ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 7, No. 11, pp. 1097-1107, November 2017
DOI: http://dx.doi.org/10.17507/tpls.0711.19

Construct Under-representation and Construct Irrelevant Variances on IELTS Academic Writing Task 1: Is There Any Threat to Validity?

Seyyed Mohammad Alavi
University of Tehran, Tehran, Iran

Ali Panahi Masjedlou
University of Tehran, Alborz Campus, Iran

***Abstract*—The study reports on the validity of IELTS Academic Writing Task One (IAWTO) and compares and assesses the performance descriptors, i.e., coherence and cohesion, lexical resource and grammatical range, employed on IAWTO and IELTS Academic Writing Task Two (IAWTT). To these objectives, the data used were 53 participants' responses to graphic prompts driven by IELTS scoring rubrics, descriptive prompt, and retrospective, rather than concurrent, think-aloud protocols for detecting the cognitive validity of responses. The results showed that IAWTO input was degenerate and insufficient, rendering the construct under-represented, i.e., narrowing the construct. It was also found that IAWTO displayed to be in tune with cognitive difficulty of diagram analysis and the intelligence-based design of the process chart, rather than bar chart, being thus symmetrical with variances irrelevant to construct; this is argued to be biased to one group: Leading to under-performance of one group in marked contrast to over-performance of another group. Added to that, qualitative results established on instructors' protocols were suggestive of the dominance of performance descriptors on IAWTT rather than on IAWTO. The pedagogical implications of this study are further argued.**

***Index Terms*—IAWTO validity, performance descriptors, construct under-representation, construct irrelevant variances**

I. Introduction

As for the merit of validity, back in 1980, the test developers have taken heed of social consequences of test and test use(Messick, 1995;Schouwstrato, 2000), i.e., the intended and unintended consequences (Haertel, 2013), and systematic evaluation of score interpretation with use of argument-relevant and evidence-based approach (Chapelle, 2012; Kane, 2013; Brennan, 2013; Sireci, 2013); therefore, application of test must be premised on testing principles and evidence not on a set of beliefs (Stevenson,1985).

Likewise, a well-positioned point is made out by researchers (Kane, 2013, 2010; Borsboom, 2013; Stevenson, 1985; Shaw & Imam, 2013), stressing the fact that the logicality of the inferences made of the test must be firmly-advocated by valid argument and theoretical rationale. As such, a very recent emphasis has been on the issue of test validity and fair assessment by Common European Framework of Reference, raising the consciousness of the principles appropriate to test validity (Huhta et al., 2014) because validity dramatically affects other issues relevant to language testing (Davies, 2011) and is applicable to all types of assessment (Messick, 1995).

However, it is prone to threat (Bachman, 1995; Shohamy, 1997) and subject to contamination associated with misapplication of tests or misinterpretation of score meaning (Henning, 1987; Xi, 2010) or boils down to so-called score pollution, as is distinct in “Lake Woebegone effect” paralleled with teaching to the test (Gipps, 1994), e.g., university preparation courses. Alderson in 2014 (as cited in Tineke, 2014) and Xie (2013) contend that teaching to the test should be taken account of with caveat since it touches more on strategy use and narrows the content leading to just score improvement because test-wise examiners get invalidly higher score (Messick, 1995), rather than developing true knowledge in the field.

In much broader terms, the threat more cogently elaborated on by Mesick (1995) is, in the main, defined and framed two-fold: Construct under-representation (CU) and construct irrelevant variances (CIV). With respect to the former, the assessment is too narrow and limited, referring to the situation in which the content is under-sampled by the assessment tools (Schouwstra, 2000). As with the latter, also, the assessment is too broad and extensive so that too many options and variables get conducive to appreciable lack of validity (Spaan, 2007).

As a general rule, the following can affect the construct to be tested, i.e., performance of test: Knowledge of subject area (Davies,2010), individual’s background knowledge, personality characteristics, test-taking strategies or test-wiseness, general intellectual or cognitive ability, instructional rubrics (Bachman, 1995), differential task functioning favored by particular test takers, and the topical knowledge (Bachman and Palmer, 2000); Bachman, 1995), all being the potential sources as construct irrelevant variances. A worth noticing definition is that construct refers to abstract nouns or psychological concepts, such as love, intelligence, anxiety, etc., characterized by being measurable and quantitatively observable (Bachman, 1995; Messick, 1995; Fulcher & Davidson, 2007; Newton and Shaw, 2013; Fulcher & Davidson, 2007; Bachman 1990; Bachman and Palmer, 2000; Brown, 2004; McNamara, 2006).

