The Journal of Applied Linguistics
Vol.9, No.18, pp.118-142, Spring & Summer 2016

Implicational Scaling of Reading Comprehension Construct: Is it Deterministic or Probabilistic?

Parisa Daftarifard[*1], Agustin Tristán [2]

1. *Faculty of Persian Literature and Foreign Languages, South Tehran Branch, Islamic Azad University, Tehran, Iran*
2. *Instituto de Evaluacion e Ingenieria Avanzada, S.C. San Luis Potosi, Mexico*

*Corresponding author: pdaftaryfard@azad.ac.ir

Received: 2016.7.25
Revisions received: 2016.9.28
Accepted: 2016.11.30

Online publication: 2016.12.2

Abstract

In English as a Second Language Teaching and Testing situations, it is common to infer about learners' reading ability based on his or her total score on a reading test. This assumes the unidimensional and reproducible nature of reading items. However, few researches have been conducted to probe the issue through psychometric analyses. In the present study, the IELTS exemplar module C (1994) was administered to 503 Iranian students of various reading comprehension ability levels. Both the deterministic and probabilistic psychometric models of unidimensionality were employed to examine the plausible existence of implicational scaling among reading items in the mentioned reading test. Based on the results, it was concluded that the reading data in this study did not show a deterministic unidimensional scale (Guttman scaling); rather, it revealed a probabilistic one (Rasch model). As the person map of the measures failed to show a meaningful hierarchical order for the items, these results call into question the assumption of implicational scaling that is normally practiced in scoring reading items.

Keywords: deterministic model of unidimentionality, Guttman scaling, implicational scaling, Rasch model, reading comprehension

Introduction

A common approach to teach and test reading comprehension of English as a Second Language is based on the use of testlets. A testlet consists of a text followed by a series of questions or items which are hypothetically believed to measure one or more subskills of reading comprehension. Sometimes, the experts categorize the skills in terms of their semantic relationship to the text and the reader (Grabe, 2009; Nuttall, 1996), but other times the judges assign them to different levels of understanding as suggested by Gray (1960). There are still others who propose different frameworks to describe these subskills.

Barrett's framework (1968) provides a guide to identify the purposes and approaches that the student may follow to read a text, organized in five reading comprehension stages (from low to high). This framework is also sometimes referred to as Kral's taxonomy (1995). Elsewhere, Cazden (1971) proposes a taxonomy for "Early Language Behaviors," mainly to assess language development in preschool and early grades of basic education, focused on cognitive and affective domains; the framework can be used for second language learning purpose with a mere modification. Other taxonomies considering cognitive and affective domains were proposed by Foley (1971), Moore and Kennedy (1971), and Valette (1971). The latter was mainly suggested for the second language learning context. Glass (N/D) suggested an adaptation of Bloom's taxonomy for German as a second language.

It is evident that most taxonomies of second language reading comprehension, many of which are believed to be derived from Munby's (1978) work on ESP and Benjamin Bloom's "Taxonomy of Educational Objectives" (Bloom, 1956; Alderson, 2000), are not empirically based (Alderson, 1991). It happens that the hierarchical taxonomical categories do not necessarily correspond to a hierarchical difficulty of the tasks to be developed by the student. A possibility of hierarchical/difficulty analysis may be represented by the SOLO taxonomy suggested by Biggs and Collis (1982), where the categories are intended to report both the complexity of the mental process and the difficulty of the task. However, those who are interested in the issue have put their efforts into describing, justifying, testifying, or refuting the possibility of empirically-based reading comprehension skill constructs (Alderson, 2000; Bloom, 1994; Grabe, 1991, 1997; Hulstijn, 1997; Roberts,

1974; Weir & Porter, 1996). Such efforts resulted in constructing different taxonomies, whose qualitative categories pretend to correspond to quantitative values or numbers describing reading comprehension constructs in second language (L2) contexts (Daftarifard, 2002).

Although different in number they are, the available reading taxonomies organize the reading subskills mostly in a similar way. Grabe (1997), for example, classified reading components into thirty-three categories: "orthographic processing, phonological coding, word recognition (lexical access), working memory activation, sentence parsing, propositional integration, propositional text-model information, comprehension strategy use, inference making, text model development, situation model development (or mental model), etc." (p. 9). Elsewhere, he categorized reading components into six more general constituents of automatic recognition skills, vocabulary and structural knowledge, formal discourse structure knowledge, content/ world knowledge, synthesis and evaluation skills/ strategies, metacognitive knowledge and skills monitoring (Grabe, 1991). Nuttall (1996), however, divided meaning into four categories of conceptual, propositional, contextual and pragmatic meaning. De Lopez, Marchi, and Coyle (1997) defined a taxonomy for reading comprehension subskills based on both text and stem factors, with thirty-three text categories and fifty-five stem categories. Finally, Daftarifard (2002) identified more than 200 subskills for reading comprehension. To the present authors, however, such lists might be (a) finely differentiated types of tasks, (b) bundles of test task characteristic, (c) processes involved in performing tasks, or (d) a combination of these three attributes.

What seems to be agreed upon among experts in the reading assessment is that certain skills seem to be more difficult than others. Therefore, students who gain higher score on the sitting tests of TOEFL and the IELTS are believed to demonstrate the higher-order ability. This implies that reading questions that are usually practiced both in L2 teaching and assessment contexts form a hierarchical continuum along which some questions are considered to measure higher-order ability and some lower-order ability (e.g., see Alderson, 2000, 1990; Weir & Porter, 1996; McNamara, 1996 for the discussion). In other words, the different layers of understanding have a hierarchical relationship, which can be imagined in the form of a pyramid. The main perpendicular axis of the pyramid locates the easiest sub-skills of reading comprehension ability at

the base, and at the top locates the most difficult ones (See Figure 1). The vertical axis may group some skills with similar and meaningful index of difficulty, which if drawn together will make the shape of a pyramid or a tetrahedron. Looking at the pyramid from above, one can see concentric circles (also represented by squares), whose center represents a separate reading ability. According to the hierarchical hypothesis, if the respondents can reach the outer layers of such a shape, they must have reached the inner too. This attribute is called scalability in the theory of measurement (Guttman, 1974; Bart & Krus, 1973; Hyltenstam, 1977; Anderson, 1978).

