Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

AWEJ Volume.5 Number.4, 2014

Pp.203-225

Fair, Reliable, Valid: Developing a Grammar Test Utilizing the Four Building Blocks

Voke Efeotor
Testing Unit
Taibah University, Madinah, Saudi Arabia

Abstract

The purpose of this study is to develop and analyse a grammar test, focusing on validity, reliability and fairness. Effective test development requires a systematic, well-organized approach, thus ensuring the veracity of the proposed inferences from the test scores, (Downing, 2006a). This systematic approach is outlined in the paper, expounding on how the test developed from the initial stage of identifying the construct to be measured, to the final stages of administering the test. This is followed by a detailed analysis of the test results which highlights possible threats to the test's validity, reliability and fairness. It also places the construct map under scrutiny, and questions whether the construct was successfully tested. The test was developed utilizing the four building blocks propounded by Wilson (2005). Each building block is explained in detail which gives insight into the process undertaken to write and analyse the test.

Keywords: validity, reliability, fairness, testing grammar, construct map, building blocks

Introduction

Whether formative or summative, high-stakes or low-stakes, assessment has become an integral part of education. Around the world, millions of students take exams each year. In many countries, success on a language proficiency test has become a prerequisite for foreign students to be admitted to a university. However, little consideration is given to the intricate process of test writing, or the numerous factors that test writers contemplate. In many educational institutions, different versions of the same test are administered. In order to ensure justice for the students, the tests, although different in content, have to be equal in difficulty. Moreover, test writers have to create assessments that minimise factors that could cause a misinterpretation of results. This paper aims to shed some light on the test writing process by expounding on one of the methods used for test writing and analysis. The four building blocks are outlined in turn, the first of which is the construct map.

The Construct Map

The test, which is the measurement instrument, had a distinct purpose. As Wilson (2005) purports, “An instrument is always something secondary: There is always a purpose for which an instrument is needed and a context in which it is going to be used” (p.6). The aim of the test was to evaluate grammar competency amongst English language learners. Thus grammar competency is the construct which is the theoretical object of interest in the participants. As Wilson (2005) states, “we assume that the construct we wish to measure has a particularly simple form- it extends from one extreme to another” (p.6). Hence, it was assumed that some of the participants would be weak in grammar and others would be stronger. In order to gauge their varying competency, it was imperative to formulate a construct map which would be a graphical representation of how the construct developed. In test development, the construct map is “refined through several processes as the instrument is developed” (Wilson, 2005, p.6). However, the purpose of this paper was to trial a test, consequently any changes to the construct map, would only be made if the test were to be developed further.

The Common European Framework (CEFR), initially published in 2001, was central to forming the construct map. The CEFR is a guideline used to validate language proficiency. The framework consists of six reference levels, which have widely been adopted in Europe as a yardstick for grading an individual’s language proficiency. Figure 1 shows the different levels of the CEFR and a description of the skills required to be categorised in each band.

Figure 1. ***The six levels of the Common European Framework***

Proficient User	**C2**	**Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.**
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.

Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Common language proficiency tests, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), use the CEFR to equate scores obtained on their tests to language proficiency. The CEFR also contains scales for the different language skills such as reading and speaking. However, there is no scale for grammar. The task of aligning the CEFR with grammar was carried out by the British Council. Their work was based on research which suggests there is an order in which grammar is acquired.

Naturalistic theories propound that there is a natural order and sequence of acquisition. Corder (1967) suggested that language learners have an inbuilt syllabus for learning grammar. The Natural Order Hypothesis claims that “the acquisition of grammatical structures proceeds in a predictable order” (Mitchell & Myles, 1998, p.12). It is upon this assumption that the construct map was made. It purports that there are items of grammar that learners at A1 level will know, and others that they will not know until much later in the learning process, as they progress along the continuum. As Larsen-Freeman (2001) states, “there has been no definitive acquisition order established, and thus teachers are still left to their own resources for judgments on how to proceed” (p.263).Thus, the British Council and the European Association for Quality Language Services (EQUALS) embarked upon a project (the Core Inventory) intending to “make the CEFR accessible to teachers” and “answer the question put by many teachers over the years of what to teach at each CEFR level” (“Core Inventory for General English”, 2010).

The British Council and EQUALS developed the Core Inventory through iterative and collaborative processes, working with partner organizations as well as with examination boards. They also drew from many sources. These included; an analysis of the language that is implied by the CEFR band descriptors, an analysis of syllabuses of EQUALS members who implemented the CEFR, a content analysis of popular textbooks, and teacher surveys. They analysed the data to find consensus points "which were common to a strong majority (80%) in each of the data sources" ("Core Inventory for General English", 2010). Examination boards (Cambridge, ESOL, City and Guilds, Trinity) provided further assistance by commenting on language points they considered to be relevant to each band. Figure 2 shows the Core Inventory and Figure 3 shows the construct map.

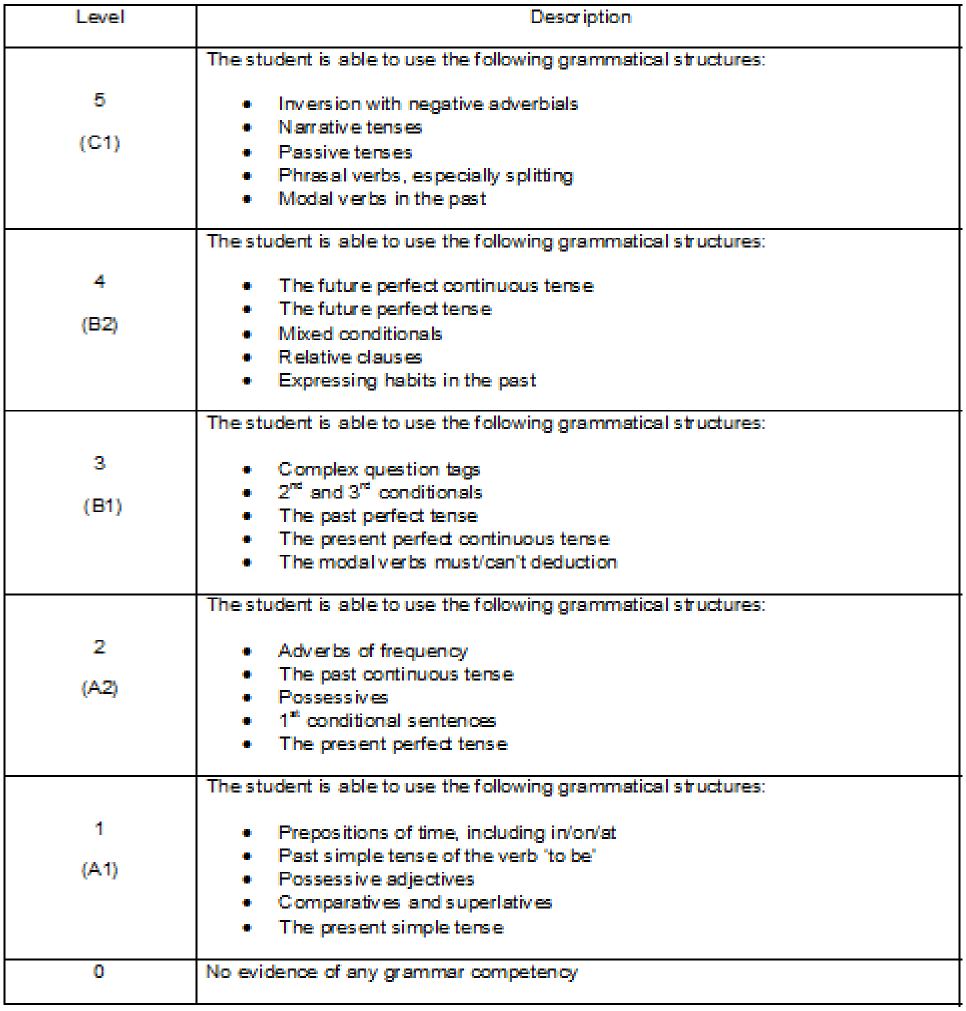
Figure 2.*The Core Inventory compiled by the British Council and EQUALS, Core Inventory for General English, 2010*

	A1	A2	B1
Grammar	Adjectives: common and demonstrative Adverbs of frequency Comparatives and superlatives Going to How much/how many and very common uncountable nouns I'd like Imperatives (+/-) Intensifiers - very basic Modals: can/can't/could/couldn't Past simple of "to be" Past Simple Possessive adjectives Possessive s Prepositions, common Prepositions of place Prepositions of time, including in/on/at Present continuous Present simple Pronouns: simple, personal Questions There is/are To be, including question+negatives Verb + ing: like/hate/love	Adjectives – comparative, – use of than and definite article Adjectives – superlative – use of definite article Adverbial phrases of time, place and frequency Adverbs of frequency Articles – with countable and uncountable nouns Countables and Uncountables: much/many Future Time (will and going to) Gerunds Going to Imperatives Modals: can/ have to/should Past continuous Past simple Phrasal verbs – common Possessives – use of 's, s' Prepositional phrases (place, time and movement) Prepositions of time: on/in/at Present continuous Present continuous for future Present perfect Questions Verb + ing/infinitive: like/ want-would like Wh-questions in past Zero and 1st conditional	Adverbs Broader range of intensifiers such as too, enough Comparatives and superlatives Complex question tags Conditionals, 2nd and 3rd Connecting words expressing cause and effect, contrast etc. Future continuous Modals - must/can't deduction Modals – might, may, will, probably Modals – should have/might have/etc Modals: must/have to Past continuous Past perfect Past simple Past tense responses Phrasal verbs, extended Present perfect continuous Present perfect/past simple Reported speech (range of tenses) Simple passive Wh- questions in the past Will and going to, for prediction

