Is ESP (EAP) a Composite or Simple Construct?
Spec- Reverse Engineering of a Specific Purpose Language Test

Corresponding Author: Razieh Rabbani Yekta
PhD Candidate in TEFL, Department of Foreign Languages
University of Isfahan, Iran
E-Mail: r_ryekta@yahoo.com

Second Author: Mansoor Tavakoli
Assistant Professor in TEFL, Department of Foreign Languages
University of Isfahan, Iran

Third Author: Abbas Eslami Rasekh
Assistant Professor in TEFL, Department of Foreign Languages
University of Isfahan, Iran

Abstract

The purpose of the present study was to detect whether items assembled in one subset in the complex tests contributed only to a simple structure or to a complex one and whether through some modifications in item assembly sub- layer of test architecture, the upgraded item -sets could represent the true dimensionality of the test, hence meaningful scores for each set. To this end a historical reverse engineering (HRE) was conducted on a joint test of ESP/ EGP on its three different versions given in 2000, 2005, and 2010 for two non-English majors, Accounting and Statistics. The research design involved a group project, participants performing HRE on the sample items of the Master degree ESP / EGP entrance exams to reach to their item specifications. From there, item assembly sub layers of these tests were critiqued in terms of their match with the intended dimensionality of the concerned tests and participants proposed some suggestions for controlling the degree of match for obtaining more meaningful part-scores out of each set .

Introduction

In many testing situations, there is no exact correspondence between the dimensions of the test and the structure of its item clusters. In most of the cases, as described by Gierl, Leighton, and Tan (2005), when the data are unidimensional, clusters of items will be found that are not homogeneous (measuring a single trait). On the other hand, there is a controversy over measuring different latent abilities in two or more independent tests or the development of a single test with composite structure (with each item set measuring more than one latent ability) for tapping those abilities (Torre &Patz, 2002; Johnson & Carlson, 1994). This is while, a number of empirical investigations have been conducted which have showed that the estimation

method for the correlated abilities yields more efficient results when it is based on the simple structure than composite structure items (ibid).

In the education system in Iran, the entrance exam for the admission of prospective students into post graduate programs is a battery of about 8 different subtests (including one English subtest plus 4 to 7 knowledge subtests of different content courses) where the raw scores of individual subtests are averaged to form an overall test battery mean and participants are ranked based on a composite percentile. As to the language subtest, after experiencing several years of upgrading, our test developers in Iran have finally reached a fixed framework for the language subtest. For almost all of the fields, this subtest consists of two parts: general English and specialized English. The first part starts with 10 vocabulary test items and continues with a cloze test of grammar with a text of non specialized content. The specialized part is composed of 3 field specific reading test-lets, each with about 5 multiple choice (dichotomously scored) items. Regarding the fact that these two parts have been examined in two independent tests for the counterpart exam for entering the PhD program, it would be of no surprise, if one catches the very challenge by asking the following questions: does such a domain involve a single multidimensional construct or multiple constructs that are correlated?

So far, several statistical approaches have been proposed in the literature to improve proficiency estimation at the sub-score level for the tests primarily designed to order individuals by a total score (Bock, Thissen, and Zimowksi,1997; Yen, 1987; Wainer, et al. 2001; Yao & Boughton, 2005). The present study, however, is a qualitative extension of those researches. The purpose was to detect whether items assembled in one subset in such complex tests contributed only to a simple structure or to a complex one and whether through some modifications in item assembly sub- layer of test architecture, the upgraded item -sets could represent the true dimensionality of the test, hence meaningful part-scores for each set. The following general questions guide the process of this research:

1. What are the item specifications on the bases of which each item set was assembled?
2. Do the items clustered in one item set come from the same Spec? In other words, is an item in a special cluster actually in the same skill domain as other items in that set or should actually be reverse engineered into a different Spec?
3. With a hindsight to the versions administered earlier, does the number of item clusters in each subtest of the last version correspond to the number of the substantively separate dimensions in the test? If not, what modifications are needed in assembling the items so that the upgraded clusters will be interpretable both at within and between-cluster level?

Regarding this questions, the present study was designed to be a historical reverse engineering of three versions of the concerned test which were given in three different years. It is a team work because, according to Davidson (2003), Spec reverse engineering should be consensus based; it is historical, because the aim was to see how these complex tests have changed over times in terms of the Spec-based assembling of their items and how the Specs for different sets have brought together the complex and simple structure items in each version (for more information about different types of reverse engineering, see Davidson and Lynch, 2002).