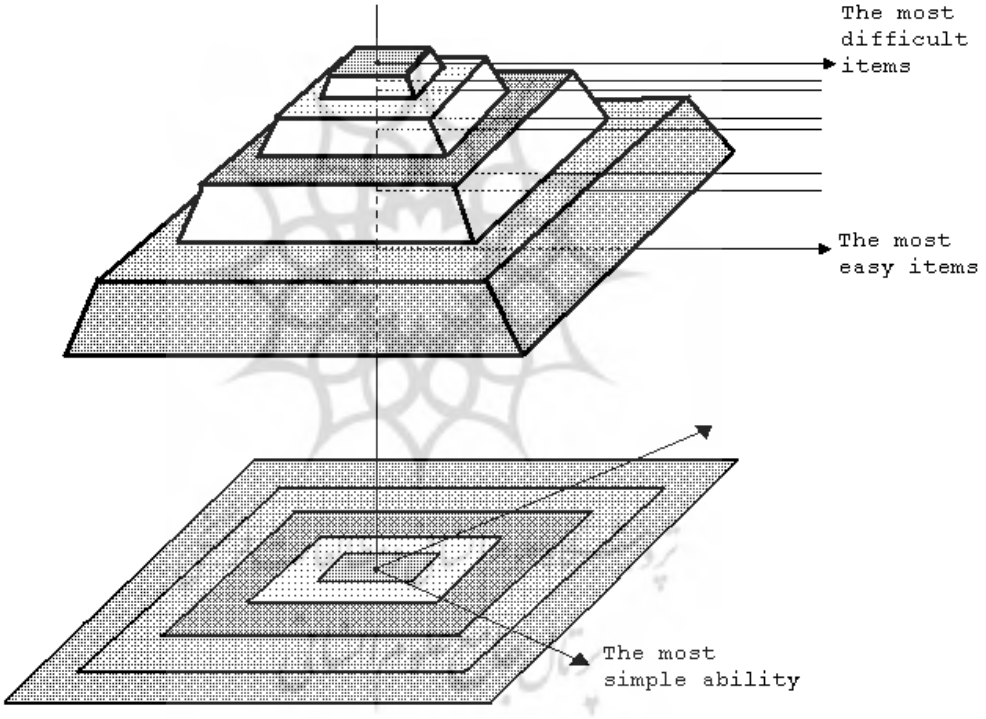

Figure 1. The Schematic Concept of Implicational Scaling in reading comprehension

The scalability assumption of reading comprehension constructs initiated certain arguments especially in the1990s (Alderson, 1990; Mathews, 1990; Weir, Hughes & Porter, 1990). For example, in trying to test the implicational scaling assumption that underlay the reading tests, Alderson (1990) could not find any meaningful concurrence among the judges of his study on what

exactly each item of the ELTS (English Language Testing System) and TEEP (Test in English for Educational Purposes) might measure. Then, he questioned the assumption of implicational scaling in reading theory. Later, the research was questioned on the basis of the concept itself as well as the methodology used to test it (Mathews, 1990; Weir, Hughes, & Porter, 1990). On the other hand, Hillocks and Ludlow (1984) and Ludlow and Hillocks (1985), in a more rigorous study on fiction, found an implicational scaling pattern among some reading comprehension questions. They classified reading ability into two main skill types of literal and inferential levels of comprehension. In this model, the literal level of understanding required identification of explicitly stated information with three question types of basic stated information, key detail, and stated relationship. In addition, the inferential level of understanding required generalizations about implicitly stated relationships in the text with four question types: simple implied relationship, complex implied relationship, author's generalization, and structural generalization (Hillocks & Ludlow, 1974). Employing both the Rasch model and Guttman scaling procedures, they found out that there was a Guttman scaling pattern among these seven skills so that lower-order ability in the hierarchy of reading has the lower item difficulty. They did not report on hierarchical ordering of items, though.

The reproducibility of reading skills is still far from clear. The hierarchical assumption underlying reading measures is so appealing that usually tests developers try to calibrate the items in terms of item difficulty while almost always nothing is mentioned regarding the type of items in terms of level of cognition (Alderson, 1991). Because of this assumption, the respondents' results on reading tests equals the sum of the correct responded items; item difficulty seems to be the best criterion based on which test designers choose the reading items. However, most of the times, difficult items belong to lower-order abilities than easier ones (Weir & Porter, 1996), which also seem to contribute less to total reading ability (Meyer, 1975 as cited in McNamara, 1996). This fact is in contradiction with the reproducibility assumption underlying the construct. As Weir and Porter (1996) mentioned, the main reason for applying the reproducibility assumption into test constructions might be 'the *practical expediency rather than . . . a principled view* of unidimensionality' [our emphasis] (p. 1).

The answers to the issues raised above have important practical implications for teaching situations and assessing reading skills. In teaching situations, second language acquisition research (SLAR) emphasizes the ordering of materials in terms of difficulty. Usually, this is practiced through readability formula while some research enquiries have found readability formulas of little use in L2 reading (Carrell, 1987). Moreover, there is less attention paid to the level of cognition of the questions or tasks usually made for each lesson.

Research mostly proves the unidimensional nature of reading comprehension (e.g., Rost, 1993), mainly through Rasch analysis or any other IRT methods; however, not much research has addresses the implicational and reproducible nature of reading scale. This would query the true starting point for teaching reading comprehension, especially in second and foreign language contexts. In assessment too, the *additivity* and weighing concepts, which are usually practiced in scoring reading tests, require more probes into the reproducibility of reading comprehension. To this end, various analyses were utilized to answer the following questions:

1. Is there any implicational scaling, in Guttman's sense, among reading questions through Guttman Scaling which is a deterministic model?
2. Is there any implicational scaling, in Guttman's sense, among reading questions though Rasch model which is a probabilistic model?
3. If there emerges any implicational scaling of reading scale, is the hierarchical ordering of reading data predictable by reading theory (that is the concept of reproducibility)?

Method

Participants

Five hundred and three Iranian students varying in their ability (151 sophomores, 140 juniors, 150 seniors, and 62 MA students) from different universities in Iran (Azad University Central Branch, Roodehen University, Tehran University, Shahid Beheshtee University, IUT University, Khatam University & IUST University) participated in this study. The participants were studying different sub-fields of English language study (literature,

teaching and translation) at both undergraduate and graduate levels (sophomore, junior, senior in bachelor degree, and first year of master of art degree) in Tehran, Iran. The reason for selecting the subjects from different levels of education was to ensure the heterogeneity of the subjects in their reading ability in order to investigate the scalability of scores for their reading comprehension (Guttman, 1974). The participants included 386 females and 117 males.