	B2	C1

Grammar	Adjectives and adverbs Future continuous Future perfect Future perfect continuous Mixed conditionals Modals – can't have, needn't have Modals of deduction and speculation Narrative tenses Passives Past perfect Past perfect continuous Phrasal verbs, extended Relative clauses Reported speech Will and going to, for prediction Wish Would expressing habits, in the past	Futures (revision) Inversion with negative adverbials Mixed conditionals in past, present and future Modals in the past Narrative tenses for experience, incl. passive Passive forms, all Phrasal verbs, especially splitting Wish/if only regrets

Figure 3.*The Construct Map*



Level	Description
5 (C1)	The student is able to use the following grammatical structures: • Inversion with negative adverbials • Narrative tenses • Passive tenses • Phrasal verbs, especially splitting • Modal verbs in the past
4 (B2)	The student is able to use the following grammatical structures: • The future perfect continuous tense • The future perfect tense • Mixed conditionals • Relative clauses • Expressing habits in the past
3 (B1)	The student is able to use the following grammatical structures: • Complex question tags • 2nd and 3rd conditionals • The past perfect tense • The present perfect continuous tense • The modal verbs must/can't deduction
2 (A2)	The student is able to use the following grammatical structures: • Adverbs of frequency • The past continuous tense • Possessives • 1st conditional sentences • The present perfect tense
1 (A1)	The student is able to use the following grammatical structures: • Prepositions of time, including in/on/at • Past simple tense of the verb 'to be' • Possessive adjectives • Comparatives and superlatives • The present simple tense
0	No evidence of any grammar competency

The Item Design

The next step in the test development process is to find a way "to stimulate responses that can constitute observations about the construct that the measurer wishes to measure" (Wilson, p.41). The aim is to operationalise dimensions of the construct into items that give an accurate representation of the ability of the participant. In order for the test to be reliable it must give an accurate indication of the ability of the student. Test reliability is discussed in detail later in the paper. However, it is worth noting that there were several decisions that had to be made with the intention of producing a reliable test. The first of these was the test item format that would be used. As Downing (2006a) purports, "The creation and production of effective test questions, designed to measure important content at an appropriate cognitive level, is one of the greatest challenges for test developers" (p.10). There are many different formats that a test can adopt. Perhaps the two most common type of item are the open-ended item format and the fixed-response format (Wilson, 2005). The choice of which format to use depends largely on the construct. It is pertinent to choose the format that will enable operationalisation of the construct.
There are various advantages and disadvantages of each format. The open-ended item format has been praised (RePass,1971; Iyengar, 1996) as it enables the examinee to express his thoughts, without giving any prompts or cues. Moreover, this format allows for detailed responses to complex issues, as well as revealing the examinees logic and reasoning. However, open-ended items do have some shortcomings. Firstly, they may not be responded to due to ineloquence rather than indifference; examinees may not respond to open-ended questions because they are unable to express their thoughts. Furthermore, open-ended items can be difficult to grade, while necessitating a certain degree of subjectivity.

The fixed-response format is often used as multiple-choice items or a Likert-type response scale. In contrast to the open-ended format, the fixed-response format is easier to grade, as possible responses are limited. The objective nature of the test limits scoring bias. In addition, students can potentially respond to many items which can permit wide sampling. When conducting a study and asking for voluntary participation, requesting long open-ended responses can deter would-be participants, whereas they may be more willing to answer fixed-response items.

There are also disadvantages of the fixed-response format. Firstly, there is the criticism that multiple-choice tests teach misinformation (Toppino & Brochin, 1989; Rees 1986), can suggest the answer to the examinee, and allow for guessing. The use of item-person matrixes can however highlight possible occurrences of guessing. Another shortcoming of fixed-response items is that it is difficult to measure certain learning outcomes such as displaying thought processes and articulating explanations. The progression of some constructs may be demonstrated by higher-order thinking skills, such as those found in Bloom's Taxonomy. It is a lot harder to demonstrate analysis, synthesis and evaluation with fixed-response items. Despite these shortcomings, the fixed response format was favourable for the purpose of testing grammar. This is due to the nature of grammar itself. It would be very difficult for an open-ended item to specify to the examinee what grammatical structure was required of them. One possible way would be to rely on the use of grammatical terminology. For example, one may ask an examinee to answer a question using the present perfect continuous tense. The problem with this approach is that it would not be testing the construct, solely, but also their knowledge of grammatical terminology. Hence, an examinee may get an item wrong, not because they are unaware of the grammatical structure, but because they are unaware of the terminology. This would have a damaging effect on the reliability of the test. Moreover, it is likely that in an open-ended item, the target grammatical structure would be used in the stem itself, and easily duplicated, which would also impact upon the reliability of the test.

The fixed-response format was advantageous for this test. However, much rests on the writing of good items. “One of the greatest limitations of the selected-response formats derives from the creation of flawed selected-response items” (Downing, 2006b, p.290). As a result, certain good codes of practice were adhered to when writing the items. Firstly, each item had four possible options. Rodriguez (2005) holds that three options are sufficient and that “using more options does little to improve item and test score statistics and typically results in implausible distracters” (p.11). However, having four options does lower the probability of randomly guessing the correct answer. Moreover, Lord (1977) demonstrated that when reliability is estimated using the Spearman-Brown prophecy formula, it increases when the number of options increases. Having five options would obviously lower the probability of guessing further, however, as more options are added it is more likely that the options would not be statistically functional (Downing, 2006b), as a result four options were opted for. Secondly, the options were homogenous in content. Hence, if the target item was an object pronoun, the three distracters were also object pronouns. The options were also similar in length in order to avoid giving a hint to testwise examinees. Thirdly, extra care was taken to ensure that the items were independent of one another. Thus, the answer to an item could not be found in the stem of another item. Finally, the items did not contain unnecessary or high-level language which could have introduced construct-irrelevant variance into the measurement. This would have affected the fairness of the test. “If some students have an advantage because of factors unrelated to what is being assessed, then the assessment is not fair” (McMillan, 2001, p.55). As Abedi (2006, p.383) states, “complex language in the content-based assessment for non-native speakers of English may reduce the validity of inferences drawn about students’ content-based knowledge.” While every effort was made to write good test items it is possible that some of the items were flawed. A discussion of the items will follow in the results section. The full test is shown in Appendix 1.

The Outcome Space

The next building block, the outcome space, is concerned with scoring the responses of the examinees. Scoring the test correctly has great implications for the validity of the test. If the test scores are to be valid a scoring key must be accurately applied to mirror the examinee item responses (Wilson, 2005). The term outcome space was introduced by Marlon (1981) to describe a set of categories which are devised to grade examinee responses. As Wilson (2005) purports, “inherent in the idea of categorization is an understanding that the categories that define the outcome space are qualitatively distinct” (p.63). This is even the case for fixed-response items, such as multiple tests and Likert-style survey questions. With open-ended items it is pertinent to develop an outcome space which provides example item responses belonging to the different categories. This requires researching the variety of responses that examinees could give. However, in fixed-response items such as multiple choice questions and true-false survey questions there are traditionally two categories; one category for choosing the correct answer and another category for choosing the wrong answer. However, this is not the only possible way to grade multiple choice items.

Multiple choice item responses may also be categorised, where an examinee scores a point for choosing a distracter which is considered a better response than the other distracters. This is highlighted in the question in figure 4.

Figure 4. ***A multiple choice item that could use polytomous scoring.***

Q. Who was the 41st President of the United States of America?

a) George Bush b) Abraham Lincoln c) Ronald Reagan d) Tony Blair

The multiple choice item in Figure 4 could be scored using two categories. If the examinee chooses the correct answer, (a), they score one point, if they choose any of the distracters they score zero. Alternatively, the distracters could be graded in order, to show partial success. As Wilson (2005) states, this is done when the difference between the distracters is large enough, and when "there is a way to interpret those differences with respect to the construct definition" (p.70). Subsequently, the outcome space may award points for each distracter accordingly; (a=3, c=2, b=1, d=0).

Another possible scoring scheme is to award negative marks for an incorrect answer. A disadvantage of this is that it causes cautious students to refrain from answering items, even if they possibly know the correct answer. The analysis focuses on the items themselves, rather than the overall score. Therefore negative marking could have had the effect of leaving certain items with very little statistical data available. Furthermore, it was believed that the measurement model would highlight possible instances of examinee guessing.

The traditional method of scoring multiple choice items was chosen. Negative marking was rejected for the reasons outlined above. Polytomous scoring was rejected because there are no significant differences between the distracters which could be interpreted with respect to the construct. In many of the items the options are homogenous. Figure 5 is taken from the test paper.

Figure 5. ***An item from the test.***

Q. Ali is tall and ___________ hair is black.

a) our b) your c) her d) his

The answer to the item in Figure 5 is a possessive pronoun, as are the other three distracters. Therefore, due to this homogeny there is no 'better' answer amongst the distracters. As a result, the most valid scoring scheme for the test was adopted, which was to give one point for a correct answer and zero for an incorrect answer.

The Measurement Model

The final building block, the measurement model aims to "relate the scored outcomes from the items design and the outcome space back to the construct that was the original inspiration of the items" (Wilson, 2005, p.85). Two main approaches to measurement have emerged, the first one has been coined Classical Test Theory (CTT). The theory purports an examinee's observed score

(X) is a combination of their true score (T) in addition to some error (E) that is made in measurement.

