***“*The National Center for Assessment in Higher Education*:***

***Its Role in Developing a Standardized Test for Arabic Proficiency in Speakers of other Languages(STAPSOL)”***

By

Dr. Amjed Al Owaidah Dr. Abdulrahman al-Shamrani

National Center for Assessment National Center for Assessment Department of measurement and Research Department of Language Testing

**a.owidha@qiyas.org****a.shamrani@qiyas.org**

**National Center for Assessment in Higher Education**

**Saudi Arabia**

**Abstract**:

 In response to requests from some academic institutions, the National Center for Assessment in Higher Education has developed a ***Standardized Test of Arabic Proficiency in Speakers of other Languages(STAPSOL)***. This test follows strict methodology in the way its items are written, reviewed, experimented, maintained and secured.

The test has four components: Verbal Comprehension (VC), Structure (ST), Listening Comprehension (LC), and Compositional Analysis (CA) as well as non-scorable trial items. A direct writing task is also included.

In order to meet internationally recognized standards, this test is made to insure reliability, validity and fairness to sex, region and level of study. In addition to achieving maximum quality assurance, the test items are written and reviewed by native speakers who are specialists in the fields of Arabic and applied linguistics and measurement. Its listening section is recorded exclusively by native speakers of Arabic. ***STAPSOL*** is also in the process of being formally linked to the Common European Framework for Languages (CEFR).

A sample of 1621enrolling in different institution in Jakarta Indonesia took the tests. However, the findings indicated that the test is reliable. Initial construct validity evidence support the validity of the test. However, More studies of the test validity need to be collected.

**Historical Background:**

In order to put the development of the Standardized Test of Arabic Proficiency for Speakers of Other Languages *STAPSOL* in its historical perspective, it is appropriate to give a brief introduction about the foundation and growth of the National Center for Assessment in Higher Education (*NCA).* The Saudi National Education Policy, written in 1969, stipulates in two of its articles the necessity of establishing independent specialized bodies that develop tools to help select and nourish able students for the admission to higher education institutions( Articles 212 and213). It was only in 2001 that the *NCA* was established as an institution solely specializing in educational measurement. A recommendation from the Council of Higher Education for the creation of this body was made then followed and supported by a decree of the Council of Ministers. This decree spells out the nature of this institution as being national, educational and non-profit. By the end of 2002 the first tool was developed by *NCA*; namely, the General Ability Test (*GAT)* which was a large scale test that became an admission requirement for any student applying for higher education. Naturally the decision to make it national and an admission requirement happened after all relevant experimentation and validity studies had been applied to it. Not too long after that the *NCA* was tasked with the development of another test, an achievement test, that became a national admission requirement as well. By the year 2008 the *NCA* had developed several other tests in different fields for different beneficiaries; considerable recognition locally and regionally was gained. It was then that some interested institutions, especially universities, approached the *NCA* and requested a test for Arabic Proficiency that could be used for academic or placement purposes*.*

**Committee Formation:**

An eight-member committee was formed to look into theuniversities’ request. The members came from different but complementary backgrounds. Three were educational measurement specialists, five were specialists in applied linguistics, language testing and the teaching of Arabic language for speakers of other languages; all members were university professors. After extensive survey, the committee reached a decision that there was a need for a large scale test that measures Arabic proficiency in speakers of other languages. There followed a thorough nine-month review which led the committee to a definition of the test objective, its components and their weights. The test is standardized and is of the same high quality that the *NCA* has developed for its entire range of tests.

**Test Objective:**

To measure the level of Arabic language proficiency in speakers of other languages that enables them to enroll in university programs which use Arabic as a medium of instruction.

**Test Components and their Weights:**

As is stated in the historical background and test objective, this test leans heavily towards the use of Arabic in an academic setting. Therefore, more weight is given to language skills that are estimated to be crucial to scholastic purposes as is reflected in the distribution below.

A given test form comprises 100 scorable and 35 non-scorable items covering the three components of the test; i. e.

1. *Reading Comprehension* which consists of 45 items
2. *Writing* which consists of 35 items
3. *Listening Comprehension* which consists of 20 items

|  |  |
| --- | --- |
| *Reading Comprehension* *Writing* *Listening Comprehension*   | 45%35%20% |

A given test form comprises also an additional 35 non-scorable items for try-out.

Items are divided among six sections to allow for internal and across section rotations.