The pattern of data collection was convenience method of sampling. No information was collected as to how they learned reading out of university; however, the method of teaching reading comprehension and the way their books approach reading comprehension were the same, which is either the content based or strategy based approach usually practiced in Iranian universities.

Instrumentation

The instrument used in this study was the reading section of the IELTS exemplar (1994) of the EAP version Module C. The IELTS exemplar consisted of three passages. Passage one included 13 short answer questions whose answers were exactly stated in the passage. Passage two encompassed eight matching questions and seven fill-in-the blank questions. Passage three included four multiple choice and four matching questions. The students were asked to read the latter two passages and answer the questions by writing the appropriate letter in the corresponding box on their answer sheet. The question type was indicated intuitively based on the definition provided in the literature.

The first 13 items seemed to measure different abilities such as understanding the implied information, searching information, interpreting the attitudinal meaning abilities, paraphrasing, understanding the propositional meaning and critical thinking. The second passage encompassed 15 items, eight of which were assumed to measure the ability to understand the main idea of different paragraphs. The last seven items on the second passage seemed to measure the ability to summarize the text. The third passage contained eight items, four of which were multiple-choice questions (MC), and the rest were transcoding information items. The four MC items seemed to measure the ability to understand the cohesive ties, cause and effect relationships, the explicitly stated information, and make inference based on the unstated

information in the text successively. The last four items in matching form seemed to measure the ability to transcode information (Table 1).

Table 1.
Organization of the IELTS Exemplar

Sectionof the test	Number of items	Item	Framework or taxonomical ability	Theoretical hierarchical level			
				1	2	3	4
Passage 1	13	V1 to V4	Factual questions	X			
		V5	Scanning	X			
		V6	Conclusion based on the word “gradually”			X	
		V7	Metadiscourse “unlike”		X		
		V8	Meta discourse “unlike”		X		
		V9	Conclusion based on the word “originally”			X	
		V10 to V13	Syntactical meaning		X		
Passage 2 Task 1	7	V14 to V20	Main idea			X	
Passage 2 Task 2	8	V 21 to V28	Summary				X
Passage 3	8	V 29	Understanding cohesive mark		X		
		V 30	Understanding cause and effect relation		X	X	
		V 31 to V32	Detail of information			X	
		V 33 to V36	Transcoding				X

Note that
Level 1 is related to scanning or searching for information
Level 2 is syntactical information and is not related to the relationships between sentences
Level 3 is related to the whole paragraph and mostly conclusion
Level 4 is related to the whole text and is summary through text or graph

Procedure

Once the test was administered to the focal students, the responses were coded into a data base and analyzed for item and test calibration. The IELTS test had an acceptable reliability internal consistency (= 0.83). Also, the average item total correlation (AITC) for the test indicates that the ELTS has a high item total correlation (AITC = 0.38, SD = 0.14) which indicates that there is a good harmony among the items.

For the main analyses, the data were subjected to both Guttman scaling using Antropoc program and Rasch model using Winsteps. Antropoc is a

computer software used to estimate Guttman scaling through several methods of measuring errors and scales. The methods in Antropoc are: (1) minimize errors (reproducibility errors minimized via heuristic algorithm), (2) graduated (scale scores determined by the hardest item checked off), (3) Goodenough-Edwards (scale scores one simple count of number of items checked off), and (4) test (matrix is tested for scalability without rearrangement or rows/columns). In this study, minimized error was selected to analyze the data due to its similarity to the concept of the original Guttman scaling.

From the deterministic point of view, the inventory is scalable if the matrix meets three criteria: A coefficient of reproducibility index (Crep) of 90% or higher, a minimal marginal reproducibility index (MMrep) of 90% or above, and the coefficient of scalability index of 0.60 and above (Hatch & Farhady, 1985).

The Rasch model is also employed using Winsteps. This model may be assumed to be the probabilistic approach to Guttman scaling. Rasch model has an advantage over Guttman's in that the former is based on a probabilistic model that accounts stochastic errors, while the latter has been criticized as being too restrictive in accounting for errors (Hessels, personal communication, 2000).

In Winsteps, two diagnostic Rasch fit statistics are provided: INFIT and OUTFIT. INFIT index provides information about misfit items when item difficulties are targeting the person abilities while OUTFIT index provides misfit information when item difficulties are far from the person abilities. In both cases, Winsteps reports a standardized error in the form similar to a standard Z value, with an expected value of 0, an acceptable interval from -2 to +2 (Z values higher than +2 report a misfit in noisy response patterns and Z values below -2 correspond to a deterministic or Guttman pattern), and a mean square error MNSQ, having an expected value of 1 although some degree of variation of 1 is usually expected, which should not be more than 1.5 to a noisy misfit.

Results

To answer the three research questions raised above, Guttman and Rasch psychometric models were utilized to test any probable implicational scaling and reproducibility underlying the data of the present measure.

Analysis One: Guttman Scaling

To answer the first question of this study, "is there any implicational scaling, in Guttman's sense, among reading questions through Guttman Scaling?", the whole response data were subjected to Antropoc using minimized error model. Table 2 shows the results.

Table 2

Guttman Results for the SBRTa, SBRTb, and IELTS Exemplar

	Crep	MMrep	Scalability	Marginal Errors	Scale Errors	Number of Students	Number of Items
IELTS	0.766	0.666	0.297	6042	4246	503	36

Note: Crept = Coefficient of reproducibility index; MMrep = a minimal marginal reproducibility index.

As is indicated in Table 2, the reproducibility coefficients of all tests were below 80% (Crep= 0.766, MMrep= 0.666, scalability= 0.297), meaning that the data in this study do not provide an implicational scaling in a deterministic way. Moreover, the low coefficient index of reproducibility indicates that the data in this study do not show being reproducible for the IELTS items. Also, the low coefficient index of scalability indicates that the errors are not randomly distributed but are meaningful and therefore, reading data do not form an implicational scale.

Analysis Two: Rasch model

To answer the second and third questions, "*is there any implicational scaling among reading questions though Rasch model?*", and "*if there emerges any implicational scaling of reading scale, is the hierarchical ordering of reading data predictable by reading theory?"* the data were subjected to Rasch model using Winsteps. Table 3 reports the results.