X= T + E

CTT deals with the relationship between these three variables in the population. CTT has a number of limitations and shortcomings. Firstly, item difficulty and item discrimination are group dependent. As a result the *p* and *r* values "are entirely dependent on the examinee sample from which they are obtained" (Hambleton & Jones, 1993, p.43). Secondly, scores obtained using CTT applications are test dependent. The true-score model which formed the foundation for CTT "permits no consideration of examinee responses to any specific item" (Hambleton & Jones, 1993, p.43). CTT is viewed as test oriented rather than item oriented. Subsequently, the theory cannot aid in predicting how well an examinee may do on a test item. Finally, as Hambleton et al. purport (1991, p.4), the assumption of equal errors of measurement is implausible since test scores are unequally precise measures for examinees with different abilities.

The second main approach to measurement is item response theory (IRT), which is also referred to as Modern Test Theory. IRT is a "general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test" (Hambleton & Jones, 1993, p.40). Here the focus is on item-level information rather than test-level information. IRT "is a measurement approach that relates the probability of a particular response on an item to overall examinee ability" (Guler, Uyanik & Teker, 2014, p.2). Subsequently IRT overcomes some of the shortcomings of CTT, namely that "IRT ability parameters estimated are not test dependent and item statistics estimated are not group dependent" (Guler et al., 2014, p.2). There are further advantages of IRT. In order to assess reliability, one does not need to conduct strict parallel tests. Furthermore, examinee ability and item statistics are highlighted on the same scale. The Rasch Model is a special case of IRT. However, there are key features of the Rasch Model which distinguish it from the item response modeling tradition. The Rasch Model fits the data to the model rather than fitting the model to the data. The Rasch Model seeks to highlight the items which are bias and for whom, the items which define the trait to be measured, and the examinees which are properly measured by the items (Wright, Mead & Draba, 1976). The Model purports that a person's ability and the item difficulty are central to a person's measure on a trait. For this paper the Rasch Model was used as the measurement model as the primary focus was on each item as it related to the examinee on the continuum of the construct map. Moreover, it was paramount to observe whether any of the items over or under discriminated relatively to the discrimination of all the items.

Sampling

Thirty male examinees undertook the test. In order to eliminate selection bias, and strengthen the internal validity of the study, the examinees were randomly selected from mixed-ability classes in a boys' school in Jeddah, Saudi Arabia. All of the participants were studying English as a foreign language, and were either aged 15 or 17. 'Random sampling is free of the systematic bias that might stem from choices, made by the researcher or others' Gorard (2013, p.79).

Item and Test Analysis

The test analysis focuses on validity, reliability and fairness, which are three important facets of any meaningful test. “The primary measurement standards that must be met to legitimize a proposed test use are those of reliability, validity, and fairness, which are also value-laden concepts” (Messick, 2000, p.3).

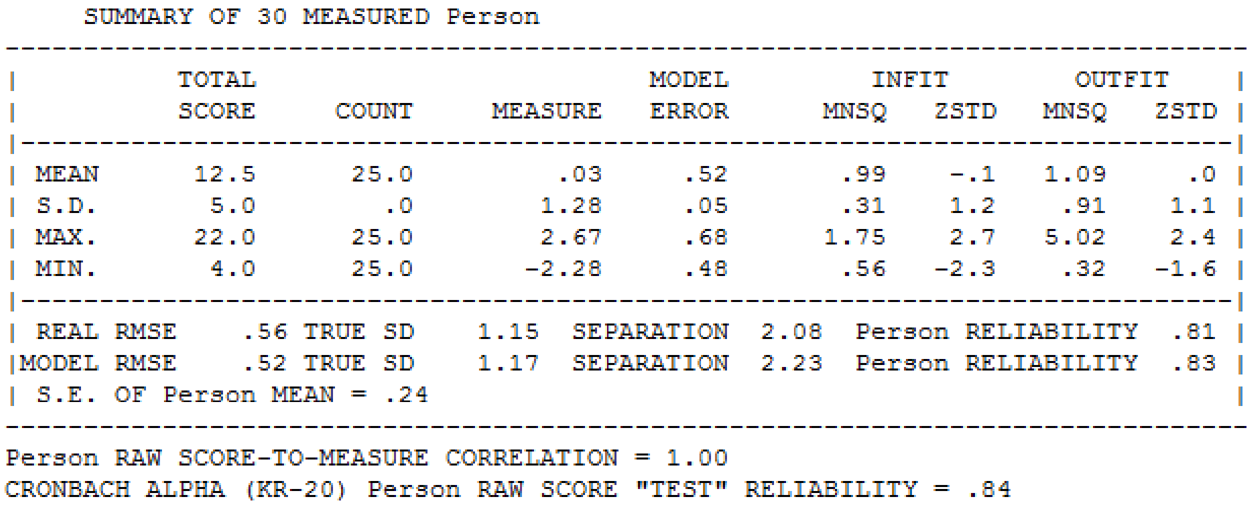
Reliability

Test reliability “refers to the consistency of scores students would receive on alternate forms of the same test” (Wells & Wollack, 2003, p.2). Test reliability is important because it ensures that test scores show more than random error. Moreover, test reliability is a prerequisite for test validity. A test cannot be valid, if it is not reliable. Subsequently, it is right to analyse the reliability of the test first. After all, “If the test is unreliable, one needn’t spend the time investigating whether it is valid–it will not be” (Wells & Wollack, 2003, p.3). There are different ways to ascertain the reliability of a test using Rasch theory. WINSTEPS Rasch Software calculates person sample reliability which is akin to the test reliability of CTT. Moreover, it calculates item reliability, which is a desired feature of this analysis which CTT does not report. Cronbach’s alpha is the most commonly used index to gauge the internal consistency of a test, which utilizes the following formula:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\hat{\sigma}_X^2}\right)$$

Cronbach’s alpha ranges from 0 to 1.00. Values close to 1.00 indicate high consistency. Values above 0.7 are good for low-stakes testing, while values less that 0.5 are unacceptable (George and Mallery, 2003). Table 1 shows the Cronbach Alpha value.

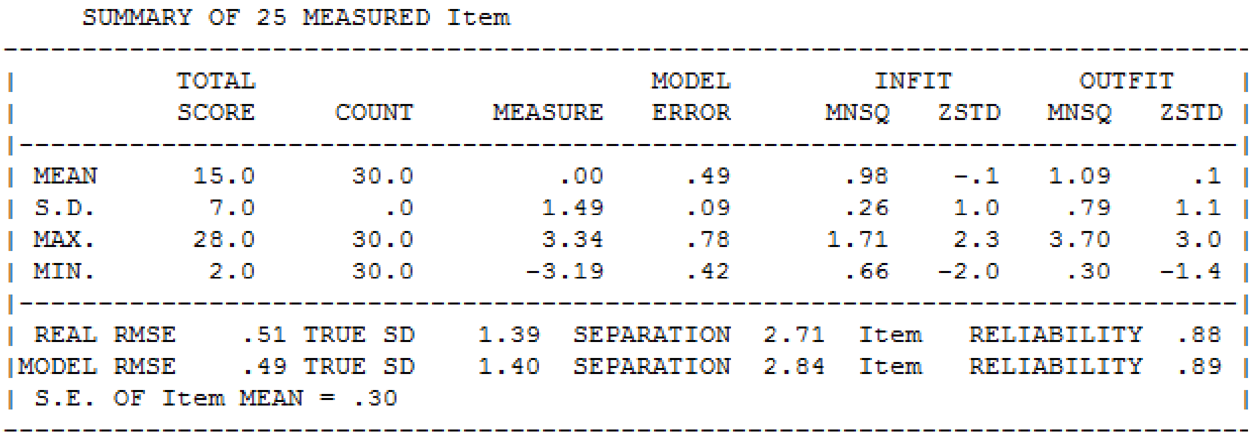
Table 1. ***The Cronbach Alpha Score & Person Reliability value***



```
     SUMMARY OF 30 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT    MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      12.5       25.0        .03     .52        .99    -.1   1.09     .0 |
| S.D.       5.0         .0       1.28     .05        .31    1.2    .91    1.1 |
| MAX.      22.0       25.0       2.67     .68       1.75    2.7   5.02    2.4 |
| MIN.       4.0       25.0      -2.28     .48        .56   -2.3    .32   -1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .56 TRUE SD    1.15  SEPARATION  2.08  Person RELIABILITY  .81 |
|MODEL RMSE    .52 TRUE SD    1.17  SEPARATION  2.23  Person RELIABILITY  .83 |
| S.E. OF Person MEAN = .24                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .84
```

The table shows that the Cronbach Alpha person reliability value is 0.84, and the model reliability is 0.83. Table 2 shows the item reliability value of 0.89. These results indicate that the estimated measures are reliable. This suggests that the test allows us to discriminate between examinees based on their ability, and also discriminate between items based on their difficulty (Bond & Fox, 2001).