Clearly, the widespread reality is that the IELTS as the most favored international English proficiency test by the Australian federal government, since the early 1990s, (O'Loughlin, 2011) has been validated. It is hence expected not to bypass evidence based issues (Chapelle, 2012; Kane, 2013) emanating from the cross-disciplinary requirements of the test takers and the experiences of the IELTS instructors, both being potentially informative source of evidence, as our tenet runs.

The main incentive for the study is implicit in this fact: There lies a rarity of study addressing the true validity of IAWTO, testing the cognitive process of writing, being established on applied cognitive psychology (Yu, Rea-Dickens & Kiely, 2007) and being embedded in graph familiarity which is itself a potential source of construct irrelevant variance(Xi,2010). Quite parallel, IAWTO requires the candidates to describe or analyze, in their own words, a graph, table, chart, diagram or a process chart (IELTS Handbook, 2007), just based on the information, i.e., of cognitive and kind of statistical nature, provided in the intended prompt. This pollutes the validity of IAWTO.

To tease apart what is at stake here, taking account of the reality that IAWTO limits the test takers to just factual content of the input diagram and does not allow any speculated explanations- and personal beliefs of the test takers- outside the given data (IELTS Handbook, 2007) sounds to us to construct under-represent. The case in attention runs diametrically counter to Harsch and Rupp's (2011) stance, holding the view that for measuring the full writing abilities of the test takers, the writing task must be open rather than limited. Therefore, a moment of reflection indicates that limiting the test takers' ability to display their full performance and too-narrowing the construct to be measured are what the present investigation taps into. Therefore, the current study sets out to report our practical experience in conjunction with IAWTO validity through the lens of introspection-driven data and descriptive protocols. By adopting a mixed-methods approach and by being situated in the context of IELTS, this study intended to investigate the following research questions:

1. Is there any poverty of input, as a construct irrelevant variance, on IAWTO affecting the performance of the test takers? In other words, to what extent does IAWTO interact the test takers' underlying construct with respect to the performance descriptors including coherence and cohesion, lexical resource and grammatical range?

2. Triangulated approach considered, i.e., interpretation of descriptive protocols of testees and experienced IELTS instructors, "Do different task types, bar chart and process chart, display the differential performances on the part of the different test takers"? Put another way, which diagram is cognitively demanding and intelligence-design based, eventually biased sampling of the content domain by the assessment instrument to the disadvantage of one group vs. another group?

II. Methodology

Study Sample

Participants (N=53) taking part in the present study included three groups. Group one included forty four test-takers, at upper-intermediate level, with various age groups, selected from two branches of Iranian English language center, in Ardebil, Iran and without any reference to differential performance related to gender. They had voluntarily registered for general English courses for either communicative objectives or potentially for academic ambitions, following IELTS certification. The logic for the selection of examinees from the cited-level is two-fold: 1) To cautiously assure that the test takers must not have any introductory familiarity with the test format which may invalidly exert impact on test score interpretation 2) The other side is connected to the limitation of the study: We could not obtain any access to IELTS participants in real test situation to investigate their introspective position towards IAWTO prompts-relevant objectives, just a simulated attempt was made.

The second group of participants consisted of four trained raters, among whom one was an experienced IELTS Instructor teaching IELTS for 13 years, training the other three raters who were experienced conversation instructors and willing to correct the writings of the test takers. This is again another potential source for the limitation of the present study since if we could gain an access to the IELTS Instructors for rating purposes, the estimated reliability would be higher.

The third group of participants was IELTS instructors corresponded though email; 85 researchers were corresponded; their email addresses were extracted from conference proceedings happening in 2014; they were expected to respond if they had at least three years IAWTO teaching experience and with relative familiarity with language testing discipline. In precise detail, some of them indicated their lack of experience with IELTS and lack of expertise in testing; others neither responded in spite of the repeated emails nor completed the questionnaire; as Dornyei(2010) put succinctly, if 50 percent of the targeted emails respond, we are lucky. Just five people out of these 85 cooperated with the research: Two of them held an M.A. in TEFL, one with three years IAWTO teaching experience and another with eight years IELTS teaching experience; the other two instructors had PhD in TEFL and were faculty members of university with their dissertation in testing, one with seven years and another with two years teaching experience. Finally, the fifth instructor affiliated with London Teacher training college and as a PhD candidate in TEFL, with four years teaching experience and with IELTS overall band score 8.5 out of 9 completed the questionnaire. A demographic representation of IELTS instructors appears below.