Table 3

Rasch Model on the IELTS exemplar (1994)

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	SCORE CORR.	ITEMS*	
1	168	501	.15	.10	.86	-3.4	.79	-2.9	.50	V1	
2	194	501	-.12	.10	.91	-2.4	.94	-.9	.47	V2	
3	215	501	-.33	.10	.82	-5.2	.79	-3.6	.55	V3	
4	195	501	-.13	.10	.82	-5.0	.78	-3.5	.55	V4	
5	50	501	1.87	.16	1.10	.8	1.35	1.5	.14	V5	

6	269	501	-.86	.10	.96	-1.1	.94	-1.1	.46	V6
7	276	501	-.93	.10	.88	-3.3	.82	-3.5	.53	V7
8	301	501	-1.18	.10	.84	-4.3	.77	-4.3	.57	V8
9	175	501	.07	.10	1.08	1.9	1.10	1.3	.32	V9
10	180	501	.02	.10	1.01	.4	1.01	.2	.38	V10
11	113	501	.79	.11	1.18	2.9	1.26	2.1	.18	V11<
12	206	501	-.24	.10	.96	-1.0	.90	-1.6	.44	V12
13	98	501	1.00	.12	.97	-.5	1.04	.3	.34	V13
14	223	501	-.41	.10	.99	-.3	.95	-.8	.43	V14
15	302	501	-1.19	.10	.98	-.5	.99	-.2	.44	V15
16	306	501	-1.23	.10	.98	-.4	.97	-.5	.45	V16
17	279	501	-.96	.10	.90	-2.7	.85	-2.8	.51	V17
18	83	501	1.22	.13	.99	-.1	1.08	.5	.29	V18
19	135	501	.52	.11	.93	-1.4	1.03	.3	.41	V19
20	176	501	.06	.10	.93	-1.9	.92	-1.1	.44	V20
21	232	501	-.50	.10	1.06	1.7	1.08	1.4	.36	V21
22	307	501	-1.24	.10	1.03	.8	1.04	.6	.40	V22
23	132	501	.55	.11	1.18	3.4	1.37	3.3	.18	V23<
24	139	501	.47	.11	.98	-.5	.91	-1.0	.39	V24
25	63	501	1.58	.14	.93	-.8	.92	-.5	.33	V25
26	137	501	.49	.11	.99	-.2	1.14	1.4	.36	V26
27	210	501	-.28	.10	.91	-2.6	.84	-2.7	.49	V27
28	268	501	-.85	.10	.91	-2.4	.87	-2.5	.50	V28
29	75	501	1.36	.13	1.11	1.3	2.04	5.0	.10	V29<
30	329	501	-1.47	.10	.97	-.6	.96	-.5	.45	V30
31	158	501	.25	.10	1.22	4.6	1.38	4.0	.18	V31<
32	140	501	.46	.11	1.27	5.1	1.60	5.3	.11	V32<
33	103	500	.92	.12	1.12	1.8	1.38	2.7	.20	V33<
34	242	501	-.59	.10	1.03	.7	.97	-.5	.41	V34
35	200	501	-.18	.10	1.01	.3	1.00	.1	.39	V35
36	105	501	.90	.12	1.12	1.9	1.32	2.4	.20	V36<
MEAN	188.	501.	.00	.11	1.00	-.4	1.06	-.1		
S.D.	77.	0.	.85	.01	.11	2.4	.26	2.4		

(< indicates a misfit of the item)

Almost all the items in Table 3 show acceptable fit indices (Z < 1.4). Only seven items have misfit indices like V29 (outfit MNSQ= 2.04, Z =5), and V32 (outfit MNSQ= 1.60, Z = 5.3). In the IELTS exemplar test, V29 and V32 seem to theoretically measure the ability to understand the function of cohesive devices and the ability to infer the implicitly stated information respectively.

To test reproducibility in the reading data under the study, the IRT person map of items in the measure used in this study were also examined. According to Hessels (personal communication, 2000, 11th September on Rasch list) "if the items generally do not follow the pattern of development as [one] expected, this means that the order of the developmental levels are not as postulated in [one's] theory. This would be [the] test of reproducibility" (p. Internet page).

Figure 2 shows the Wright (Person) map of the results, demonstrating that various item types in the IELTS do not occur in a particular hierarchical order;

the item classifications are mixed. According to test instructions of the IELTS exemplar as shown in Table 1, items V15, V14, V16, V17, V18, V19, and V20 were assumed to measure the ability to understand the main idea of the text, but some of them occurred above (are harder than) other items (for example, V18 appears above V19), and some of them occurred below (are easier than) other items (for example, V17 occurs below V19). Moreover, some items appeared in unexpected orders as V32 (inference) occurred below (is easier than) V5 (scanning or information search) and V11 (the ability to understand syntax). Therefore, the reading data in these three tests do not conform to the reproducibility assumption underlying the reading test used in this study.