Participants

Three groups of EFL teachers were selected for this study. They were sampled from among a number of teachers in Isfahan University as well as Isfahan University of Technology based on their availability at the time of conducting this project. One group were PhD candidates of TEFL (n=4) with about two to three years of EFL classroom experience who had already passed some special courses both in language assessment and ESP. A second group consisted of 2 assistant professors in TEFL who had been teaching language assessment on average for 5 years and a third group was made up of EGP (n=4) and ESP (n=2) teachers. Table 1 shows the teacher participants background and experience by group.

Table 1. Teacher participants training and teaching background

Groups	Experience Details		
	EFL teaching	**Assessment –related Teaching or training**	**ESP –related Teaching or training**
G1. PhD candidate in TEFL (n=4)	2 to 3 years	Three semesters training in language testing	Two semesters training in ESP
G2. Assistant Professors (n=2)	About 8 years	On average for 5 years teaching experience in language assessment	Two semesters training in ESP
G3. EGP / ESP teachers (n=6)	About 8 years	Three semesters training in language assessment	On average for 6 years ESP/EGP teaching experience

Methodology

This study was designed as a case study with evaluative features (for more information regarding this research design see: Nunan, 1992; Peara-Hernandez, 2010). The process comprised four general phases which followed approximately the steps in the workshop model used by Scott Walters (2010) in his standard reverse engineering project. In phase one, teacher participants were introduced to Davidson and Lynch (2002)'s model of test specifications adapted from Popham (1978). They were also given some kind of worksheets on which they analyzed the items in different sets across different years of administrations. The only tools used by the participants in their item analysis were: Nation (2001, 2003)'s Vocabulary-Profiler for determining the words frequency in vocabulary item sets, Purpura (2004)'s coding scheme for

grammar subsection, and Weir et al. (2000) parameters of EAP reading and Brown (1988)'s item types for specialized and general reading test-lets. The results of the analysis are presented in tables 2 to 5 for different sub components of the joint EGP /ESP tests across three different years.

Table 2. Frequency analysis of the words selected as the options in the subtests of General vocabulary Knowledge administered in three different years for two non-English fields of Study

Year ------ *field* / *frequency*	2000		2005	2010
	Statistics	Accounting	Statistics / Accounting	Statistics / Accounting
1st-1000 words	53.85%	37.5%	76.92%	7.5 %
2nd -1000 words	30.77%	37.5%	15.38%	5%
Academic Vocab	15.38%	16.67%	5.13%	30%
*Off-list Vocab	0%	8.33%	2.57%	57.5%

*Off-list vocabulary refers to the technical or subject specific words.

In 2000 version, general vocabulary subtests were different for different fields of study, but, it seems that multiple forms of a same test were used (as the percentages of words selected from each frequency level showed a relatively similar distribution across both fields). In 2005 and 2010 versions, however, such a differentiation was stopped.

Another interesting observation by the participants here was the test developers' completely reverse trends in their word selection patterns for 2000 and 2010 versions. While in 2000, most of the word options were from the common core of high frequency words, in 2010, this was the technical vocabulary knowledge of different *ranges* (Read, 2000) which was mostly tapped by the items.

The only observation that is not reported in the vocabulary section was the existence of a separate cluster of items measuring examinees' conceptual knowledge and terminology of their specific field. These special clusters which were only assembled in 2000 version comprised one third of the whole items and so the part-score obtained from these sub components could have a strong bearing on the total summed score.

Table 3. Average percentage of teacher participants' allocation of the items in cloze and error identification (EI) Sub- sections to the components of grammatical knowledge based on Purpura (2004)'s model

	Error Identification		Cloze	
Year / ***field*** / ***components***	2000		2005	2010
	Statistics	Accounting	Statistics / Accounting	Statistics / Accounting
Lexical form (LFORM)	% 66.6	% 40	% 25	% 60
Lexical meaning (LMEAN)	-------	------	% 50	----------
Morphosyntactic form (MSFORM)	% 33.4	% 60	% 25	% 40
Cohesive form (CFORM)	-------	------	------	% 40
Cohesive meaning (CMEAN)	-------	------	------	% 40

In Table 3, again, some remarkable differences were reported in item assembling across different years. The first one is in the last two rows of the table where the participants did not code any of the items as the measures of Cohesive form and meaning for 2000 and 2005 versions while for 2010 version these two components had a reading of 40%. This non-existence of the items measuring the textual understanding was reported by the participants to be the consequence of the item types differences, i.e., sentence level error identification vs. cloze test.