Table 2. ***The Item Reliability Value***



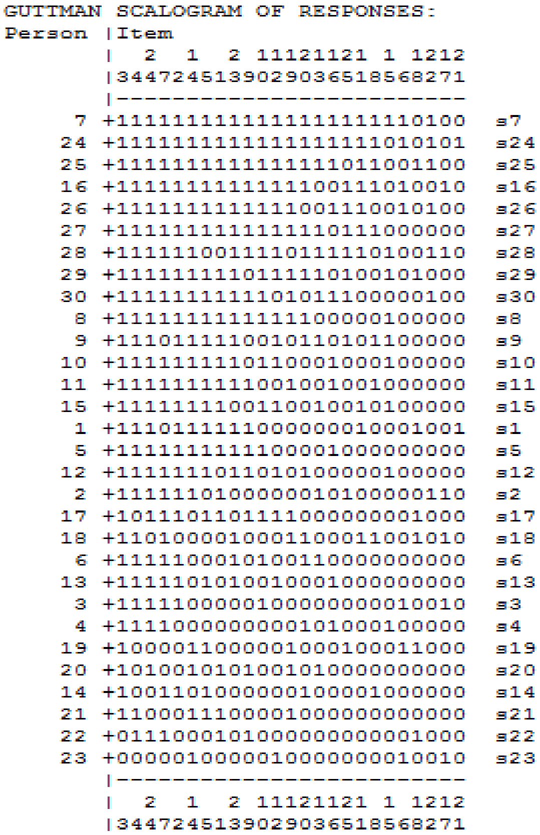
```
     SUMMARY OF 25 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      15.0      30.0         .00     .49       .98    -.1   1.09     .1 |
| S.D.       7.0        .0        1.49     .09       .26    1.0    .79    1.1 |
| MAX.      28.0      30.0        3.34     .78      1.71    2.3   3.70    3.0 |
| MIN.       2.0      30.0       -3.19     .42       .66   -2.0    .30   -1.4 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .51 TRUE SD    1.39  SEPARATION  2.71  Item   RELIABILITY  .88 |
|MODEL RMSE    .49 TRUE SD    1.40  SEPARATION  2.84  Item   RELIABILITY  .89 |
| S.E. OF Item MEAN = .30                                                     |
-------------------------------------------------------------------------------
```

Validity

The second step in the process is to check for test validity. The meaning of validity in testing has evolved over time. In 1954, the American Psychological Association purported four types of validity: concurrent, construct, content and predictive. Predictive and concurrent validity were later combined to form criterion-related validity (Smith, 2001). For Messick (1989, 1995), “validity is a unitary concept realized in construct validity and has six facets of content, substantive, structural, generalizability, external, and consequential” (as cited in Baghaei & Amrahi, 201, p.1052). It is this definition of validity that will form the basis of the analysis, with the most important aspects discussed. The data was analysed using WINSTEPS Rasch Software.

A glance at the item person matrix yields some preliminary observations possibly affecting validity, such as overfit.

Table 3.*Item Person Matrix*

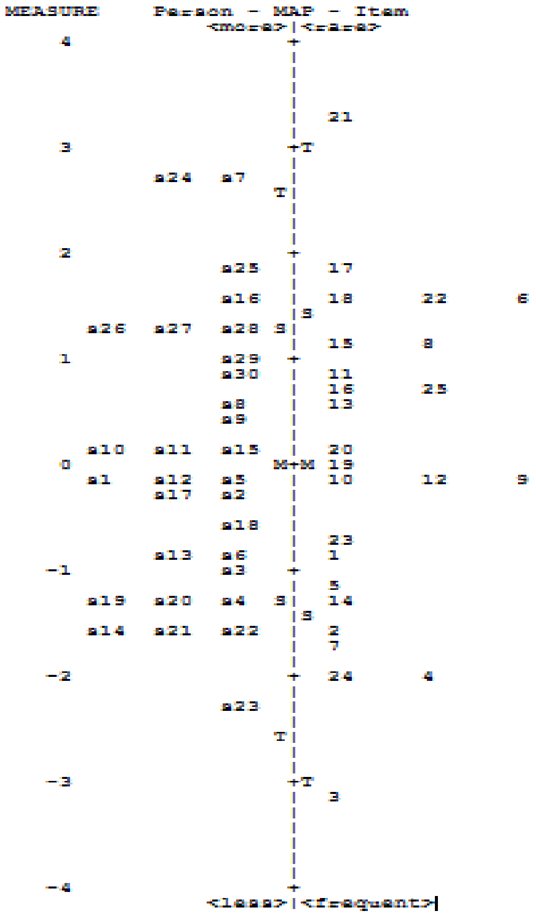


```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |  2  1  2 11121121 1 1212
       |3447245139029036518568271
       |-------------------------
     7 +1111111111111111111110100  s7
    24 +1111111111111111111010101  s24
    25 +1111111111111111011001100  s25
    16 +1111111111111100111010010  s16
    26 +1111111111111001110010100  s26
    27 +1111111111111111011100000  s27
    28 +1111110011110111110100110  s28
    29 +1111111110111110100101000  s29
    30 +1111111111101011100000100  s30
     8 +1111111111111100000100000  s8
     9 +1110111110010110101100000  s9
    10 +1111111110110001000100000  s10
    11 +1111111111001001001000000  s11
    15 +1111111100110010010100000  s15
     1 +1110111111000000010001001  s1
     5 +1111111111100001000000000  s5
    12 +1111111011010100000100000  s12
     2 +1111110100000001010000110  s2
    17 +1011101101111000000001000  s17
    18 +1101000010001100011001010  s18
     6 +1111100010100110000000000  s6
    13 +1111101010010001000000000  s13
     3 +1111100000100000000010010  s3
     4 +1111000000000101000100000  s4
    19 +1000011000001000100011000  s19
    20 +1010010101001010000000000  s20
    14 +1001101000000100001000000  s14
    21 +1100011100001000000000000  s21
    22 +0111000101000000000001000  s22
    23 +0000010000001000000010010  s23
       |-------------------------
       |  2  1  2 11121121 1 1212
       |3447245139029036518568271
```

The table does not show a perfect Guttman pattern. An initial look at the matrix immediately highlights a problem with item 6 and item 24. Item 6 is from level 2 of the construct map, yet only 7 examinees answered the question correctly. However, item 24, which is from level 5 on the construct map has a facility of 25. Such results threaten the validity of the measurement. The matrix does however show that the students with less ability tended to get only the easier items correct. This is important for the validity of the test, even if the difficulty of the items does not match what was initially expected from the construct map.

The concern of the content facet of validity is “that all the items or tasks as well as the cognitive processes involved in responding to them be relevant and representative of the construct domain to be assessed” (Baghaei & Amrahi, 201, p.1052). An item person map (Figure 4) is used to help provide evidence for the content facet of validity.

Figure 4. ***The Item Person Map***

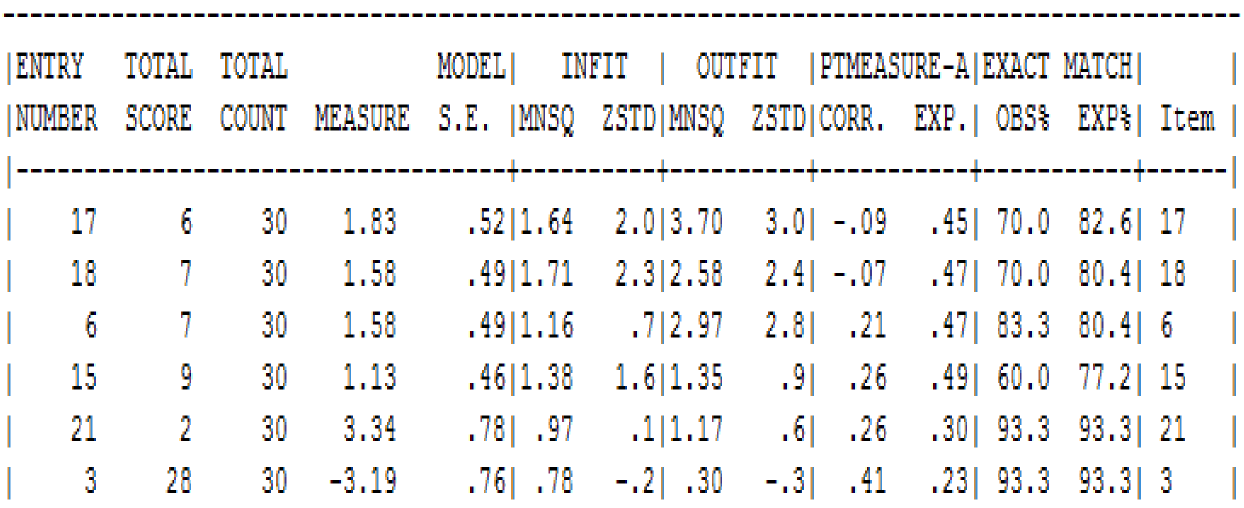


```
MEASURE    Person - MAP - Item
               <more>|<rare>
    4                +
                     |
                     |
                     |
                     |
                     |  21
                     |
    3                +T
                     |
           s24  s7   |
                    T|
                     |
                     |
                     |
    2                +
                s25  |  17
                     |
                s16  |  18     22     6
                     |S
      s26  s27  s28 S|
                     |  15     8
    1           s29  +
                s30  |  11
                     |  16     25
                s8   |  13
                s9   |
                     |
      s10  s11  s15  |  20
    0               M+M 19
      s1   s12  s5   |  10     12     9
           s17  s2   |
                     |
                s18  |
                     |  23
           s13  s6   |  1
   -1           s3   +
                     |  5
      s19  s20  s4  S|  14
                     |S
      s14  s21  s22  |  2
                     |  7
                     |
   -2                +  24     4
                     |
                s23  |
                     |
                    T|
                     |
                     |
   -3                +T
                     |  3
                     |
                     |
                     |
                     |
                     |
   -4                +
               <less>|<frequent>
```

The item person map in Figure 4 scales the ability of the examinees, on the left hand side, as well as the difficulty of the items on the right. As the item person matrix highlighted, item 24 was one of the easier items on the test. Students did not need to be placed very highly on the scale, in terms of ability, to have a predicted chance of 50% of getting the item correct. Moreover, the map suggests that item number 3, with a facility of 28 was too easy. There are a few gaps in the item difficulty continuum, which suggests that some areas of the construct were not covered by the test. The item difficulty did not fully correspond with what was expected from the construct map.