IELTS Instructors	Education	IELTS Teaching experience
1	3 years	M.A in TEFL
2	8years	M.A in TEFL
3	7 years	PhD in TEFL
4	2 years	PhD in TEFL
5	4 years	PhD candidate in TEFL

Instruments

In the present study, the following instruments were employed:

1. Four graphic prompts, i.e., two bar charts and two process charts, extracted from Cambridge IELTS were administered to the examinees. Two of the charts were administered at the beginning of training as exercise charts in non-test situation and the other two at the end of the course in test situation for independent writing purposes. A word worth citing is that the first two charts, in addition to being explained to the examinees, had also sample analysis which was supposed to be advantageous in the consistency of the score on the part of raters in rating process and on the part of examinees in the writing process.

2. One writing prompt for IELTS academic writing task two for placement purposes. The prompt was extracted from Cambridge IELTS (2011).

3. Five IAWTO-related eliciting prompts, i.e. descriptive questionnaire, were designed and administered both to the test takers and to the IELTS instructors; the purposes of these prompts were to extract the introspective attitudes of the test takers towards poverty of input on IAWTO, clarity of instruction, cognitive load, similarity of performance descriptors and the intelligence based design of the tasks.

4. A consent letter indicating that the test takers agreed with participation in the project was also made use of.

5. Scoring rubrics sheet called performance descriptors sheet typical of analytical rating, extracted from IELTS Handbook (2007) were given to both raters and examinees. It was submitted to the raters in order mainly to avoid rater's bias (Henning, 1987; Weigle, 1999; Eckes, 2012; Ecks, 2008) and to free them from evaluative personal judgment (Weigle, 1994, 1998) and background variables (Johnson & Lim, 2009) potentially affecting their scoring process. In the case of the examinees, the objective was to engage them in the related prompts (Bachman & Palmer, 2000) and to keep them cognizant of the underlying construct to be tested. More tangibly, a use was made of a set of scoring rubrics being taken account of both holistically and analytically; as regards the former, IELTS Organization scale (IELTS Handout, 2007) considers kind of a holistic scale, however four areas of assessment system have been specified, i.e., task-achievement, coherence and cohesion, grammatical rage and lexical level for IAWTO; but to make the rating process more homogeneous and that so as to reduce the variation among raters, holistic frames were analyzed into concrete elements; that is to say, the lexical and grammatical features characterized by range and level, passive voice and active voice, conditional sentences, conjunctions and connectives were all among the linguistic elements.

A caution in order was that so as to put more trust in the process of scoring rubrics and to move in approximately exact line with the simulated criteria of IELTS assessment, in explaining the details to the raters, some samples of IAWTO and IELTS academic writing task two extracted from Cambridge IELTS (2011) and IELTS Test Builder were analyzed in the short training course. Moreover, an appropriate language-related handout (McCarter & Ash, 2003) including two pages was administered and instructed to both raters and examinees; the former took account of these appropriate language in their rating process, as the scoring rubrics on IELTS (IELTS Handbook, 2007) suggest and the latter was to be instructed the language-appropriate materials of handout during training session, based on which they analyzed the bar chart and process chart.

Procedure

To accomplish the objectives of the present study, the following procedures were undertaken. At the outset, the test takers signed a letter of consent for cooperation with the study; they were also appraised of the significance of IAWTO so as to be motivated for interaction with the trait to be measured (Bachman & Palmer, 2000). They then were placed into a suitably homogeneous level. Of course, the examinees were invited from a context, i.e., English language Educational department, where term-by-term achievement test serving the purpose of placement test for shift of level objectives was routinely administered. For putting more confidence in the interpretation of the findings and in order to deter from the potentially intervening and confounding variables, IELTS academic writing task two taken from Cambridge IELTS (2011) was administered to the examinees for the surety of their homogeneity.

In the second phase, the study was informally piloted for a small number of examinees for controlling the possible intervening and confounding variables and also for ironing out any sources of confusion in connection with instruction, etc. Then, the examinees were given and instructed two graphs for four sessions, every session lasting for 90 minutes. Next, the examinees were given and instructed both the handout related to IELTS task one for strategy training objectives and the performance descriptors. The reason why training sessions – for graph training - were conducted is traced evidentially back to the IELTS Organization advice that “it is a good idea to thoroughly prepare for the test" (IELTS Handbook, 2007, p. 3).