Another interesting issue was that the percentages in 2010 column add up to more than 100%. This was explained by the participants as an indication of the items in the cloze subsection which measure more than one component at the same time.

For analyzing the specialized reading test-lets, at first, participants rated the passages on the scheme of textual parameters developed by Khalifa and Weir (2009). The results of this analysis were reported in table 4 below.

Table 4. Qualitative and Quantitative features of the texts selected as the stimulus in specialized reading test-lets

	2000		**2005**				**2010**				
Field	**ST**	**AC**	**ST**	**AC**			**ST**			**AC**	
Textual Features	**P1**	**P1**	**P1**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	**P1**	**P2**

Grammatical	*Length (No. of words)*		340	237	263	190	207	172	192	201	300	261	274
	Vocabulary (lexical density)		0.58	0.56	0.54	0.54	0.62	0.59	0.62	0.45	0.59	0.63	0.59
	Readability (Flesch Reading Ease)		45.3	3.4	28.1	21.7	11.7	34.0	18.3	54.5	33.0	5.0	25.5
Discourse	*Genre*		Textbook	Textbook	Magazine art	Textbook	Report	Textbook	Journal art	Magazine art	Textbook	News Art	Textbook
	Rhetorical task		Exposition	Exposition	Exposition	Exposition	Exposition	Exposition	Argument	Argument	Exposition	Exposition	Exposition
Content	*Subject area*		St	Ac	St	Ac	Ac	Ac	St	St	St	Ac	Ac
	Subject specificity	*% AWL words*	15.71	12.24	9.54	11.46	15.64	15.70	18.48	8.33	19.02	15.15	13.28
		% off-list words	9.14	7.59	12.60	6.77	7.58	4.65	7.61	3.43	3.93	15.53	7.75

After this initial analysis, participants in different groups were asked to analyze the test-let items separately in terms of the skills and strategies they tapped. For this part, Weir et al. (2000)'s parameters of EAP reading were used as a framework. But, as the results of a pilot study showed that the participants may have difficulty reaching consensus regarding Weir et al. categories, a modified version was used in the main study in which Weir's Explicitly Stated Main Idea (EXMI) was replaced by Kim's (2009) Reading for Literal Meaning; Syntax component was replaced by Rhetorical Function (Brown, 1988); and In Inferring Lexical Meaning category, a distinction was made between Sub-technical and technical vocabulary. Table 5 presents the modified version. Each cell of this table shows the number of items which tap a special skill or strategy by passage by field of study and year of administration.

Table 5. Pattern of assembling of the item types in different test-lets as marked by the participants

	2000	**2005**	**2010**

Field / Item Type		ST	AC	ST	AC			ST			AC	
		P1	P1	P1	P1	P2	P3	P1	P2	P3	P1	P2
Scanning				///////								
Skimming for the main idea					/	/		/	/		/	/
Understanding Factual information/literal meaning		////	////	////	///	//	/////	/		///	////	///
Vocabulary	Sub-technical			///			/	/			//	/
	technical					/					/	
Inference	propositional		/	/				//	//	/	/	/
	pragmatic								//			
Rhetorical Functions								/				
Total		4	5	15	4	4	6	6	5	4	9	6

After the initial observation, participants tried to compare the overall structure of test-lets in terms of the distribution of their item types. They also participated in a five-point semantic differential task to rate the amount of within-test-let dependency (to what extent answering a question within a test-let helps answering the next ones in the same test-let) and between-testlet dependency (to what extent answering the questions in one testlet helps answering the question in the follow-up test-lets or vice versa). They unanimously reported a kind of simple structure in2000 version with almost most of the items in both fields measuring a kind of surface understanding of the factual information and a wider distribution of the item types in 2005 and 2010 version with, again, the surface understanding of the text as the primary skill tapped. As to the semantic differential scale of dependency with 1 to 5 points between Dependent and Independent extremes, participants showed the overall 1 rating for within item dependency for all of the test-lets except for Statistics 2005 version where they rated the dependency of the items in the second test-let as 3 or 4. There was, also a case of between-testlet dependency in Accounting 2010 version where the first passage was selected to be a report of audit failure and the second passage was a definition of the auditing process. This content overlap between passages was explained by the participants to be a potential source of dependency between test-lets.