Another way of viewing the content facet of validity is to ascertain whether the students' responses align with their abilities, by looking at the point-measure correlations (PT-MEASURE). Negative correlations usually mean that the responses to the items contradict the continuous latent variable. Table 4 shows the items with the lowest correlations. The correlations for items 17 and 18 were negative, at -0.09 and -0.07 respectively. This is because, although the items were difficult, students with lower ability answered them correctly. Such items threaten the validity of the measurement. The majority of the items do, however, have positive correlations.

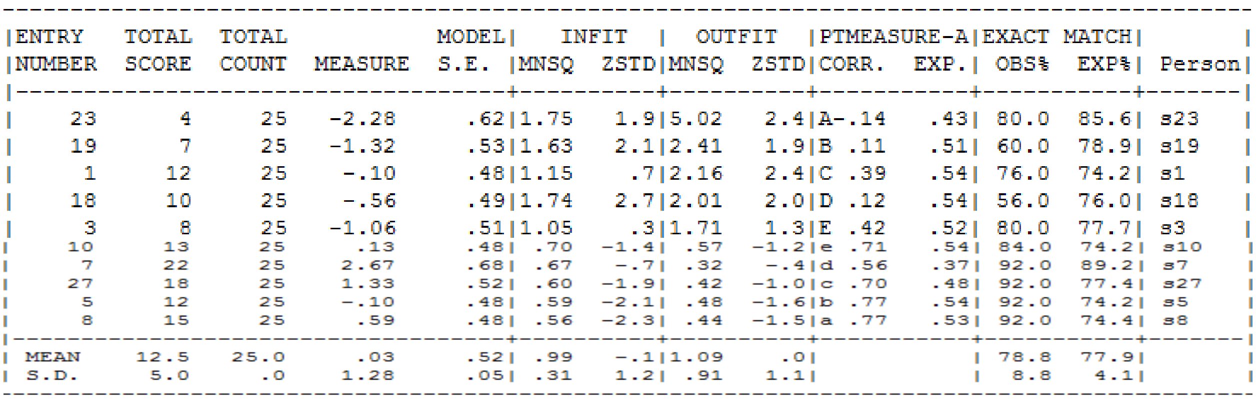
Table 4. ***The items with the lowest correlation***



```
-------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|    17      6     30     1.83    .52|1.64   2.0|3.70   3.0| -.09   .45| 70.0  82.6| 17   |
|    18      7     30     1.58    .49|1.71   2.3|2.58   2.4| -.07   .47| 70.0  80.4| 18   |
|     6      7     30     1.58    .49|1.16    .7|2.97   2.8|  .21   .47| 83.3  80.4| 6    |
|    15      9     30     1.13    .46|1.38   1.6|1.35    .9|  .26   .49| 60.0  77.2| 15   |
|    21      2     30     3.34    .78| .97    .1|1.17    .6|  .26   .30| 93.3  93.3| 21   |
|     3     28     30    -3.19    .76| .78   -.2| .30   -.3|  .41   .23| 93.3  93.3| 3    |
```

The substantive aspect of validity "deals with finding empirical evidence to assure that test-takers are actually engaged with the domain processes provided by the test items or tasks" (Baghaei & Amrahi, 2011, p.1052). One way of evaluating this is by doing a multiple choice distracter analysis, looking at the P-values, which would show "the degree to which the responses to the distracters are consistent with the intended cognitive processes around which the distracters were developed" (Wolfe & Smith, 2007, p. 209). However, as was previously mentioned, the distracters are mainly homogenous, and so no conclusions relating to the construct map were drawn from the distracters examinees chose. Another indicator of the substantive aspect of validity is person fit statistics. Person fit statistics concern "the extent to which a person's pattern of responses to the items correspond to that predicted by the model" (Smith, 2001, p. 296). Table 5 shows the highest and lowest INFIT and OUTFIT mean-square statistics.

Table 5.***The highest and lowest INFIT and OUTFIT mean-square and Z-Standardized statistics***



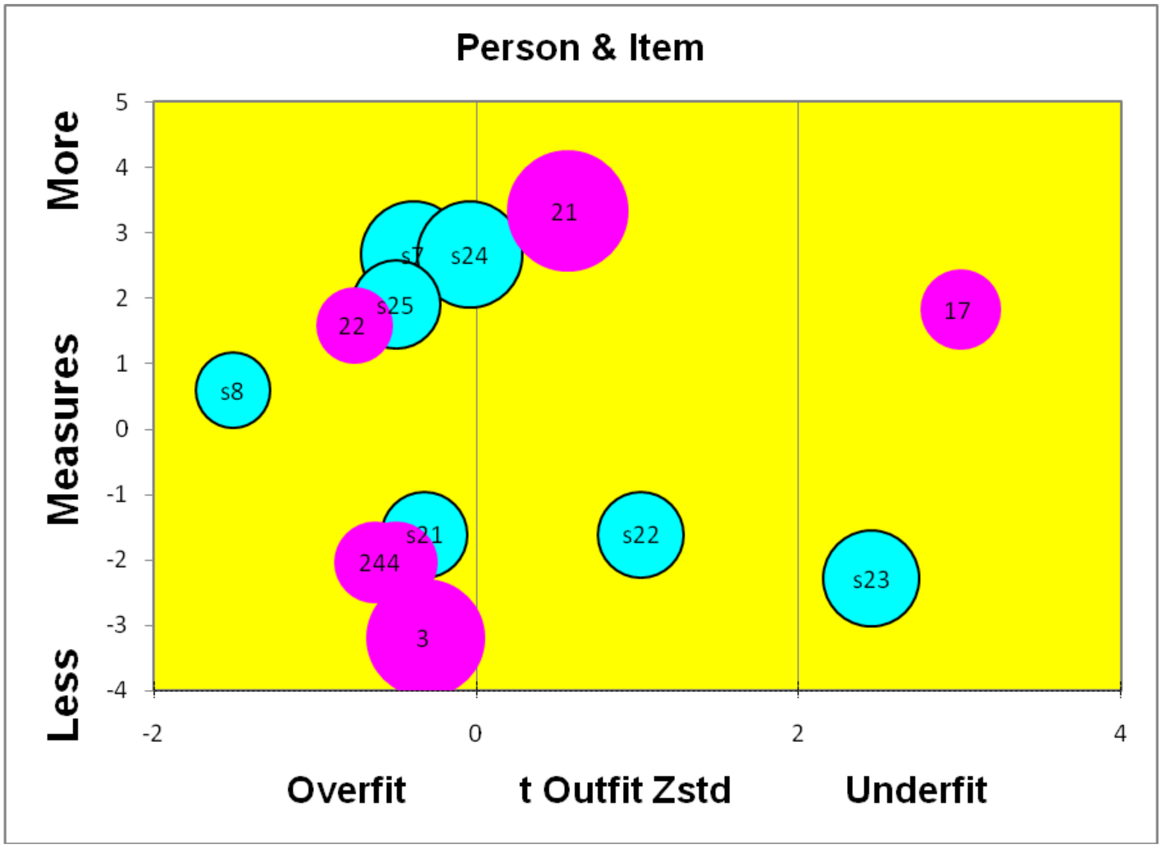
```
---------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Person|
|------------------------------------+----------+----------+-----------+-----------+-------|
|    23      4     25    -2.28    .62|1.75   1.9|5.02   2.4|A-.14   .43| 80.0  85.6| s23   |
|    19      7     25    -1.32    .53|1.63   2.1|2.41   1.9|B .11   .51| 60.0  78.9| s19   |
|     1     12     25     -.10    .48|1.15    .7|2.16   2.4|C .39   .54| 76.0  74.2| s1    |
|    18     10     25     -.56    .49|1.74   2.7|2.01   2.0|D .12   .54| 56.0  76.0| s18   |
|     3      8     25    -1.06    .51|1.05    .3|1.71   1.3|E .42   .52| 80.0  77.7| s3    |
|    10     13     25      .13    .48| .70  -1.4| .57  -1.2|e .71   .54| 84.0  74.2| s10   |
|     7     22     25     2.67    .68| .67   -.7| .32   -.4|d .56   .37| 92.0  89.2| s7    |
|    27     18     25     1.33    .52| .60  -1.9| .42  -1.0|c .70   .48| 92.0  77.4| s27   |
|     5     12     25     -.10    .48| .59  -2.1| .48  -1.6|b .77   .54| 92.0  74.2| s5    |
|     8     15     25      .59    .48| .56  -2.3| .44  -1.5|a .77   .53| 92.0  74.4| s8    |
|------------------------------------+----------+----------+-----------+-----------+-------|
| MEAN    12.5   25.0      .03    .52| .99   -.1|1.09    .0|           | 78.8  77.9|       |
| S.D.     5.0     .0     1.28    .05| .31   1.2| .91   1.1|           |  8.8   4.1|       |
---------------------------------------------------------------------------------------
```

The majority of the INFIT values are within the acceptable range of 0.6 to 1.4 (Bond & Fox, 2007). Only s19 and s23 exceeded this value. INFIT values are more sensitive to responses on items that are roughly targeted at the examinee's ability level. Thus, s23, who got the easiest items wrong, triggered a mean-square value of 1.75. However, none of the examinees have an INFIT mean-square value of greater than 2.0 which would suggest the measurement is inaccurate. The majority of the OUTFIT values are also within the acceptable range of 0.6 to 1.4. However, there is great underfit in the OUTFIT values of some of the examinees, meaning their responses are too unpredictable. S23 has an OUTFIT mean-square value of 5.02. This is because he got the second hardest item correct. It can be assumed, since he got all of the easier items incorrect, that this was a guess. Nearly all the Z-Standardized values (ZSTD) are within the acceptable range of -2.0 to 2.0 (Bond & Fox, 2007).