*Reading Comprehension:* items are usually based on passages of different lengths. There are short passages (perhaps a two or three sentence paragraph), medium size passages of 2-3 paragraphs and longer passages of about 300 words each. These passages cover a wide range of non-technical topics dealing with science, the arts, social customs and/or traditions, the environment and the like. Passages that are longer than one paragraph are usually numbered and so that questions can refer to these numbers for convenience. Items are meant to test abilities like finding information, logical relations, reasoning, inference, referentials, titles and sub-titles…etc. Normally questions follow paragraph ordering; e.g. questions relevant to the 1st paragraph come in order before questions relevant to the 2nd paragraph. One consequence of this is that the usual ordering of items by difficulty gives way to ordering the questions sequentially according to the flow of information in the passage itself.

 *Writing:* items are of two types:

1. Multiple choice items in which the focus is on the actual usage of the language, not on discrete grammatical points. In other words, the correctness of the usage is the focal target. Therefore, such items focus on word order, permutation, structural agreement forms, correctness of adjectival, adverbial, attribution, predication, noun-constructs, pronominal, prepositional and numeral constructions. Most questions of this type measure understanding and application abilities.
2. A free writing task in which the test-taker is given a question that requires the writing of an argumentative essay in no less than 200 words. Normally the test-taker is asked to express his/her opinion for or against an issue and logically support whatever side is taken.

One thing that might pose a challenge to *NCA* regarding the free writing task is how to score it. Rating can be done by machine or by humans; both mechanisms are used now in the field. For Arabic, machine rating is still in the preliminary stages and much progress must be made before that is fully incorporated by *NCA*. The only option for *NCA*, therefore, is to take the human rating route. Since that is the case, two four-day workshops were planned. In preparation, *NCA* surveyed the standards in use for essay rating and crafted a draft of what is thought to be acceptable standards. This draft proposed 7 dimensions and invited 12 university professors for the first workshop. These dimensions were thoroughly scrutinized and revised using actual free writing responses. The workshop participants were then given about 1000 answer sheets that were distributed equally. Every batch was independently rated by 2 persons. This allowed for calibration of the performance of each of the 2 raters. The results showed that there was need to do more revisions because the gap between every two raters and among the total number of raters was wider than expected. The range was between .60 to .70.

The second workshop was convened with 12 new raters. More revisions were made and the shortcomings of the first workshop were fixed. Twenty free writing response sheets were given for try-out rating. An analysis of this batch showed a greater convergence in results. The range now is well above .90. The rest response sheets (about 1300) were equally distributed among the raters.

*Listening Comprehension:* Listening itemsare based on monologues (announcements, advertisements and parts of lectures or news readings) or dialogues of different lengths. The items are intended to measure abilities like long term retention of information, finding information relevant to time and place, directions and so on. In the listening section there are usually fewer questions dealing with higher order abilities than in the reading section, for example.

The recordings were made in state-of-the-art professional labs by native speakers (both men and women) who are acknowledged for the clarity, precision and accuracy of their pronunciation.

**Item Writing:**

Prospective item writers for the STAPSOL are carefully selected and then invited to participate in a five-day workshop, in which they are well familiarized with standardized testing. Basic principles like the clarity of test objectives, strategies and characteristics of writing multiple choice questions, item difficulty, content, bias, analysis, test validity and reliability are all introduced. This is followed by hands-on experience in item writing where every participant writes items, reviews and discusses them with colleagues under supervision of test specialists. The *NCA* then asks the workshop participants to write questions for the various test components according to the assigned requirements and objectives of the test. When items are sent to *NCA*, each item is assigned a special code number that remains with it in the item bank. Likewise, each item writer is also given a code number in order to insure anonymity and avoid any kind of personal bias. With anonymity of the item writer assured, each item is reviewed by a committee which consists of three members: (1) a field specialist, (2) a specialist in measurement and (3) a neutral person who may be an advocate for the test-taker or the society at large. The Reviewing Committee scrutinizes each item for: content, domain, bias towards sex or region, difficulty, creativity, suitability of each stem and/or choices. The items are then entered into an item bank. When an item is selected from the item bank for use in a test, it is reviewed for a second time to insure accuracy and adherence to the test requirements. Test forms are made and tried out, including the conciseness and clarify of the instructions to the test takers. Test items are largely in the multiple choice format. After try-out, test forms are electronically read and scored. The characteristics of each and every item, i.e. difficulty, discrimination, bias and representation of domain are then definitively known.

**Language:**

The Arabic language has a long history and literary traditions. Like all living languages it has undergone changes due to social, areal and scientific factors. Through its history Arabic has had varieties which still continue until present day. Arabic dialects are used and spoken today in all Arab countries and regions. However, there is one Arabic variety that may be considered the *lingua franca* ofall educated Arabs. This variety is mostly used in formal functions like meetings, speeches, sermons, lectures and media. It is the medium of instruction in all educational levels. Internationally, it is the sixth language used in the United Nations. This variety is given different names in English such as Modern Standard Arabic, Modern Literary Arabic, Modern written Arabic, Modern Literary Written Arabic and Arabic Language Today. Regardless of what name is given, this variety is the one to be used in this test. One benefit the test-taker is going to make out of this is that if he/she is proficient in this variety, he/she will spend much less time picking up dialectal variations.