One interesting aspect which was not subsumed under the item analysis for reading component, was the general-specific differentiation in reading ability in 2000 and 2005 versions as opposed to the 2010 version; this was explained by the participants with regard to table 3. They reached to the consensus that the insertion of the separate test-lets which measure content-general reading ability in 2000 and 2005, could be regarded as an alternative to testing supra-sentential understanding in the cloze passage of general content in 2010 (see: table 3).

For the next step, after having a strong mind of the item features and item clustering in different versions as well as Spec writing principles, participants in different groups were asked to reverse

engineer the items Specs in 2010 version with a view toward the differences observed between this last version and the other two.

For this phase, given time constraint, items were assigned differentially across groups commensurate with teacher participants experience. Although according to Scott Walters (2010), this "differential item assignment . . . lessened an element of control", it facilitated the reverse engineering process where the items were more relevant to the respective participants situations. In this way, the vocabulary subsection of the concerned test was assigned to the PhD candidates (G1), cloze subsection to the assistant professors with experience in teaching language assessment, and specialized reading test-lets to the ESP/EGP teachers.

Figures 1 to 3 present the reverse engineered Specs crafted in different groups, each working on one subsection of the joint test of EGP/ESP, 2010 version.

GD: Examinees should recognize the vocabulary which best fits the provided context in the stem sentence. They will demonstrate a knowledge of vocabulary which ranges from common core to academic to technical level.

PA: Students will be given 10 stem sentences each with only a single blank which must be filled with a key or stimulus content word. The Key or stimulus must not necessarily be from the academic or sub-technical words, while distracters must be mostly technical or field specific. A few common core or high frequent words will also be included among options.

RA: Students will read the stems and the options for the first time and then recognize the meaning which completes the content of the sentence. Then they will try to select the form which relates to the meaning through rereading the stem or deleting the wrong options one by one.

FIGURE 1. PhD Candidates group (G1) reverse engineered Spec for the first sub-part of the test which was supposed to measure General vocabulary knowledge

Note. GD = General Description; PA = Prompt Attribute; RA = Response Attribute

GD: Examinees will demonstrate knowledge of the lexico-grammatical features both at sentential and supra-sentential level of discourse by filling the gaps with the appropriate options. At supra-sentential level, they must have an integration of textual competence and knowledge of morphosyntactic rules or lexical forms which decode the grammar.

PA: A short passage of sub-technical content will be given to the examinees from which some parts have been deleted selectively. The deleted parts will be selected in a way to include one or more of the following cases:

__ Multi-word phrases with co-occurrence restriction

__ Pronoun referents

__ Connectors

__ Verbs in special mood or voice

__ Choice of a particular part of the speech

RA: Examinees will read the passage and the options once and for the selection of the answer which makes the passage grammatically correct, they will reread either a few words around the blank or the previous sentences of the passage completely. For some items, both the words around the blank and the previous sentences must be considered as the clue.

FIGURE 2. Assistant professors (G2)' reverse engineered Spec for the Cloze subtest of the grammar

GD: Examinees will demonstrate their ability to read the texts of technical content. The specific reading skills on which examinees will demonstrate their ability are (in order of priority):

__ Understanding the literal meaning of the text

__Using explicit statements in the text to form an inference

__ Inferring the meaning of Academic and subject specific words and expressions

__ skimming the text for the main idea

__ Interpreting and evaluating the text using their own schemata

PA: Examinees will be given two to three content-specific passages of about 200–300 words, each followed by 4 to 9 multiple choice questions. Texts must be selected from among expository texts of the specialized textbooks or journals /newspapers articles. Items in each test-let must be developed in a way that answering one of them does not influence answering the other; they must tap different range of the reading skills and parameters with Understanding the Literal Meaning of sentences as the primary and understanding Implicit Meaning and vocabulary as the secondary level skills (within test-let imbalance of item types); at the same time, distribution of item types must not vary greatly between test-lets (between-testlet balance of item types)

RA: Examinees will read the text and use their own skills and strategies either at the global level or at the local level with the latter as the primary requirement for choosing the right answer. These skills include:

At Global Level (Reading between the lines)

__ inferring writer's intention based on what is stated in the text

__ interpreting the writer's message

At Local Level

__ using their knowledge of language to decode the meaning of the key words and relating them in a cohesive manner to get a surface meaning of the text at sentence and supra-sentential level.