The consequential aspect of validity "addresses the actual and potential consequences of test score use, especially in regard to sources of invalidity such as bias, fairness, and distributive justice (Wolfe & Smith, 2007, p.244). This is expounded on in the discussion on fairness.

Another useful tool for indicating the construct validity of the instrument is a bubble chart. The bubble chart shows how well the estimated measures for the examinees and the items fit the Rasch model. Figure 5 shows the bubble chart for the items and persons at both extremes.

Figure 5. ***The Bubble Chart of persons and items***

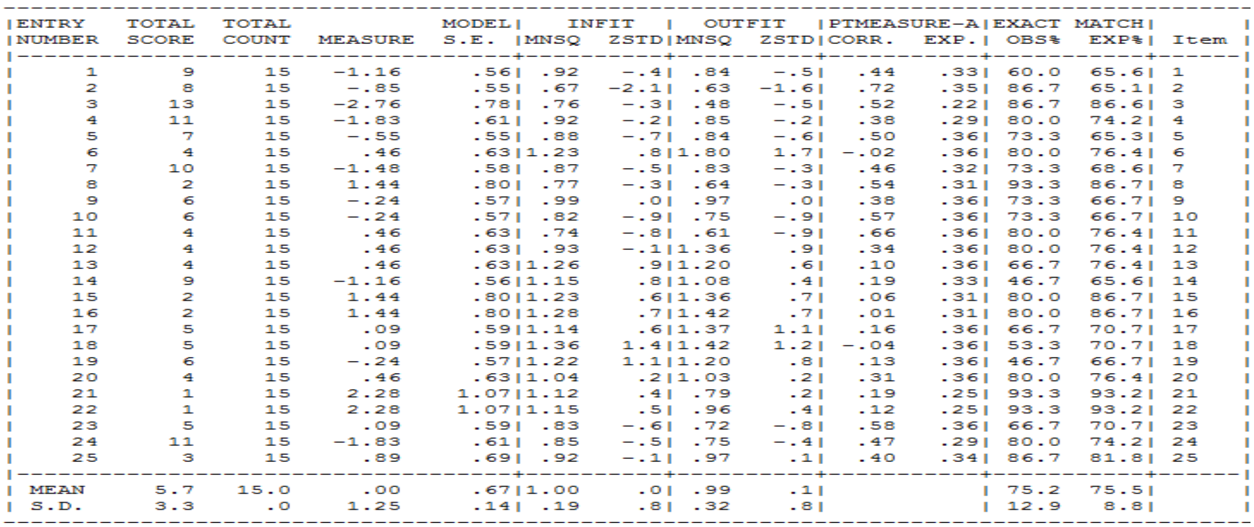


The sizes of the circles show the accuracy of the measure along the latent variable. As previously mentioned, and confirmed by the bubble chart, nearly all of ZSTD values are within the acceptable range of -2.0 to 2.0. There is however major underfit for examines s23 and item 17, meaning that their outcomes are too unpredictable. The reason for s23's underfit has been mentioned previously. Item 17's underfit was due to the fact that it was the second hardest item on the test, and was answered correctly by only 1 of the top 6 students in terms of ability. However, the item was answered correctly by the student with the least ability. This caused the major underfit and the labeling of the item as unpredictable. Overall the data shows that the measurement was valid.

Fairness

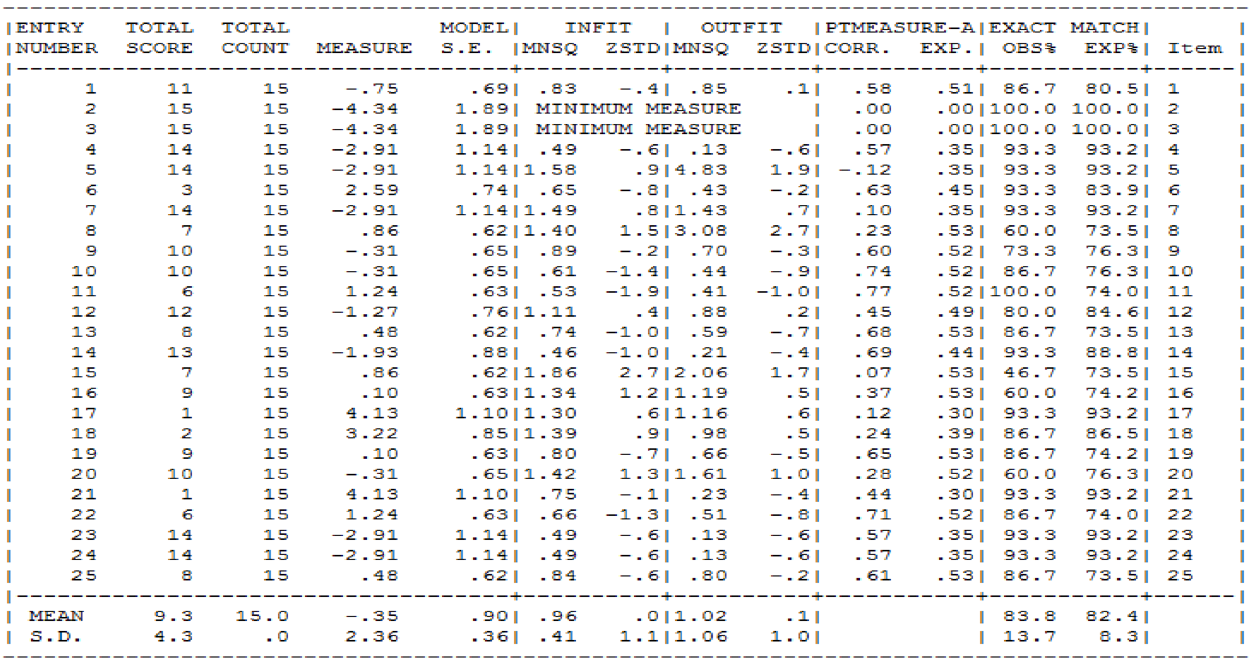
There is no single meaning of fairness in testing, rather there are different aspects that need to be considered. The first of these is whether there was equal treatment of all the examinees with regard to the test conditions and other features of test administration. All of the students in the trial were given one hour to answer the questions. This was ample time for the students to ensure that they answered all of them. Other conditions of the test remained consistent for all the examinees. Another important aspect of test fairness is that the test is free from any kind of bias. The test should not advantage or disadvantage any group of examinees over another. The discriminating factor in the test should be the ability of the student. A fair test is one that produces comparably valid scores from person to person, group to group, and setting to setting (Willingham, 1998). There should be no bias amongst any subgroup. In the sample used to trial this test, all of the students came from the same school, and were all male. The only differentiating characteristic was their age. Half of the participants were 15 years old (grade 9), and the other half were 17 years old (grade 11). The question remains whether there was any bias towards one of these two groups. In order to try to prevent any bias, the vocabulary in the items was taken from grade 5 graded readers which are specifically written for second language learners. As a result, both sets of students should have been very familiar with the vocabulary. In order to determine whether the test, or any individual item, biased one group over another one looks at differential test functioning (DTF) and differential item functioning (DIF). Tables 6 and 7 show the item difficulties for grade 9 and 7 respectively.

Table 6. ***Item details for the grade 9 students***



```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT  MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     1      9     15    -1.16    .56| .92   -.4| .84   -.5|  .44   .33| 60.0  65.6| 1    |
|     2      8     15     -.85    .55| .67  -2.1| .63  -1.6|  .72   .35| 86.7  65.1| 2    |
|     3     13     15    -2.76    .78| .76   -.3| .48   -.5|  .52   .22| 86.7  86.6| 3    |
|     4     11     15    -1.83    .61| .92   -.2| .85   -.2|  .38   .29| 80.0  74.2| 4    |
|     5      7     15     -.55    .55| .88   -.7| .84   -.6|  .50   .36| 73.3  65.3| 5    |
|     6      4     15      .46    .63|1.23    .8|1.80   1.7| -.02   .36| 80.0  76.4| 6    |
|     7     10     15    -1.48    .58| .87   -.5| .83   -.3|  .46   .32| 73.3  68.6| 7    |
|     8      2     15     1.44    .80| .77   -.3| .64   -.3|  .54   .31| 93.3  86.7| 8    |
|     9      6     15     -.24    .57| .99    .0| .97    .0|  .38   .36| 73.3  66.7| 9    |
|    10      6     15     -.24    .57| .82   -.9| .75   -.9|  .57   .36| 73.3  66.7| 10   |
|    11      4     15      .46    .63| .74   -.8| .61   -.9|  .66   .36| 80.0  76.4| 11   |
|    12      4     15      .46    .63| .93   -.1|1.36    .9|  .34   .36| 80.0  76.4| 12   |
|    13      4     15      .46    .63|1.26    .9|1.20    .6|  .10   .36| 66.7  76.4| 13   |
|    14      9     15    -1.16    .56|1.15    .8|1.08    .4|  .19   .33| 46.7  65.6| 14   |
|    15      2     15     1.44    .80|1.23    .6|1.36    .7|  .06   .31| 80.0  86.7| 15   |
|    16      2     15     1.44    .80|1.28    .7|1.42    .7|  .01   .31| 80.0  86.7| 16   |
|    17      5     15      .09    .59|1.14    .6|1.37   1.1|  .16   .36| 66.7  70.7| 17   |
|    18      5     15      .09    .59|1.36   1.4|1.42   1.2| -.04   .36| 53.3  70.7| 18   |
|    19      6     15     -.24    .57|1.22   1.1|1.20    .8|  .13   .36| 46.7  66.7| 19   |
|    20      4     15      .46    .63|1.04    .2|1.03    .2|  .31   .36| 80.0  76.4| 20   |
|    21      1     15     2.28   1.07|1.12    .4| .79    .2|  .19   .25| 93.3  93.2| 21   |
|    22      1     15     2.28   1.07|1.15    .5| .96    .4|  .12   .25| 93.3  93.2| 22   |
|    23      5     15      .09    .59| .83   -.6| .72   -.8|  .58   .36| 66.7  70.7| 23   |
|    24     11     15    -1.83    .61| .85   -.5| .75   -.4|  .47   .29| 80.0  74.2| 24   |
|    25      3     15      .89    .69| .92   -.1| .97    .1|  .40   .34| 86.7  81.8| 25   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN     5.7   15.0      .00    .67|1.00    .0| .99    .1|           | 75.2  75.5|      |
| S.D.     3.3     .0     1.25    .14| .19    .8| .32    .8|           | 12.9   8.8|      |
-------------------------------------------------------------------------------------------
```