In the third phase, the examinees were administered two writing charts, each lasting for 20 minutes, as is IELTS Chart analysis time (IELTS Handout, 2007), with five minutes break time between. The examinees were recommended to pay attention to the performance descriptors in their analysis. Then, they were given five introspection eliciting prompts immediately after their writing was over and they went through the retrospective, rather than concurrent, process of think-aloud protocols. The fundamental rationale for the use of retrospective think-aloud protocol is traced back to an attempt to control some variables since it would otherwise affect their writing invalidly and on the other hand, the time-time variable- allocated for the task, i.e., 20 minutes, could have been more carefully controlled (IELTS Handbook, 2007). The point in case has been well elaborated by Green (1998) who contends that concurrent protocols less prone to confounding variables than are retrospective ones, but for the purposes of study, the retrospective protocols can also be used. A noteworthy point is that the participants who sufficed for yes/no answers without providing any reasons for their claim, were excluded from the study.

In phase four, the four raters were invited to cooperate with scoring process. Then, they were given and explained the handout and scoring rubrics. Next, they were delivered the writings of the test takers for scoring. In phase five, the IELTS instructors were asked to complete the researchers' self-designed descriptive questionnaire so as to get assured of their perceptions and attitudes towards the prompts mentioned in the questionnaire.

III. Results and Analysis

As Table 1 indicates, a statistically significant difference was observed at the P=.05 level for the four groups of raters: *F* (3,189) = 3.23, *p*=.02. Despite reaching statistical difference, Post-hoc comparisons using Duncan test, as is indicative in Table 2, revealed that the actual difference in mean scores between groups 1, 2, 3, and 4 was quite small, standing at 4.78, 4.58, 4.16 and 4.12, respectively.

Table 1
One-Way ANOVA results for comparing the scoring of the raters on placement test

Source	Sum of squares	df	Mean Square	F	Sig
Between Groups	14.254	3	4.751	3.235	02
Within Groups	277.585	189	1.496		
Total	291.839	192			

Table 2 indicates that a higher degree of consistency is observed in the scoring system of the raters (4.78; 4.58; 4.16; 4.12). As afore-cited, the raters were provided with both detailed guidance and holistic instruction in connection with rating system: analytical scale and holistic scale both presented in a checklist so as to make sure of the inter-rater consistency.

Table 2:
Post Hoc Test indicating the areas of difference between the raters' scoring

Raters	Mean of scores out of 9	Test takers
1	4.78	44
2	4.58	44
3	4.16	44
4	4.12	44

As Table 3 suggests, 27 test takers (61.4%) agreed with the insufficiency of input task, but regarding the clarity of task rubrics, moderate dissatisfaction was observed. A noteworthy point is that in connection with task input, two missing responses were observed and also concerning task rubrics, one missing response was evidenced.

Table 3:
Descriptive Statistics for the attitudes of test takers towards task input and task rubrics

Prompts	Response	Frequency	Percent
Task input	Yes	15	34.1
	No	27	61.4
Task rubrics	Yes	20	45.5
	No	23	52.3

As Table 4 reveals, 63.6% of the test takers stated the the first task of IELTS Academic Writing is cognitively demanding; 61.4% of the test takers indicated that is potentially possible to use the performance descriptors on task two of IELTS academic writing rather than on task one. Additionally, 70.5% of the test takers' introspected response indicated that chart analysis is of intelligence-based design, as this has been evidentially advocated by Yu, Rea-Dickens and Kiely(2007).

TABLE 4:
THE ATTITUDES OF TEST TAKERS TOWARDS COGNITIVE LOAD, PERFORMANCE DESCRIPTORS AND INTELLIGENCE-BASED DESIGN

Prompts	Response	Frequency	Percent
Cognitive load	Yes	28	63.6
	No	10	22.7
Performance descriptors	Yes	7	15.9
	No	27	61.4
Intelligence-based design	Yes	31	70.5
	No	10	9.1

Table 5 indicates that our participants were not real participants on IELTS and that there are a range of higher and lower scores, i.e. those for more proficient and less proficient test takers, are not typical of lack of confidence in the attitude-driven data. The analysis of the attitude-relevant data in light of the sufficiency of diagram input, thus, indicated that Group A and B, i.e., those getting a score of relatively 3.5-4.5 and those relatively with score of 4.5-7, both agreed with the insufficiency of diagram input. Statistically, as regards the former group, 41.7% and 58.3% disagreed and agreed, respectively, with the poverty of input in the diagram. On a similar scale, Group B viewed that 27% the diagram is informative and 72% the input is insufficient, both groups' attitude being in line with poverty of input on the diagram.

