	.90
82	.24	-.01	**.31**	-.01	.204	.80
83	**.23**	.01	.09	-.14	.086	.91
85	.06	.01	**.26**	.25	.165	.83
86	.10	.15	**.56**	-.06	.443	.56
87	.06	-.05	**.58**	-.07	.345	.65
88	.05	.02	**.65**	-.03	.455	.54
89	-.11	-.09	**.56**	.05	.274	.73
91	.04	.03	**.50**	-.01	.280	.72
92	.18	**.25**	.04	-.18	.151	.85
93	-.10	.13	**.48**	.06	.261	.74
94	.00	.06	**.54**	-.08	.306	.69
95	-.05	.06	**.48**	.04	.241	.76
96	-.02	-.02	**.48**	.06	.226	.77
97	**.25**	.11	-.06	-.19	.103	.90
98	.10	.02	**.38**	-.02	.192	.81
99	.07	.02	**.30**	.22	.183	.82

Note. The greatest factor loading of each item is shown in boldface.