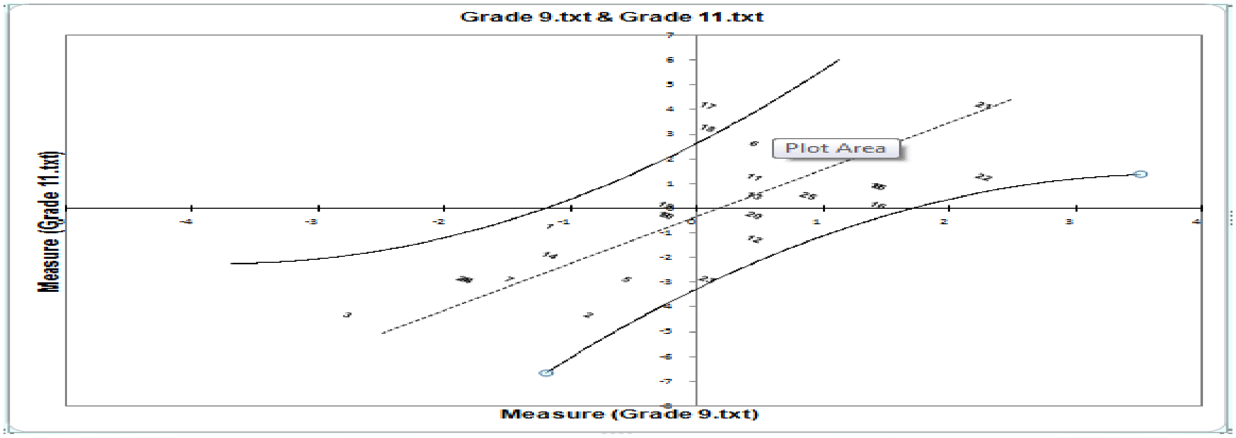
Table 7. ***Item details for the grade 11 students***



```
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     1     11     15     -.75    .69| .83   -.4| .85    .1|  .58   .51| 86.7  80.5| 1    |
|     2     15     15    -4.34   1.89| MINIMUM MEASURE     |  .00   .00|100.0 100.0| 2    |
|     3     15     15    -4.34   1.89| MINIMUM MEASURE     |  .00   .00|100.0 100.0| 3    |
|     4     14     15    -2.91   1.14| .49   -.6| .13   -.6|  .57   .35| 93.3  93.2| 4    |
|     5     14     15    -2.91   1.14|1.58    .9|4.83   1.9| -.12   .35| 93.3  93.2| 5    |
|     6      3     15     2.59    .74| .65   -.8| .43   -.2|  .63   .45| 93.3  83.9| 6    |
|     7     14     15    -2.91   1.14|1.49    .8|1.43    .7|  .10   .35| 93.3  93.2| 7    |
|     8      7     15      .86    .62|1.40   1.5|3.08   2.7|  .23   .53| 60.0  73.5| 8    |
|     9     10     15     -.31    .65| .89   -.2| .70   -.3|  .60   .52| 73.3  76.3| 9    |
|    10     10     15     -.31    .65| .61  -1.4| .44   -.9|  .74   .52| 86.7  76.3| 10   |
|    11      6     15     1.24    .63| .53  -1.9| .41  -1.0|  .77   .52|100.0  74.0| 11   |
|    12     12     15    -1.27    .76|1.11    .4| .88    .2|  .45   .49| 80.0  84.6| 12   |
|    13      8     15      .48    .62| .74  -1.0| .59   -.7|  .68   .53| 86.7  73.5| 13   |
|    14     13     15    -1.93    .88| .46  -1.0| .21   -.4|  .69   .44| 93.3  88.8| 14   |
|    15      7     15      .86    .62|1.86   2.7|2.06   1.7|  .07   .53| 46.7  73.5| 15   |
|    16      9     15      .10    .63|1.34   1.2|1.19    .5|  .37   .53| 60.0  74.2| 16   |
|    17      1     15     4.13   1.10|1.30    .6|1.16    .6|  .12   .30| 93.3  93.2| 17   |
|    18      2     15     3.22    .85|1.39    .9| .98    .5|  .24   .39| 86.7  86.5| 18   |
|    19      9     15      .10    .63| .80   -.7| .66   -.5|  .65   .53| 86.7  74.2| 19   |
|    20     10     15     -.31    .65|1.42   1.3|1.61   1.0|  .28   .52| 60.0  76.3| 20   |
|    21      1     15     4.13   1.10| .75   -.1| .23   -.4|  .44   .30| 93.3  93.2| 21   |
|    22      6     15     1.24    .63| .66  -1.3| .51   -.8|  .71   .52| 86.7  74.0| 22   |
|    23     14     15    -2.91   1.14| .49   -.6| .13   -.6|  .57   .35| 93.3  93.2| 23   |
|    24     14     15    -2.91   1.14| .49   -.6| .13   -.6|  .57   .35| 93.3  93.2| 24   |
|    25      8     15      .48    .62| .84   -.6| .80   -.2|  .61   .53| 86.7  73.5| 25   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN     9.3   15.0     -.35    .90| .96    .0|1.02    .1|           | 83.8  82.4|      |
| S.D.     4.3     .0     2.36    .36| .41   1.1|1.06   1.0|           | 13.7   8.3|      |
-------------------------------------------------------------------------------------------
```

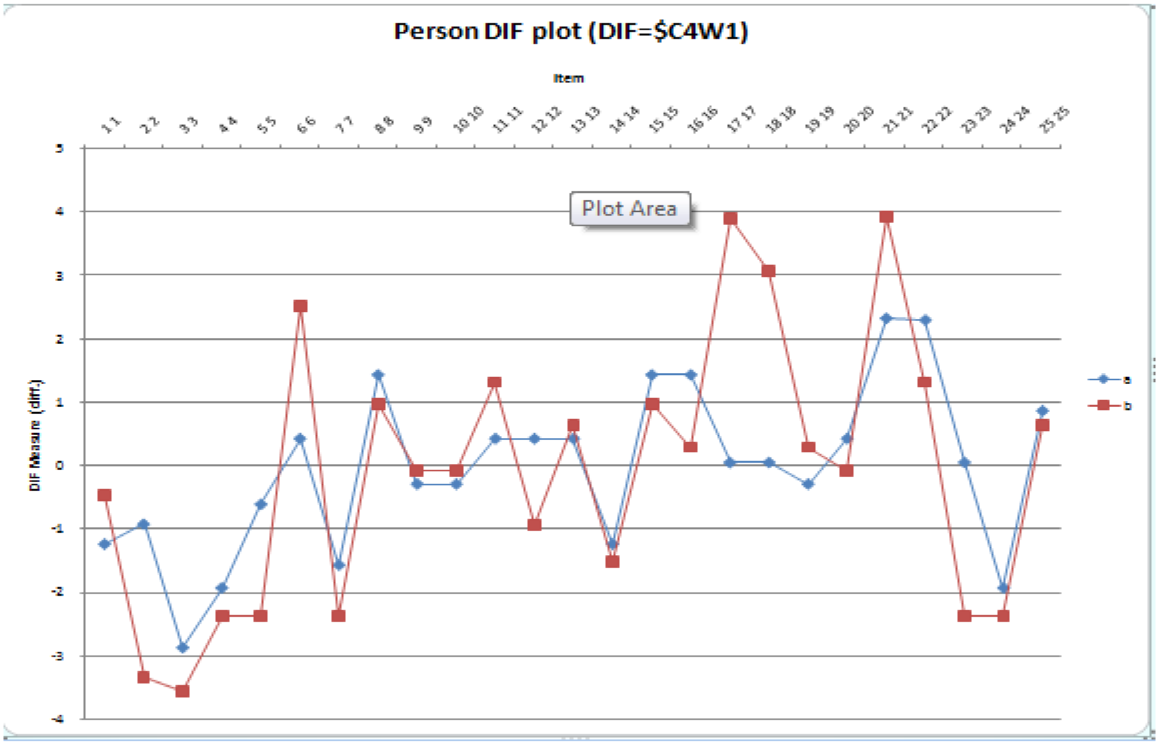
A comparison between the two age groups is made easier using a scatter plot as shown in Figure 6.