TABLE 5:
DESCRIPTIVE STATISTICS FOR THE LOW/HIGH SCORING TEST-TAKERS' ATTITUDES TOWARDS POVERTY OF INPUT, COGNITIVE LOAD, AND INTELLIGENCE-BASED DESIGN OF THE CHART

	Poverty of input		Cognitive load of chart		Intelligence-based design	
Band scores out of 9	Yes	No	Yes	No	Yes	No
3.5- 4.5	10	14	17	5	16	3
	41.7%	58.3%	77.3%	22.7%	84.2%	15.8%
4.5 - 7	5	13	11	5	15	1
	27.8%	72.2%	68.8%	31.3%	93.8	6.3%

As regards cognitive load of chart, no statistically significant difference was observed between the attitudes of more proficient and less proficient test takers; that is to put, 77.3 percent indicated that the chart is cognitively demanding and the other 22.7 disagreed with the cognitive load of the chart. On the other hand, those getting higher scores on IELTS writing, i.e., 68 percent agreed with the cognitively demanding nature of the chart with the exclusion of the 31.3 percent disagreeing with the cognitively demanding load of chart. As it was mentioned in the analysis of Table 5, the logic behind the inclusion of this comparison is traced back to the validation evidence providing us with the relative certainty of the attitudes compared to the proficiency level of the test takers.

With respect to intelligence-based design, as Table 5 stands, two groups of test takers have less or more similar attitudes towards the intelligence-based design of the test. That is to say, 84.2 percent of the group performing lower (3.5-4.5) on writing came to the consensus that chart analysis has an intelligence-based design and in the same vein, 93.8 percent of the test takers agreed with the intelligence-based design of the test, both being statistically on the logic that both more proficient test takers and less proficient ones had similar attitudes concerning prompt design. On the opposite hand, just 15.8 and 11.4 percent of the test takers, a very negligible number, stated that the chart does not have intelligence-based design.

TABLE 6:
DESCRIPTIVE STATISTICS FOR COHESION AND COHERENCE AND LEXICAL RESOURCE

Scoring rubrics	Frequency	Percent
Cohesion and coherence		
Task 1	2	4.5%
Task 2	32	72.7%
Task 1 and 2	10	22.7%
Range and level of words		
Task 1	2	4.5%
Task 2	42	95.5%

As it is clear in Table 6, 72.7% of test takers have indicated that it is feasible to have a coherent and cohesive performance on writing associated with task two rather than task one. Clearly, 4.5% of the test takers have advocated task one being coherent and cohesive and on the contrary, 22.7% have associated the coherence and cohesion with both task one and task two, keeping a neutral stance. Added to that, 95.5% of test takers have advocated the fact that task two is more engaging with respect to the use of range and level of the words than task one, so that just 4.5% of the test takers have agreed with easy use of word range and word level in task one.

TABLE 7:
GRAMMATICAL RANGE, CONDITIONAL SENTENCES AND CONNECTIVES/ CONJUNCTIONS

Grammatical range	frequency	Percent
Passive voice		
Task 1	5	11.4%
Task 2	33	75%
Task 1 and 2	6	13.6%
Conditional sentences		
Task 1	4	9.1%
Task 2	36	81.8%
Task 1 and 2	4	9.1%
Connectives/ Conjunctions		
Task 1	4	9.1
Task 2	23	52.3
Task 1 and 2	17	38.6

As it is clear in Table 7, the results obtained from the test takers' response prompted by stimulated recall indicated the following: 75% of the test takers-33 test takers- agreed with the fact that it is more feasible to use passive voice on task two rather than on task one. On the contrary, just five test takers addressed the practicality of using passive voice on task one and furthermore, 6 test takers held stated that it is possible to use passive voice on both tasks. Also, just four of the test takers addressed that it was feasible to use connectives/conjunction in task one.As regards conditional sentences, as with the stances towards passive voice, 36 test takers indicated the possibility and dominance of employing conditional sentences on task two. By marked contrast, the other two groups-just four of them- agreed with the feasibility of using conditional sentence in task one. With respect to the use of connectives and conjunctions, as Table 7 illustrates, 17 test takers stated the possibility of connectives/conjunction use in both task one and task two, while 23 test takers' position stood at their use in task two.

So far we have reported the qualitative result; now, the qualitative result is reported: The attitudes of the test takers and IELTS Instructors. As regards the former, the responses were simplified and were then adapted to the framework of performance descriptors elaborated on before; the genuine format of their descriptive protocols were, of course, kept untouched for qualitative report. Since the test takers were not well-acquainted with the design of IELTS writing, they were orally explained and non-technically presented with the prompts required, in the main, to motivate their introspective and cognitive attitudes towards the performance descriptors present in task one and task two. Some of the responses are reported below.

With respect to the first question that either input included in the chart is informative enough to motivate the test takers' to write and to use their lexical resources and grammatical range or not, we suffice for some of the responses appearing below.