**Theoretical Framework:**

Two main frames of reference were surveyed; namely, (Common European Framework of Reference for Languages, Council of Europe, 2004) and (ACTFL Proficiency Guidelines, American Council on the Teaching of Foreign Languages, 1986). After profound comparison was made, the academic committee selected the CEFR as its theoretical paradigm because of (1) the richness of experience it enjoys in its field, (2) the frequent maintenance it undergoes, (3) its transparency to different linguistic backgrounds and (4) its broadness and neutrality since it transcends linguistic variations. The CEFR posits six levels of proficiency, ranging fromA1 at the lowest level to C2 at the highest level.

However, The purpose of the current study is to have more evidence of its reliability and validity. It intends to answer the following questions:

1. What is the reliability of the STAPSOL?
2. Does the factor structure of the test support the allocation of its items in the different subtests?
3. Does the test differentiate different levels of proficiency?
4. Does the item-person fit analysis conform to the known criteria in the scientific arena?

**Methodology**

**Samples:**

The try-out sample was drawn from Arabic Language institutes in four universities in Saudi Arabia, namely King Saud University, Imam University, and Umm al-Quraa University. The total sample was 952 students enrolling in middle to upper levels of Arabic language classes.

The current sample was 1621 students enrolling in advance courses of Islamic and Arabic studies in four Indonesian universities and an academic institute of Islamic and Arabic studies (affiliated with Imam University located in Jakarta).

Two sample tests ( 7050a and 7050b) were formed from the item pool, each containing 100 items, for trial. Four Saudi universities participated in the trial process. Each of the four has an Arabic language institute whose graduates are expected to compete for admission to university academic programs where courses are taught in Arabic. More than 950 students participated in the experimentation. These students represent more than 50 nationalities and range in their Arabic programs from medium level to advanced level.

**Instrumentation:**

Two tests , (*7051* and *7052)* were developed from the tried items in the sample tests. For validity purposes, these two tests were administered in four Indonesian universities. More than 1260 students (males and females) took the tests. The large majority of these students have already finished their language programs and are now attending college using Arabic as a medium of instruction. The levels of the rest of students range in their Arabic programs from medium to advanced.

**Statistical Analysis:**

Rasch Measurement Model as an approach to Construct Validity: Rasch model is a mathematical formula that specifies the form of relationship between persons and the items that operationalize one attribute. That means the probability of obtaining higher scores should be associated with the persons with higher ability on that specified and targeted attribute (Green,2002). A key feature of the Rasch model is that of unidimensionality or the requirement that a set of test items measure a single construct. Hence, if the observed test data fit the Rasch model expectations, then the test data are measuring only one construct and that would imply the test data have an evidence of internal construct validity.

 Given the fact stated above, the research questions stated in this study were answered using Rasch analysis. Specifically, the overall fit statistics, person and item separation, STRATA property, and the principle components factorial analysis of the residuals were all used to investigate the internal construct validity of STAPSOL.

Results

 Before interpreting the results of Rasch analysis, it is important to provide an overview of the overall fit statistics of STAPSOL. First, the mean infit and outfit for person and items should be 1 or close. As shown in table1 the data of STAPSOL indicate that the mean infit and outfit for both persons and items was .99. the standardized mean infit and outfit are expected to be 0 or close. For persons the mean infit and outfit was 0. For items the mean infit and outfit were 0 and .1 respectively. This suggests that the data in hand fit the Rasch model very well. The standard deviation of the standardized infit used to assess the overall misfit. Using 2 as a cutoff point ( green, 2002). The data indicate that the standard deviation for person was 1.1, and for items 1.2 . these values are less than the cut-off suggested by green(2002) of 2.0. This, in turn, provides evidence that the data fit the measurement model used (see appendix A).

 Other overall fit statistics is the person and item separation. This statistics is a measure of the spread of the persons and items position. Separation index more than 1 suggests that the data has sufficient breadth in terms of its position (green, ? ). The STAPSOL data show that the person separation index was 3.30.This result suggests that the persons measure has an adequate breadth position on the linear continuum from less proficient to more proficient in Arabic. However, item separation index was 12.87 which suggests that the STAPSOL items were adequately dispersed (Appendix A).