Figure 6. ***Scatter Plot showing grade 9 and grade 11 results***



The scatter plot confirms that overall there was no bias in the test towards either of the two grades. However, items 17 and 18 were relatively more difficult for the grade 11 students. Both of these items have a high difficulty level on the test. With only a very small number of students getting the items correct it cannot be said that the items were biased. The DIF graph which is shown in Figure 7 also shows that there was no bias in the test. Both groups performed very similarly on each item. This is with the exception of items 17 and 18. However, with the relatively small sample used to trial the test, and then even smaller number of examinees who answered the item correctly, no credible claim of bias can be made.

Figure 7. ***Person DIF plot***



Conclusion

In conclusion, an analysis of the data shows that the test was mostly reliable, valid and fair. The items did test the construct, grammar competency. In general, the students needed more ability to get the items with higher difficulty correct. However, the item difficulty of each item does not match with what was expected from the construct map. This may be because the leveling of grammar items by the British Council and EQUALS is inaccurate. It may also be because certain items were not written well. Before such conclusions could be made, and the construct map altered, the test would need to be developed further. During this development, further items would need to be added for each grammar topic. Instead of having just one question on the present perfect continuous tense, the test would contain at least two. This would enrich the data for analysis. Moreover, the developed test could be given to a larger sample which includes females.

One possible reason why the item difficulties were not as expected is the fact that the first language of every student impacts upon their learning. As a result, there are certain grammar items which are particularly difficult. Grammatical structures which are familiar to the learner in his mother tongue are easier to grasp. However, those which are alien to the student may cause great difficulty, even if the topic is considered easy. Two examples of this for Arabic speakers is the present continuous tense and the verb *to be.* In Arabic the structure for the present simple tense and the present continuous tense is the same. Moreover, the verb *to be* is omitted in the present simple tense. This is highlighted in Figure 8.

Figure 8. ***Grammar topics which cause specific difficulty for Arabic speakers***

Arabic	Literal Translation	Meaning
أحمد يلعب كرة القدم	Ahmad plays football	Ahmad plays football Ahmad is playing football
أنا طالب في جامعة درهم	I student in Durham University	I am a student at Durham University.

In my experience teaching in the Middle East, learners whose first language is Arabic tend to find difficulty with the grammatical structures highlighted in Figure 8. Using the verb *to be* is placed at A1 level by the British Council and EQUALS. However, due to the lack of the structure in Arabic, learners may take more time to grasp its use. This may explain, for example, why item 1 had a higher difficulty than item 7. Item 7, which tests the past continuous tense, is familiar to the students in Arabic. However, item 1 uses a different preposition to the one they are familiar with. This is shown in Figure 9.

Figure 9. ***Item 1 from the grammar test***

1. I am going to Riyadh ___________ Friday.

   a) on b) at c) in d) by

In English the preposition *on* is used with days. However in Arabic the preposition في is used, which translates as *in*. The item is likely to be easier for students who also use the preposition, *on*, in their mother tongue. To examine this possibility, a developed test should be given to English learners whose first language differs.

Acknowledgement: I would like to thank Mr. Saabir Mark Alaji for his help administering and grading the test.

About the Author:

Voke Efeotor is a language instructor and test item writer from the United Kingdom. He has been working in Saudi Arabia for eight years, and is presently working at Taibah University in Madinah, Saudi Arabia. At Taibah University he is part of the Testing Unit which is responsible for producing exams for the preparatory year English language program. He is a doctorate student at Durham University in the UK, and his area of interest is assessment. Prior to embarking upon his doctorate, he completed an MA in Linguistics, and he has a bachelors degree in law from King's College London.

References

Abedi, J. (2006). Language Issues in Item Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Baghaei, P. & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching and Research,* 2 (5), 1052-1060.

Bond, T. & Fox, C. (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Bond T. G. & Fox, C.M. (2007). (2nd ed.) *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics 5*: 160–170.

Core Inventory for General English. (2010). Retrieved April 5, 2014, from http://www.teachingenglish.org.uk/sites/teacheng/files/Z243%20E&E%20EQUALS%20BROCHURErevised6.pdf

Downing, S. M. (2006a). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Downing, S. M. (2006b). Selected-Response Item Formats in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide andreference*. 11.0 update (4th ed.). Boston: Allyn & Bacon.

Gorard, S. (2013). *Research Design: Creating Robust Approaches for the Social Sciences.* Thousand Oaks, CA: SAGE Publications, Inc.

Guler, N., Uyanik, G. & Teker, G. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education,* 2 (1), 1-6.

Iyengar, S. (1996). Framing responsibility for political issues. *Annals of the of Political and Social Science*, 546 (1), 59-70.

Hambleton, R., & Jones, R. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38-47.

Hambleton, R., Swaminathan, H., Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications, Inc.

Larsen-Freeman, D. (2001). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language.* (3rd ed.) (pp. 251-266). Boston: Heinle & Heinle.

Lord, F. M. (1977). Optimal number of choices per item--a comparison of four approaches. *Journal of Educational Measurement*, *14,* 33-38.

Marlon, F. (1981). Phenomenography—Describing conceptions of the world around us. *Instructional Science*, *10,* 177-200.

McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Publishing Company.

Messick, S. (1989). Validity. In R.L. Linn (ed.) *Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (2000). Consequences of Test Interpretation and Use: The Fusion of Validity and Values in Psychological Assessment. In D. Jackson & R. Goffin (Eds.), *Problems and Solutions in Human Assessment: honoring Douglas N. Jackson at seventy*. (pp. 3-21). Norwell, MA: Kluwer Academic Publishers.

Mitchell, R. & F. Myles (1998). *Second Language Learning Theories*. London: Arnold.

Rees, P. J. (1986). Do medical students learn from multiple choice examinations? *Medical Education*, *20,* 123-125.

Repass, D. E. (1971). Issue Salience and Party Choice. *The American political science review 65* (2), 389.

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian S, Albertson B, Rand DG (In press). Structural topic models for open-ended survey responses. *American Journal of Political Science.*

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24* (2), 3- 13.

Smith, E. V. Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2 (3), 281-311.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true–false examinations. *Journal of Educational Research*, *83*, 119 –124.

Wells, S.C. & Wollack, J.A. (2003). *An Instructor's Guide to Understanding Test Reliability*. Testing & Evaluation Services publication, University of Wisconsin.

Willingham, W.W. (1998) A systemic view of test validity. In *Assessment in Higher Education* , S. Messick, ed. Mahwah, NJ: Erlbaum.

Wilson, M. (2005). *Constructing Measures*: *An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. W. & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement*, 8 (2), 204-234.

Wright, B. D., Mead, R. and Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. *MESA Research Memorandum Number 22*, Chicago: University of Chicago, MESA Psychometric Laboratory.

Appendix 1- The grammar test

GRAMMAR QUIZ (60 minutes)	
Name	

Circle the correct answer for each item: a, b, c or d.

1. I am going to Riyadh __________ Friday.

a) on b) at c) in d) by

2. Ahmad __________ late for class yesterday.

a) were b) is c) was d) are

3. Ali is tall and __________ hair is black.

a) our b) your c) her d) his

4. Jeddah is _______________ than Madinah.

a) big b) more big c) bigger d) more bigger

5. Ibrahim _____________ to eat rice and fish.

a) like b) liking c) likes d) to like

6. Which sentence is **NOT** correct?

a) They ran quickly. b) We worked hardly.
c) The boys slept heavily. d) He spoke loudly.

7. I ___________________ until 7 o'clock yesterday.

a) was working b) were work c) was work d) were working

8. The three ____________ house was very nice.

a) brothers b) brother's c) brothers' d) brothers's

9. If you ___________ online, you save a lot of money.

a) order b) ordered c) orders d) have ordered

10. He has _____________ this program before.

a) sees b) see c) saw d) seen

11. Michael loves his car, _____________?

a) isn't he b) doesn't he c) does he d) is he

12. If I had more money, I ______________ that bag.

a) would buy b) will buy c) buy d) bought

13. When I got home, the children had already ______________.

a) eating b) eat c) eaten d) ate

14. What _______________________ doing?

a) you have been b) have you been c) you been d) was you been

15. Omar _________________ the bus.

a) can have missed b) can missed c) must has missed d) must have missed

16. By October, I will __________________ English for five years.

a) be studying b) study c) have studying d) have been studying

17. I think the match __________________ by the time we get home.

a) will have finished b) has finished c) will has finished d) will be finish

18. If he ______________ your car, you should pay him.

a) had washed b) has washed c) would wash d) will have washed

19. We walked to the middle of the park, ___________ we stopped to play football.

a) where b) that c) which d) who

20. My grandfather ____________________________ at the airport.

a) was used to work b) was use to work c) use to work d) used to work

21. __________________________ such well-behaved children.

a) Never have I met b) Never I have met c) I never have met d) I have met never

22. In Ghana, a two-year-old British girl ___________________ with her parents after being freed by kidnappers in Southern Ghana.

a) has been reunited b) had reunited c) has reunited d) had been reunited

23. The Eiffel Tower ________________ by millions of people this year.

a) has visited b) has been visited c) has been visiting d) was visiting

24. Which sentence is **NOT** correct?

a) Tomorrow, I will wash up the dishes. b) Yesterday, we ran of water out. c) Please turn on the light. d) Please pick your toys up.

25. He _______________________ his mathematics test, if he'd really tried.

a) could pass b) could have passed c) could has passed d) could passed

THE END