1. I think the information in the chart is not enough, it could be more.
2. The chart is not related to writing; it just includes some statistics with little information to motivate me to write.
3. The chart provides me with nothing to write about. Since you told us not to write our idea, what could we write about?
4. I cannot interpret the chart and I have to write repeated sentences that I have learned for chart analysis.

The descriptive responses of such kind are typical of the test takers' dissatisfaction (61.1%) with the amount of information required for written performance. On the contrary, some test takers, as it is clear in Table 3, have viewed the chart of enough input. With relation to the second question that either the guideline or instruction of the chart is clear or not some sample responses are more illuminating:

1. I need another person to explain the chart to me because it confuses me.
2. It is not clear at all.
3. It I was clear.
4. If there were more explanation, it would be better.
5. I did not know what to do.

The responses continue in this form so that as it is clear in Table 3, the attitudes of those agreeing with the clarity of test rubrics are less or more closer to those of the disagreeing ones. Sample responses for question three that either the chart was cognitively demanding or not, the following randomly selected responses, among many, appeared:

1. "It challenged my mind", eight test takers had written.
2. It took my time more since I had to just think.
3. Like statistics I did not know what is what; because I am weak at math.
4. Process chart was more challenging than the Table charts, seven test takers responded.

With a closer attention to Table 4, the marked comparison of these sample responses shed light on the fact that it statistically and descriptively looks to be cognitively demanding. Furthermore, as it is explicitly interpreted from the last response, process chart sounds to be more cognitively demanding and is based on reasoning and intelligence than the other kinds of charts. As with question four that either we can use the same grammatical and lexical elements in IELTS academic writing task one and task two similarly or not, more clear-cut responses were provided:

1. The grammatical structures I knew I could not use because I was just thinking of what to do with the chart and how to organize the data on chart. But in task two, I could easily use the structure and vocabulary I had learned.

2.In task one, I could use repeated words and phrases I had memorized before; but in task two I needed extensive number of grammatical and lexical points", eleven test takers gave an answer similar to this one.

3. Task one needed finding the relation between the numbers and memorized words and phrases, but task two needed a real witting. I could easily use my knowledge in task two.

The samples of descriptive responses coming from the vey cognitive reservoir of the test takers are compatible with those quantified in Table 8 and Table 9. What these all lead us to are more explicitly manifested and specifically summarized in the responses to the question that either task one is of intelligence-based design or not? The descriptive responses below cast light on:

1. Task one was related to math and I think it wanted to check how intelligent I am.

2. Task one puzzled me like math because I hated math always.

3. Whatever was difficult for me was task one however I had practiced it during training session.

4. Process chart was more difficult than bar chart.

5. I couldn't understand the use of statistics and IELTS writing.

The above protocols lend support to the fact that task one has intelligence-based design. These points are made from the view of test takers. Along the same line, descriptive protocols of well-experienced IELTS instructors reported their response below as to weather the diagrams were biased or not.

Subject A: Yes, it is biased. Based on my studies on test method facet, the specific content of charts can function as a facet and affect students' performance differently…

Subject B: Yes. Due to specific nature of charts 1 and 2, some test-takers will be more advantaged…..

Subject C: It demands world knowledge. Hence, it is biased.

Subject D: I think, yes. It is biased…… since they are already familiar with the so called jargon in their major they can write more confidently.

Subject E: Yes; since the topic falls within the expertise of the test taker…

More attentively considered, the IELTS instructors responses all indicate the fact that the diagrams are biased to particular test takers; they have less or more set the logic for world knowledge, their major or expertise, familiarity with the content, etc. Moreover, regarding the second research question that either the test takers can use the same range and level of the words and grammar, i.e., passive voice, conditional sentences, connectives and conjunctions in the prompts-task one and task two-similarly or not; the following responses cast light on the fact.

Subject A: Yes. I think both require almost the same range of lexical and grammatical resources.

Subject B: To me, the test takers are most likely to utilize higher range of grammar (e.g. coherence, cohesion, and passive voice) in the first writing task while higher range of vocabulary in the 2nd academic writing task.

Subject C: Due to the nature of argumentative writing, task 2, is more challenging than task one. So, it needs higher grammatical range and level.

Subject D: No…… Each feature of task needs different students'
repertoire; students may utilize various lexis and forms. It's claimed
that the similar content is not the sole factor to be considered for
employment of the same vocabularies and structures.