FIGURE 3. ESP /EGP teachers (G3)' reverse engineered Spec for the Specialized Reading Test-lets

So far, the first question has been answered. To answer the second question, a third phase was designed in which different groups tried to examine their crafted Specs together.

One of the G1 Participants (G1P1) started the discussion from the Spec reported in FIGURE 1, while a participant from G2 (G2P1) approached the issue from a contrastive perspective by comparing the 2000 version assembling of common core vocabulary in an independent set partitioned from the field specific conceptual knowledge and terminology set. G3 participants also focused on the need for the differentiation of the common core and technical vocabulary in Spec writing. The discussion on FIGURE 1 Reverse engineered Spec was transcribed as follows:

G1P1: *All this does not put under one Spec.*
G2P1: *2000 version seems to be more construct relevant; there, we could have one Spec for common core and one for technical vocabulary.*
G2P2: *Distracters must have been of the same level of frequency as the key, and the key must have been selected from among common core or academic vocabulary.*
G3P1: *Technical vocabulary must be omitted from general vocabulary subsection and tested in another set.*
G3P2: *What is intended to be measured is general vocabulary knowledge; including vocabulary from different ranges is irrelevant.*

In whole, the consensus reached by the groups was the report of a kind of "item to objective incongruence" (Popham, 1984) for general vocabulary section.

As to the cloze test of grammar, because of so many differences in item assembling between three versions, participants put forward a contrastive oriented discussion on the issue:

G2P1: *In 2005 version, in addition to measuring conjunctions and relative pronouns in vocabulary subsection, we have: 1)a cloze test which measures both vocabulary and grammar at the sentence level, and 2) a reading test-let of a general content.*
G2P2: *So, it seems that the cloze test in 2010 version was intended to be a substitute for Vocabulary, Grammar and Reading (VGR) in 2005.*
G1P1: *But, we have only a short cloze passage with only four items; most of the items measure only the knowledge of syntax at the surface level; that is, higher order reading processes are not tapped in most of the cases.*
Researcher: *But, the previous research have shown that cloze test has the potentiality of being an integrative test of language proficiency if texts and gaps are selected rationally.*
G3P1*: Here, because we have a few items, each of them must tap a composite of lexico-grammatical features at both sentence and supra-sentential levels. Of course this may bring about a higher level of dependency between items.*

In wrapping up the discussion on this subsection, participants all agreed on the need for designing the multi-factor items for the cloze sub-test, each of them assembled under a separate Spec of a composite triple nature of VGR.

As to the Specialized subsection, participants started their discussion when they had a relatively clear picture of the General component and the shortcomings of its architecture. For this part, G3 participants led the discussion:

G3P1: *Local level of understanding could be tested under the general reading Spec. What is of primary importance for the specific purpose sub-component is measuring the inference ability for getting the implied meaning beyond the surface words; something that makes them refer to their background knowledge.*

G3P2: *In each test-let we can see that in most of the items both right answers and distracters are the exact paraphrase of what is explicitly stated in the texts. No inference is involved. Sub-technical vocabulary items, as mentioned before, must be measured in General vocabulary item set not here.*
G3P1: *What is important to be tested in this part is the ability of the examinees to infer the meaning inherent in the rhetorical functions of expository and argumentative texts.*
G2P1*: Items in this section must exclusively measure those reading skills which are not tapped in General subsection.*
G1P1: *That is inferring propositional meaning behind special rhetorical features and using their background to evaluate what is stated.*
G2P2: *in the existed version, now, the crafted Spec is highly inclusive. It can become more specific by excluding those parts which have already covered under previous Specs.*

At this point of the discussion, one of the participants in G3 mentioned a point by which the study entered its last phase.