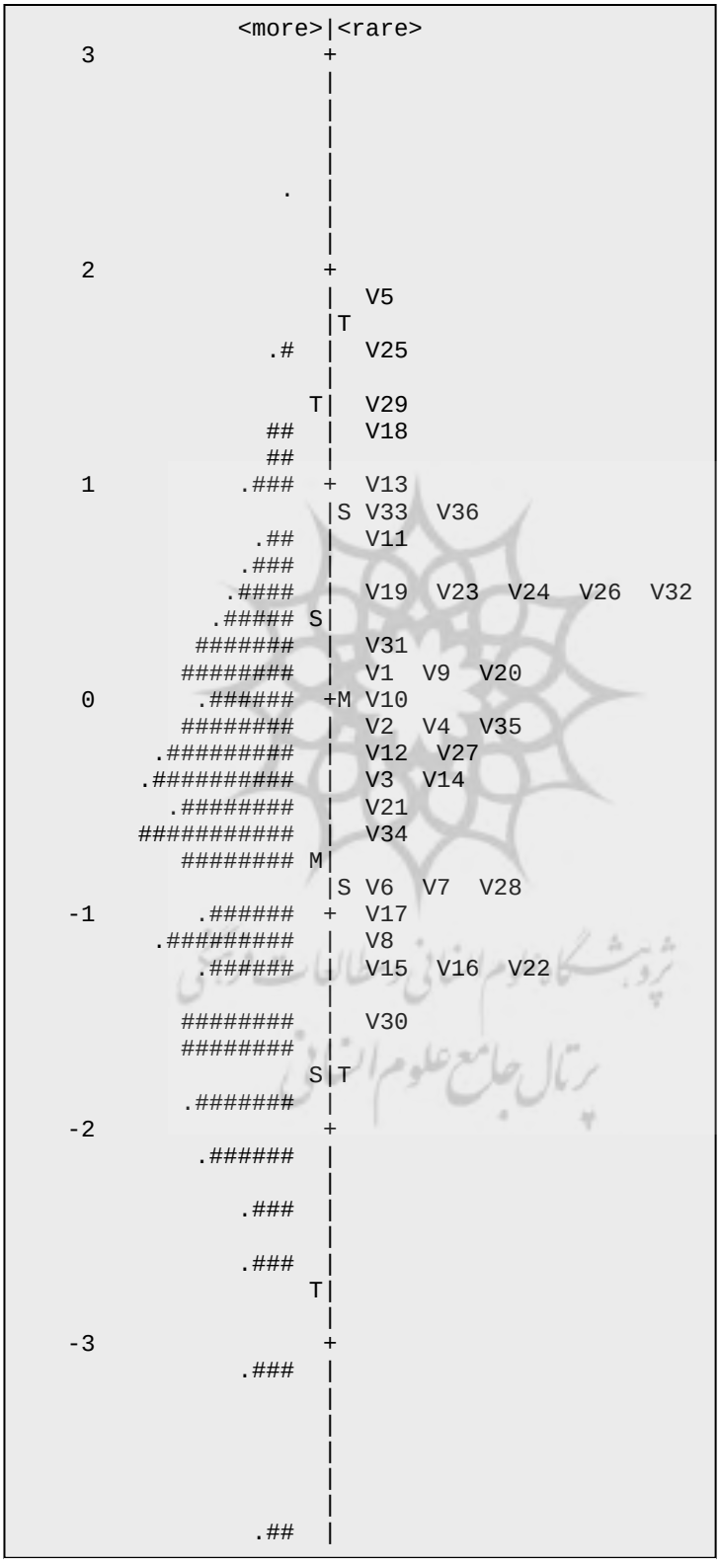

Figure 2. Wright Map of the IELTS Exemplar Items

These results may also call into question the additive nature of the items that is usually assumed when scoring. In scoring a reading test, almost always no weighting is applied or followed. Usually, items in a test are selected through item difficulty measures coming from an IRT calibration. This implies that the items are inherently reproducible in nature, in the sense that from the total score of these items, one can understand about the students' position on the reading scale. In other words, it is supposed that students with lower score have less reading ability than students with higher scores, where "reading ability" is organized in hierarchical levels in the sense of a learning framework or taxonomy. Therefore, the present study did not support the assumption of reproducibility and additivity because there is no one to one correspondence among the ability level supposed to be measured by an item and its difficulty (Daftarifard & Lange, 2009).

Moreover, The Test Design Line (TDL) proposed by Tristan and Vidal (2007) shows a very clear pattern of the items (Figure 3), following very closely the theoretical design of the test, with a Mean Absolute Difference (Mad=0.11) well below the quarter of logit thumb´s rule to accept that the test provides a reasonable measurement scale.

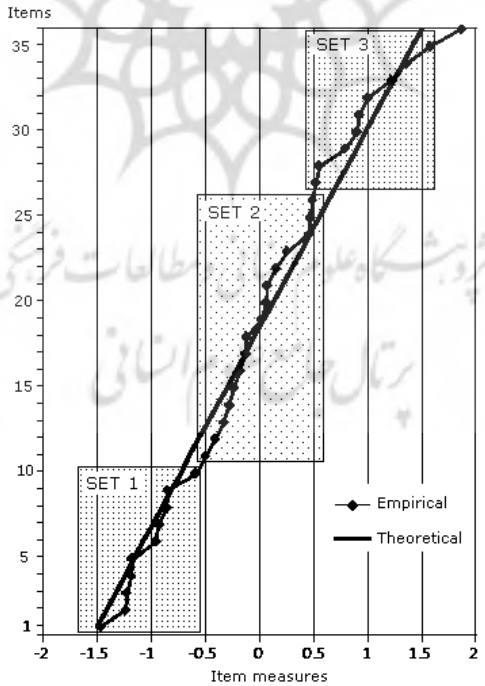

Figure 3. Test design line of the IELTS Exemplar.

With the TDL, it is feasible to define sets of items (or anchor scales) in terms of difficulty, and provide the set to a group of experts who may define constructs as proposed by Beaton and Allen (1992) and Beaton and Johnson (1992). Three sets may be chosen, for instance, in one logit ranks (See Table 4).

Table 4
Logit Ranks for the IETLS

Set 1, from -1.5 to -0.5		Set 2, from -0.5 to +0.5		Set 3 from +0.5 to +1.5	
ITEM	Difficulty	ITEM	Difficulty	ITEM	Difficulty
V30	-1.47	V21	-0.5	V19	0.52
V22	-1.24	V14	-0.41	V23	0.55
V16	-1.23	V3	-0.33	V11	0.79
V15	-1.19	V27	-0.28	V36	0.9
V8	-1.18	V12	-0.24	V33	0.92
V17	-0.96	V35	-0.18	V13	1
V7	-0.93	V4	-0.13	V18	1.22
V6	-0.86	V2	-0.12	V29	1.36
V28	-0.85	V10	0.02	V25	1.58
V34	-0.59	V20	0.06		
		V9	0.07		
		V1	0.15		
		V31	0.25		
		V32	0.46		
		V24	0.47		
		V26	0.49		

This way of judging can identify the constructs that are measuring the items. In this case, instead of organizing items as factual or procedural, remembering or analyzing, Judges may identify procedures such as "Basic" through "Proficient" levels. The reason is that categories like factual or procedural and the like are taxonomical categories that are involved in a different dimension. Probably, the items in Set 1 (See Table 4) are focusing on simple sentences with direct meaning, whereas the items in Set 2 concern paragraphs with compound sentences and dangling modifiers. Or perhaps the items in Set 1 measure learners' grammatical competency whereas the items in Set 3 are focused on a strategic competency to understand complete texts in specific contexts. Even, the judges may identify that Set 1 may be answered by non-expert students even if they have a low level of understanding of the language, in very concrete aspects of daily topics using very common words

whereas Set 3 may refer to sentences highly related to a abstractions and diverse contexts, including dexterities to handle metaphors and multiple meanings according to the context, with uncommon words. This analysis will define a hierarchical complexity description coherent with the difficulty of the tasks.

Discussion

It is common among reading specialists to divide reading ability into different layers of cognition (Grabe, 2009, 1997, 1991; Alderson, 2000) such that hypothetically labeled lower-order abilities are assumed to be followed by higher-order ones (Alderson, 1991). The concept of unidimensionality of reading comprehension (Mathews, 1990; Weir, Hughes, & Porter, 1990; Hillocks & Ludlow, 1984) leads scholars to believe that there might be a reproducible nature in latent trait of reading comprehension so that one can understand about the respondents' positions by looking at their scores on a test.