Subject E: I suppose task two demands different lexical and structural resources

On closer scrutiny, the responses are quite illuminating; with the exception of one instructor, the others have indicated that task two needs much richer range and level of grammar and lexis; this is in parallel line with whatever the test takers indicate in both qualitative responses resulting from protocols and descriptive prompts. In conjunction with the fact that either there lies any poverty of input or not, the responses read below.

Subject A: No. Unlimited abilities of test takers cannot be measured by restricted characteristics of input. The extent of interaction relies on students' repertoire which can be used to activate their language knowledge, topical knowledge, schema, and cognitive processes... the input as one indicator is not sufficient.

Subject B: I think…….the test taker is in need of more language resources for the chart analysis.

Subject C: Task1 intends to measure the examinees ability in interpreting diagrams and illustrations and does not focus on the examinees potential to argue for or against a certain issue.

Subject D: As far as I am concerned, despite their proficiency in English, some test takers …… may not be able to excel the ones who depend on the prefabricated and pre-planned patters while doing the first writing task.

Subject E: What is included in the prompt can easily be revealed by looking closely at the charts…….Regarding eliciting a full range of language ability, it truly depends on the type of the prompt.

As the responses illustrate, the answers are various, but lean more towards incomplete information pyramid associated with task one. All subjects support the fact that the input on task one is degenerate and cannot interact the full range of the test takers in written competence. To also follow the fact that which chart is of intelligence based design, the following protocols are more evidential:

Subject A: It may seem that chart 3 is more intelligence-based,…..the examinees need to draw on their cognitive resources to initially find a logical relationship between different segments of the illustrations.

Subject B: Again table 3 includes intelligence design…….

Subject C: The chart 3 is mostly based on the pictorial intelligence instruction while the table 5 measures the mathematical intelligence profile of the test takers. However, the test-takers with higher spatial intelligence profile may not be able to cognitively interpret the chart. This is due to the fact that this intelligence has something to do more with learning than interpreting (to my humble option). In contrast, the mathematical intelligence, to me, can facilitate both learning and interpreting abilities. Therefore, the test takers would outperform table 5.

Subject D: I think Table five seems to be more cognitively demanding due to some computations which test takers need to do, having intelligence-based design.

Subject E: Chart 3 is more cognitively demanding and of intelligence-based design. To be explained appropriately, it puts more burdens on test taker's shoulder applying more technical vocabulary and needs more strategic competencies.

More attentive consideration of the protocols at hand reveals that that process chart taps more into the cognitive ability of the test takers and is of intelligence-based design rather than bar chart. This is in line with whatever the test takers have stated. To address the other question as to either the diagrams are useful for academic purposes or not, the following is more informative.

Subject A: I guess so. Each section of the test has been designed for a particular reason. This part of the test, in my opinion, does discriminate between an individual knowing how to interpret illustrations and someone who doesn't.

Subject B: Yes. The test calls for different skill and ability categories….required for success in all academic settings. In other words, the students of most, if not all, majors need to have such knowledge and skills at their disposal.

Subject C: Yes, it surely does. As afore-mentioned, the test takers with different intelligence profile are most likely to perform differently while interpreting and analyzing the diagrams and the tables.

Subject D: Interpretation of data is an integral and natural part of academic writing. However, the extent to which the IAWTO mirror academic writing in target language use domains is not quite clear to us.

Student E: Yes it depends on learners' topical knowledge……For example, students are able to use their biology knowledge learned at school to respond such tasks.

From the perspectives of the IELTS instructors, task one seems to be academically effective and instructive; acceptance of the academic value of the task one is taken with reservation as it is implied from the responses. As regards the fact that either task one is strategic-competence based or knowledge based, the responses appear below.

Subject A: Yes, it is strategy based. What is tested is the ability to decode nonverbal information regardless of the content.

Subject B: Surely they do. To be able to analyse diagrams and write reports require different strategies….

Subject C: They can be appropriate for the test -takers who are some kind familiar with the relevant knowledge and have sort of background information.

Subject D: I think yes; they should employ specific strategy to interpret the data.

Subject E: I accord with the statement. It is the strategic competencies of the learners enabling them to carry out the task….

The responses are much clearer. All instructors agreed with the fact that the task one of IELTS needs strategy more than knowledge. So, this can also lend support to the bias-relevant issues of the task one and it can also lend support to the fact that input is not enough to engage them. The last issue to be tapped into is framed within the lines of Gipps'(1994) statement that teaching to the test pollutes score as score polluting variance. The following responses indicate that either task one of IELTS can contaminate the performance of the test takers or not, giving support to or weakening the previous issue relevant to use of strategic competence on task one.