G3P3: *If we can have two or three content specific passages, through content balancing and between- testlet item balancing we can have two or three unites under separate specific purpose Specs.*
G2P3: *this item clusters can be best assembled with a more inclusive cloze test of VGR and sub-technical vocbulary item set to form our exact joint test of EGP/ESP.*

The last question asked about the degree of match between structure of item clusters and dimensional structure of the test. In answering this question a degree of comparison was involved:

G1P1: *In 2000 version, there is field specificity not only in specialized reading subtest but also in General vocabulary subsection, while almost all of the words are from high frequency level.*

G1P2: *In new versions, if sub-technical vocabulary, as it was argued, are going to be tested in a separate set, it's better that the common core vocabulary, i.e. up to 2000 level of frequency, will*

be separated from the academic or sub-technical set with the former as a common set between all the fields and the latter differs between different fields of study.

G2P1: *Cloze subtest seems to be a pivot, with its item mapped on all other components VGR.*
G3P1: *In Specialized part, the first test-let must contains only those items which measure the inference ability for getting the propositional meaning of the rhetorical functions and a second passage must have items which tap the background knowledge of the examinees for interpretation and evaluation of the text.*

At the end, for actual dimensional structure to be more meaningfully represented, two checklists were developed in which participants coded the items in different categories based on the ability dimensions they tapped. Tables 6 and 7 show the categorization of each item which is based on the coding of that item in that category counting to more than 80% of the participants (remaining 20% coding in other categories were ignored).

Table 6. Participants' coding of the items in different categories based on the ability they tapped.

	English for Accounting			
Items	**Linguistic Factors**	**Content Factors**		
		Field related	**Other fields**	**Common-core**
1			√	
2		√		
3			√	
4			√	
5		√		
6		√		
7			√	
8			√	
9		√		
10			√	
11	√			
12	√			
13	√			
14	√			
15	√			
16		√		
17		√		
18		√		
19		√		
20		√		
21		√		

22	√			
23				√
24				√
25		√		
26	√			
27		√		
28		√		
29	√			
30				√

Table 7. Participants' cross-joining of items across different components of EGP-ESP test

	English for Accounting					
Test Subsections	**(GE)**		**EGAP**		**ESAP**	
	structure	**vocabulary**	**vocabulary**	**reading**	**vocabulary**	**reading**

GE	**Structure**				16,17,18, 19,20		22,29
	Vocabulary						
EGAP	**Vocabulary**				1,5,6		19,23,24,26, 30
	Reading						
ESAP	**Vocabulary**				2,3,4,7,8, 9, 10		
	Reading					16,17,18,20,21, 25, 27,28	

As it is clear from the table 7, item responses in this version of the test can be characterized by primary and secondary dimensions; the test data, therefore, can be considered as multidimensional (Ackerman, 1992; Roussos & Stout, 1996a; Shealy & Stout, 1993 Boughton et al. 2000).

Discussion and Conclusions

In Iran ESP context, ESP tests are administered at two high stake contexts, one as the Master Degree Entrance Exam and the second for the admission of the graduates into PhD programs. In this latter case, test was traditionally administered in two stages: 1) EGP as the prerequisite for the second stage which was 2) the Specialized module. These two modules were separate from each other in all the phases of design, administration, scoring and reporting. Something that makes the Master Degree counterpart worth investigating and at the same time challenging is the joint nature of these two components at all of the above mentioned phases.

Answering the first two questions, almost in all of the sub-components of the concerned tests (in this case 2010 version), expert judges pointed to the non-simple structure (more than one latent trait) behind item clusters; even in one case, one of the reverse engineers (G3P2) reported that there is a kind of construct irrelevancy in vocabulary subsection of accounting language test (with non-technical and sub-technical vocabulary as the relevant constructs and off-list vocabulary items as irrelevant constructs). In specific purpose component of the test also, reported within test-let imbalance of the item types with the technical vocabulary knowledge and content knowledge of the candidates as the secondary level abilities in all the test-lets items, can

be considered as an evidence that in this component also items are suspected of being multi-dimensional. In the cloze component, reading and lexico-grammatical knowledge were alternatively introduced by the experts as the primary and secondary abilities. Generally, participants at several points during the reverse engineering process showed inclination toward assembling the simple structure items under separate Specs with more specificity of the construct.

To restate the discussion in technical terms, with the exception of one case (G3P2), all the participants accepted that there are multiplicity of construct in different components and that this multiplicity is intentional with all of the secondary traits as the auxiliary rather than nuisance dimensions of the test (e.g., Shealy and Stout, 1993a), which must be assessed as a part of construct not as something irrelevant to it.