Moreover, mathematically speaking, hierarchy is defined in terms of difficulty. Urquhart and Weir (1998) mentioned that item difficulty is the criterion by which items are located on a hierarchical continuum, which starts with the simplest questions and ends in the most difficult one. Elsewhere, Hajipurnezhad (2001) stated that, in his study, "respondents [judges and students] recognize a parallel between perceived complexity and perceived factuality/ inferentiality" (Internet Page). The hierarchy assumption is so appealing that usually test developers try to calibrate the items in terms of item difficulty. Because of the unidimensionality assumption underlying the reading measures, item difficulty seems to be the best criterion based on which one chose the items. However, most of the time, difficult items belong to lower-order abilities than the easier ones (Weir & Porter, 1996), items that also seem to contribute less to reading ability (McNamara, 1996). This is in contradiction to the reproducibility assumption underlying the construct. As Weir and Porter (1996) mentioned, the main reason for applying the reproducibility assumption into test constructions might be 'the practical expediency rather than . . . a principled view of unidimensionality' (p. 1).

It is necessary to identify that taxonomies or learning frameworks that concern with the complexity of the mental procedure necessary to perform a

specific task, whereas difficulty is defined in terms of the frequency of correct responses obtained from a set of persons of a focal group. Therefore, it is incorrect to refer to the simpler levels of taxonomy as the easier tasks, or say that the highest taxonomical levels are also the hardest. Figure 4 shows what was expected: difficulty in the same axis or dimension than taxonomical complexity, whereas Figure 5 shows what is really happening: the taxonomical categories correspond to a different dimension than difficulty of the items.

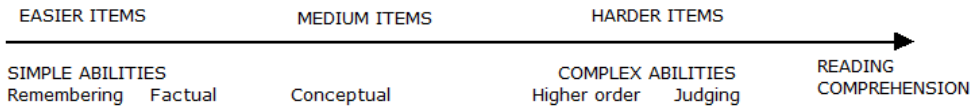

Figure 4. Complexity and difficulty as a single dimension

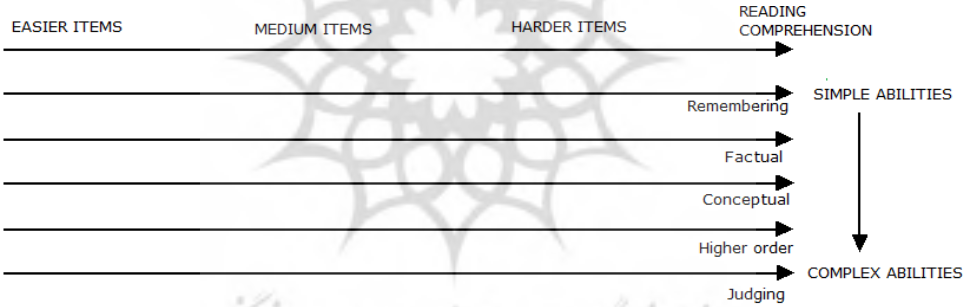

Figure 5. Complexity and difficulty as two different dimensions.

Tristan (1998) and Tristan and Molgado (2006) explain the difference between complexity and difficulty. Complexity is an attribute of the mental process (it can also be a psychomotor, conative or affective process) that the person has to perform to develop or show a competency or ability. In short, complexity is related to a theory or framework and hence it is conceptually defined and its empirical verification must consider other procedures not linked to the frequency of responses of the persons.

Complexity involves three characteristics in its definition:

A)It is defined a priori. A group of experts define the construct of the competency or ability and establish the level according to a taxonomy or learning framework.

B)It is subjective. The group of experts do need to agree in the identification of the level of complexity and there are multiple possible ways to classify an item, depending on the context of the task, the expertise of the judge in the field and the appreciation of the way the person must perform the task, among other sources of subjectivity. A technique to establish the agreement among judges is needed to classify the items of a test.

C)It cannot be calculated. Once the judges arrive to an agreement, they may define a number, but there is not a formula or algorithm to calculate the complexity level for a specific item.

Difficulty is a quantity associated to the frequency of correct or wrong responses to a specific task explored by an item administered to a focal group of persons. Difficulty has to be obtained from a pilot test or any other source of responses from the focal group. Following Tristan and Molgado (2006), difficulty involves three properties:

A) It is defined a posteriori. The item must be administered to a group of persons and their responses are needed to calculate the difficulty. If the item has not been administered it may not be calibrated in its difficulty.

B) It is objective. Item difficulty can be defined following a specific model (in classical test theory, IRT, Rasch and so forth), it does not depend on the opinion of judges proposing difficulty as a definition or by consensus. Objectivity is a main attribute of measurement.

C) It can be calculated. Depending on the model, difficulty may be calculated using the proportion of correct responses, the probability of response of a person facing the item and so forth. It is clear that there is a formula or algorithm to calculate the difficulty of an item.

The three characteristics or properties of complexity and difficulty are complementary or even opposite. That is why test designers cannot get a good result trying to make equivalent difficulty and complexity. Difficulty and complexity are not synonymous but two different and independent dimensions. This different meaning does not represent a problem or contradiction once their definitions have been done and their attributes or properties have been identified.

Figure 6 shows the complete framework of a test, with, at least, three different dimensions: complexity, difficulty (already discussed) and the content dimension appropriate to a specific domain of competencies, abilities or knowledge (English as a Second Language in this case, Mathematical topics, Science areas or History chapters in other tests).

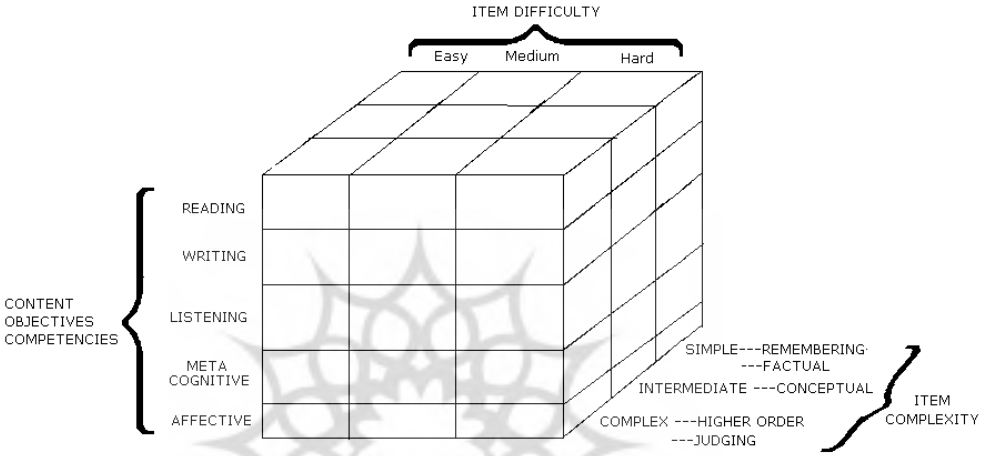

Figure 6. Framework for the three minimal dimensions contained in a test.