 Moreover, the 4-STRATA was used. According to Schumacher (2004), The Strata is a property that assists in understanding how well person measures dispersed along the latent trait of STAPSOL because it shows the number of distinct ability level separated by 3 errors of measurement, thus, it is very useful in determining group differences. Using the Strata formulae as follows:

STRATA = (4Gp + 1)/3 = (4\*12.87+1)/3 = 17.49

 means that the STAPSOL developers (NCA) should be able to distinguish 17.49 different groups of non-Arabic speakers based on the STAPSOL Scale. Given the results of both persons and item separation indices and the Strata property provide an answer to the research question of this study that the STAPSOL scale does in fact differentiate between different levels of proficiency.

 To judge whether or not the STAPSOL is adequately reliable, the person separation reliability index was used. This index is analogous to the classical Cronbach's alpha. This index should fall between 0 and 1. An index of 1 means perfect reliability and 0 means random data. Table 1 indicates that the person separation reliability is .92. this result suggests that the respondents of STAPSOL were being separated by the items along the linear continuum of the construct targeted. Also, it provides evidence of the replicability for person placement using different items measuring the same trait.

 Another advantage of using Rasch analysis is the item-person variable map. This map provides the STAPSOL developers with an overall picture about the position and the distribution of the persons to STAPSOL items. Appendix B explains the result of this map. The left side represents the examinees on interval logit scale continuum and the upper side represents the more proficient examinees in Arabic whereas the lower side indicates the less proficient examinees. The right side of the map; however, represents the STAPSOL items. Difficult items are located at the top and more easy items are located at the bottom. The “M” letter indicates the mean of the distributions for both items and persons whereas the “SD” represents one standard deviation of both items and persons. “T” is two standard deviation units. The mean of the items distribution is always an arbitrary set to 0. Examining the item – person map suggests that the ability of persons distributions are higher than the difficulty of item distributions by 1 standard deviation and is negatively skewed. This result implies that a future version of the STAPSOL has to include harder items. Furthermore, the map indicates that the distribution of both persons and items has a wide range that roughly goes from + 2 and – 2 standard deviation. This result suggests that there is an adequate match between the STAPSOL items and persons taking the test. The no gaps on the items linear continuum explains the wide and sufficient coverage of the test items which supports the previous result (appendix B).

 Finally, as an attempt to find evidence for the internal construct validity, principle component analysis of residuals was used. The Rasch factor analysis of residuals has the same idea as the factor analysis which expects that the Rasch factor analysis does factor analysis on the residuals leftover after fitting the data to the Rasch model. The factor analysis of these residuals is then used to identify if there is a common variance shared among the unexplained data by the Rasch model. If a dominant variance was found among items not explained by the Rasch model then a second dimension other than the intended dimension is interfering. This will affect the internal construct validity of the measure. Those items have to be removed from the scale to form a new subscale. However, in order to test for the dimensionality of the STAPSOL Winsteps program (2009) version 3.69.1.10 was used to calibrate the STAPSOL data. Results in appendix C indicate that the variance explained by the Rasch measurement is 31.2 % whereas the unexplained variance by is 68.8%. More importantly, the unexplained variance by the first factor was 2.8 %. In item unit strengthen, it was 3.7. Is that factor good enough to form another dimension or is it only due to some randomness in the test data? Linacre (1998) suggested that 2 item unit strengthen could be due to randomness in the data. In the case of STAPSOL data the strength was 3.7 which is more than the suggested criteria. The factor loading in appendix C suggests that there were two items having positive loading close to .4 or above and they were item 10 , and item 14. On the negative loading side there were also two items close to .4 or above. Those items are item 55 and item 77. A rule of thumb is that any item loading close to .4 and above should be investigated (Bond & Fox, 2007). As shown in appendix C these positive and negative loading are those common variance found in the first factor and it is worth only 3.7. In light of the present finding, further investigation of unidimensionality was done through cross plot person measures on both 4 items stated above. Graphically speaking, the person measure on these items has to be within certain confidence interval range. Therefore, if they fall outside the specified boundary then they are telling a different story, hence, confirming multidimensionality problem. Figure 1 shows that there is no data point falling outside the specified boundary, thus supporting the unidimensionality of STAPSOL. However, the noise items were, in fact, only random errors in the data which in turn provide an evidence that the STAPSOL has a sufficient internal construct validity.

Figure 1: Cross Plot person measures

**References**

Bond, T. G., & Fox, C. M.(2007). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Green, K. E., & Frantom. C. G. (2002). Survey development and validation with the Rasch model. A paper presented at the International Conference on Questionnaire Development .

Linacre . J. M. (2009) Winsteps Rasch Measurement. Version : 3.69.1.10

Schumacher, R. E. (2004). Rasch measurement : The dichotomous model. In . Smith, E, V. & Smith, R. M.  *introduction to Rasch measurement*.