As to Dimensionality-based partitioning of the test items, Zhang (2006) said:

> if test data are multidimensional, and if the multidimensional structure reflects itself in the content classification of items, the test assembly process that controls content specification can be expected to maintain a given dimensional structure in the test. By the same token, if test data are in fact multidimensional, and if the multidimensional structure fails to reflect itself in the content classification of items, the test assembly process will not maintain the dimensional structure in the test when correct content classification is mistakenly assumed. Whether the dimensional structure is maintained or not should have some impact on ability estimation and important decisions based on ability estimates.

In other words, if multidimensionality of the test is certain, prescribing particular item-to-dimension mapping structures in assembling the items will control the size and nature of the multidimensionality and will have contribution to the precision and accuracy of the measure in yielding meaningful part-scores (ibid). Given multidimensionality of the test data in the context of the present study and with regard to the continuity of general and specific purposes language traits (Dudely-Evans & S.T. John, 1998), the suggested content specification (item classification) by the reverse engineers which was reported in tables 6 and 7 should be used as a framework for reconfiguring the item-to -dimension mapping structure, hence optimizing test blueprint, if EGAP and ESAP components are going to be tapped jointly in one module with small size clusters.

It is crystal clear that addressing this issue which is beyond the scope of the present study demands some psychometric involvement on the part of practitioners and researchers; they are now, surely on their way to solve the dilemma of multi-purpose tests in language assessment, in general, and in ESP/EAP tests, in particular.

References:

Bock, R. D., Thissen, D., & Zimowski, M. F.(1997). IRT estimation of domain scores. Journal of educational measurement, 34, 197–211.

Boughton, K.A., Yao, L., & Lewis, D.M. (2006). Reporting diagnostic subscale scores for tests composed of complex structure. In *Meeting of the national council on measurement in education*. San Francisco, CA, April.

Brown, J. (1988). Components of engineering-English reading ability. *System*, 16 (2), 193-200.

Davidson, F. (2003). Reverse engineering in language test development. SCALAR 6 CSULA. University of Illinois at Urbana-Champaign.

Davidson, F. & Lynch, B. (2002). Testcraft: A teacher's guide to writing and using language test specifications. New Haven, CT: Yale University Press.

Gierl, M. J., Leighton, J. P., Tan, X. (2005). Evaluating DETECT classification accuracy and consistency when data display complex structure. *Journal of Educational Measurement.*

Johnson. E. G., & Carlson, J. (1994). *The NAEP 1992 Technical Report* (Report No. 23-TR-20). Washington, DC: National Center for Education Statistics.

Khalifa, H and Weir, C J (2009).Examining Reading: Research and practice in assessing second language reading. Cambridge: UCLES/Cambridge University Press.

Kim, A. (2009). Investigating second language reading components: Reading for different types of meaning. Working Papers in TESOL & Applied Linguistics, 9 (2).

Nation, I.S.P. (2001, 2003), Vocab Profi le (software downloaded 4 May 2001), <http://www.vuow.ac.nz/nation_p/ software.download>.

Nunan, D. (1992). *Research Methods in Language Learning.* Cambridge: CUP.

Peara- Hernández,J.(2010).Teacher evaluation of item formats for an English Language Proficiency assessment. A thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts inTeaching English to Speakers of Other Languages, Portland State University.

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Scott Walter, F. (2010).Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, 7(4), 317-342.

Torre, J., & Patz, R. (2002) A Multidimensional item response theory approach to simultaneous ability estimation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, April, New Orleans, LA

Wainer, H., Vevea, J.L., Camacho, F., Reeve, B, Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented scores-"borrowing strength" to compute scores based on small numbers of items. In D. Thissen H.Wainer (Eds), Test Scoring (pp. 343-387). Mahwah, NJ: Erlbaum.

Weir, C J, Yang, H, and Jin, Y (2000) An empirical investigation of the componentiality of L2 reading in English for academic purposes, Cambridge: Cambridge University Press.

Yao, L., & Boughton, K. A. (2005). A Multidimensional item response modeling approach for improving subscale proficiency estimation in cognitive diagnostic assessments. Paper submitted for publication in APM.

Yen, W. M. (1987, June). A Bayesian/IRT index of objective performance. Paper presented at the annual meeting of the Psychometric Society, Montreal, Qubec, Canada.