It has to be noticed that among these three basic dimensions, having the Guttman's scalability concerns only with difficulty dimension. The Rasch model provides a satisfactory explanation in terms of the probability of response to an item depending on the person's measure of ability that runs in the same axis or dimension than difficulty; hence, it is possible to obtain a measure for each competency or individual ability such as remembering, conceptual handling or higher order thinking. Also, it is possible to have different measures of the ability of a person on macro-abilities such as reading, writing or listening for ESL.

Conclusion

The purpose of the present research was to empirically probe the possibility of implicational scaling assumption usually assigned to reading questions both in teaching classrooms and assessment. The results of this study suggested that, unlike what Hillocks and Ludlow (1984) suggested, it is unrealistic to expect the items in a reading test to form a deterministic (Guttman) hierarchy in which

one skill always precedes other ones; rather, it appeared that such hierarchies are better represented by a probabilistic approach. However, the data do not support the reproducibility nature of reading measures; it is not possible to model in a single dimension the taxonomical ordering as equivalent to the item difficulty. In fact, it is possible to obtain an empirical hierarchy of the sub-scales of the learning framework. The interpretation of taxonomical complexity and item difficulty in two distinct dimensions makes clear the implications of the ability needed to perform a task and the mental (or psychomotor) process involved in a response.

Various item types contained in the measures of this study do not occur in a particular and predictable order; in fact items are mixed in their taxonomical classification. Therefore, it is not possible to predict the level of cognition of the item in terms of the respondents' total score. It can be concluded that complexity and difficulty are two distinct dimensions.

It has been shown that the difficulty of an item does not reflect a one-to-one relationship with the framework categories: a hard item does not guarantee that it is measuring a higher-order ability like inferential process; on the contrary, it might measure lower-order ability like syntax or scanning. This finding has been previously reported in McNamara (1996) and certainly explains why it cannot permit to get a consensus among judges trying to match items in taxonomy and difficulty, as Alderson tried to do (1990). The results confirm that theoretical abilities must be represented at least in a three dimensional space: the content of the academic or professional matter (organized in topics, subjects, chapters, & areas), the difficulty of the items (from easy to hard) and the cognition process of the persons (from simple to complex or from low level to higher order). This conclusion may be translated in other fields (Rense, 2000, personal communication); for instance, in mathematics: subtraction is cognitively a less demanding operation than division but in both operations, there are items measuring easier to harder tasks. That is why "5/1" is just as easy as "35-44", even though division is the hardest operation.

Often it is claimed that reading comprehension test items are measuring reading ability; however, almost always no research indicates which aspects of reading are measured. If such assumptions are not observed, the inferences will be misleading.

References

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language, 6*, 425-438.

Alderson, J. C. (1991). Language testing in the 1990's: How far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing: Anthology series* (pp.1-26). Singapore: SEAMEO Regional Language Center.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Allerson, S., & Grabe, W. (1986). Reading Assessment. In F. Dubin & D. E. Eskey & W. Grabe (Eds.), *Teaching second language reading for academic purposes* (pp.161-181). NY: Addison-Wesley Publishing Company, Inc.

Anderson, R.W. (1978). An implicational model for second language research. *Language Learning, 28,* 221-278.

Bachman, L. F., (1995). *Fundamental considerations in language testing*. London: Oxford University Press.

Baker, D. (1989). *Language Testing*. London: Edward Arnold.

Bart, W. M., & Krus, D. J. (1973). An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement, 33*, 291-300.

Barrett, T.C. (1968). The Barrett Taxonomy of the cognitive and affective dimensions of reading comprehension. In H. M. Robinson, *Innovation and change in reading instruction* (pp.1-30). NY: the National Society for the Study of Education

Baudoin, E. M, Bober, E. S., Clarke, M. A., Dobson, B. K., & Silberstein, S. (1977). *Reader's Choice: A reading skills textbook for students of English as a second language*. Michigan: The University of Michigan Press.

Beaton, A.E., & Allen, N.L. (1992). Interpreting scales through scale anchoring. Journal of Educational Statistics, *(17),* 191-204.

Beaton, A.R., & Jonson, E.G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement, (29)*2, 163-175

Biggs, J.B., & Collis, R.E. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Bloom, B.S. (1957). *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I, cognitive domain.* New York-Toronto: Longmans-Green.

Bloom, B. S. (1994). Reflections on the development and use of the taxonomy. In L.W. Anderson & L.A. Sosniak (Eds.), *Bloom's taxonomy a forty-year retrospective* (pp.1-8). NY: the National Society for the Study of Education.

Carrell, P. L. (1987). Readability in ESL. *Reading in a Foreign Language, 4*, 21-40.

Champeau De Lopez, C.L., Marchi, G. B., & Coyle, M. E. A. (1997, April-June). A taxonomy evaluating reading comprehension in EFL. *Forum, 35(2),* 30, from http://www.exchanges.state.gov/forum/vols/vol35/no2/p30.htm

Cazden, C.B. (1971). Evaluation of learning in preeschool education : Early language development. In Bloom B., Hasting J. & Madaus G. (Eds.), *Handbook of formative and summative evaluation of student learning* (pp. 345-398). NY: McGraw Hill.

Daftarifard, P. (2002). *Scalability and divisibility of the reading comprehension ability*. Unpublished master's thesis, Iran University of Science and Technology, Tehran.

Foley, J.J. (1971) *Evaluation of learning in writing.* In Bloom B., Hasting J. & Madaus G. (Eds.) *Handbook of formative and summative evaluation of student learning* (pp. 767-814). NY: McGraw Hill.

Glass, G.V. (N/D) *Building tests that make students think.* In Test and grades. Chap.1. Available in internet: http://glass.ed.asu.edu/TG/chp1.htm

Grabe, W. (1986). The Transition from Theory to Practice in Teaching Reading. In F. Dubin, D. E. Eskey, & W. Grabe (Eds.), *Teaching second language reading for academic purposes* (pp. 25-48). NY: Addison-Wesley Publishing Company.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*, 375-406.

Grabe, W. (1997). *Reading research and its implications for reading assessment.* LTRC paper.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Gray, W. S. (1960). The major aspects of reading. In H. Robinson (Ed.), *Sequential development of reading abilities* (pp. 8-24). Chicago: Chicago University Press.

Guttman, L. L. (1974). The basis for scalogram analysis. In G. M. Maranell (Ed.), *Scaling: A sourcebook for Behavioral Scientists,* (pp. 142-171). NY: Aldine Publishing Company.

Hajipournezhad, G. R. (2001, Oct.). Reading complexity judgments: Episode 1. *Shiken: JALT Testing & Evaluation SIG Newsletter 5* (pp. 2 – 5) from http://www.jalt.org/test/haj_1.htm

Hillocks, G. JR., & Ludlow, L. H. (1984). A taxonomy of skills in reading and interpreting fiction. *American Educational Research Journal, 21*, 7-24.

Hatch E., & Farhady, H. (1981). *Research design and statistics for applied linguistics.* LA: University of California.

Henning, G. H. (1977). A developmental analysis of errors of adult Iranian students of English as a foreign language. *Language learning, 28*, 387-397.

Hulstijn, J., (1997). Mnemonic methods in foreign language vocabulary learning: Theoretical considerations and pedagogical implications. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp.203-224). Cambridge: Cambridge University Press.

Hyltenstam, K. (1977). Implicational patterns in inter-language syntax variation. *Language Learning, 27*, 383-411.

Jensen, L. (1986). Advanced reading skills in a comprehensive course. In F. Dubin, D. E. Eskey, & W. Grabe (Eds.), *Teaching second language reading for academic purposes* (pp. 103-124). CA: Addison-Wesley Publishing Company, Inc.

Karami, F. (2000). *The effect of task variation on the reading comprehension ability of the learners*. Unpublished master's thesis, Iran University of Science and Technology, Tehran, Iran.

Kral, T. (1995). *Selected approaches from the creative English teaching forum 1989-93*. United States Department of State, EUA

Ludlow, L. H., & Hillocks, G. Jr. (1985). Psychometric considerations in the analysis of reading skill hierarchies. *Journal of Experimental Education, 54*, 15-21

Maranell, M. G. (1974). Introduction. In G. M. Maranell (Ed.), *Scaling: A sourcebook for Behavioral Scientists* (pp. xi-xix). Chicago: Aldine Publishing Company.

Matthews, M. (1990). Skill taxonomies and problems for the testing of reading. *Reading in a Foreign Language, 7*, 511-517.

McNamara, T. (1996). *Measuring Second Language Performance*. NY: Addison Wesley Longman.

Moore W.J., & Kennedy L.D. (1971). Evaluation of learning in the language arts. In B. Bloom, J. Hasting, & G. Madaus(Eds.), *Handbook of formative and summative evaluation of student learning* (pp. 399-446). NY: McGraw Hill.

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. Hong Kong: Macmillan Publishers Limited.

Pretorius, E. J. (2000). Reading and the Unisa student: Is academic performance related to reading ability? From http://www.unisa.ac.za/dept/bmi/resrep00/arts/linguist/publicat.html

Roberts, N. (1974). Further verification of Bloom's taxonomy. Journal of Experimental Education, *45(1)*, 16-19.

Rost, D. H. (1993). Assessing different components of reading comprehension: Fact or fiction. *Language and Education, 6*, 79-91.

Stauffer, S. A. (1974). An overview of the contributions to scaling and scale theory. In G. M. Maranell (Ed.), *Scaling: A sourcebook for Behavioral Scientists,* (pp. 131-141). Chicago: Aldine Publishing Company.

Trimble, L. (1985) *English for science and technology: A discourse approach.* London: Cambridge University Press.

Tristan, L.A. (1998) *Test blueprint techniques* (in Spanish: Tablas de validez de contenido) Instituto de Evaluación e Ingeniería Avanzada. San Luis Potosí, Mexico.

Tristan, L.A. & Molgado, R. D. (2006) *Handbook of taxonomies* (in Spanish: Compendio de taxonomías. Clasificaciones para los aprendizajes de los

dominios educativos). Instituto de Evaluación e Ingeniería Avanzada. San Luis Potosí, Mexico.

Tristan, L.A. & Vidal, U.R. (2007) *Linear model to assess the scale´s validity of a test.* AERA Meeting, session: "New Developments in Measurement Thinking", SIG-Rasch Measurement. Available through ERIC: ED501232.

Urquhart , A.H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. NY: Longman

Valette, R. M. (1971) *Evaluation of learning in a second language.* In B. Bloom, J. Hasting & G. Madaus (Eds.), *Handbook of formative and summative evaluation of student learning* (pp. 815-854). NY: McGraw Hill.

Weir, C.J., Hughes, A., & Porter, D. (1990). Reading skills: Hierarchies, implicational relationships and identifiability. *Reading in a Foreign Language, 7*, 505-510.

Weir C. J., & Porter D. (1996). The multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language, 10*, 1-19.

Biodata

Parisa Daftarifard (PhD) is an assistant professor in TEFL program in Iran, which is affiliated with the department of Persian Literature and Foreign Languages at the IAU (South Tehran Branch). She has held different workshops on statistical programs such as SPSS, Rasch Model, and LISREL. She has published a considerable number of articles related to Assessment, Second Language Acquisition, and Reading Comprehension. Also, she has presented a number of articles in different international conferences. Her main areas of interest are assessment, reading comprehension, first language acquisition, second language acquisition, and cognitive development.

Agustin Tristan Lopez (PhD) has got his PhD in Mechanics of Materias from National School of Bridges and Roads of Paris. He has published a numerous number of articles in Assessment and Evaluations, Among his most recent publications are: Compendium of Taxonomies; Manual of Correlation Formulas and Test Quality Standards Objectives. He gives technical support from the design of a test to item analysis. He produces the family of KALT programs (including item banking, classical item analysis, on-line testing including Rasch model and criterion-referenced test and item analysis) and offers technical support and training (test design, Rasch analysis, item design and analysis). He is the technical Editor of the journal "Advances in Measurement" and Scientific Editor in

the magazine. He is currently the Director of the Institute of Evaluation and Engineering Advanced, S.C. Also he is a Consultant in Rasch, logistic models and classical models for educational evaluation and biostatistics at the Institute of Objective Measurement, Chicago, USA; NAFEMS - ISO, Great Britain; KNOW.