Subject A: Diagrams require the knowledge of decoding illustrations and interpreting them. Knowledge, if the content is considered, is not the assessment element…

Subject B: I agree to the idea that intensive courses, i.e. teaching to the test, teach strategies at the expense of content.

Subject C: I think some kind of background information can pave the way for the outperformance of the test takers. The strategies can be more fruitful when analyzing the tables.

Subject D: If test takers are oblivious of IELTS task requirements they might not be at their best in IELST test….. it is believed that the nature of IELTS writing is predictable and can easily be achieved by introducing the so called pre-fabricated sentences. I suppose in such courses light be shed on this issue.

Subject E: I believe that strategic competence (metacognitive) is a part of language knowledge. Therefore, instructing learners how to perform strategically in IELTS exam do not interfere with the knowledge they have to be taught.

As it is directly inferred, the responses are in consonance with the fact that IELTS intensive course instruction leads to score pollution since content is sacrificed to strategy use or so. What is more significant, with some pre-fabricated patterns and techniques the test takers can achieve a score under-representing the true level of test takers.

IV. Discussion and Conclusion

This study set out to investigate the debated validity of IAWTO-related amount of input on IAWTO, being designed as a follow-up to an earlier study by Yu, Rea-Dickens and Kiely (2007). Two major issues in the current study are: The use of think-aloud protocols driven by stimulated recall to extract cognitively existing attitudes of the test takers towards the performance descriptors in task one and two framed in comparison and evaluation and tapping into the cognitive process of well-experienced IELTS instructors and their perception of task one and two of IELTS.

With an eye towards the design and administration of the task one testlet of IELTS, no denial of the validity of the before-design nature of the test is made. Rather, we stand on the position that a test of whatever essence should be on a continued cycle validated, being so will assist the stakeholders to gain productive results from the safe and valid inferences of the test. That is why the present study has considered both test takers and IELTS instructors' perception and experience of the case at issue. This is what has been evidentially advocated by researchers (Moss, 2013), contending that for making a decision, evidence of students' performance is not per se sufficient; one must consider information about the conceptual and material resources, the teaching and learning process, material resources, etc., considering the attitudes of the stakeholders.

More apparently, both test takers and IELTS instructors have agreed with the poverty of input, the biased and intelligence-design based nature of task one and the cognitively demanding nature of task one. More significantly, as is axiomatic in the protocols, use of lexical range and level as well as grammatical level sound to be more pragmatically and suitably evaluative in task two rather than task one. Namely, the majority of test takers have addressed the fact that use of passive voice, conditional sentences, advanced vocabulary level, etc., for instance, seem to be more practical in task two since task one does not include the needed input to interact these abilities. On the other hand, as it comes from the instructors cognition as well as experience, task one of IELTS demands strategy more than knowledge and content, hence the curriculum gets reduced and narrowed (Gipps, 1994) to a situation where drawing parallel between test situation and non-test situation (Bachman & Palmer, 2000) may seem unlikely.

With a view to Table 7, for instance, it gets vivid that characteristics of task design exert some dramatic influence on the nature of teaching and test preparation (Green, 2006); this indicate that the intelligence-based design will push the teachers to teach those content relevant to intelligence. This is in effect misuse of the test and the reason for the misuse should be tackled (Fulcher, 2009). Alderson states that in designing our test, we are required to consider the needs of the test takers and courses… (as cited in Davies, 2014). So, the table indicates that there is a hidden trace of the fact that the chart has intelligence-based design, with more sample Size, more confidence can be put in the investigation: An expert task left to the testers. Otherwise, the design of the test will follow its chaotic life, more likely not suitable for the intended purpose; as Fucher (2012) views, there will be a design chaos and this will lead to the validity chaos. Therefore, if the test takers under-perform on task one due to biased nature of task and poverty of input as well as for the intelligence-based design of the tasks, since the tasks are interdependent in terms of the reality that performance on one task can emotionally affect the performance of the test takers on another task, thus they may even underperform on task two by the agency of task one. Our reservation is implicit in the reality that either the method associated with IELTS Academic Writing Task One- measuring the cognitive process of thinking and demanding a description and analysis based on the chart- can meet the requirements of basic testing practices or not. The point is well addressed by Reath(2004), stating that the testing methods used should meet the standards of testing practices.

To exert elaboration on the matter more concretely, let's take as an example the response that "It took my time more since I had to just think….”; it is inferred that thinking took the test taker's time more than writing because it was based on reasoning and the numbers need to be more organized in the form of a content; since the test taker thinks of connecting the ideas, how can he/she think of form? So, this is more suggestive that taking account of lexical range and level as well as grammatical resource will limit the test takers to use their full range of ability. This is what can be a potential source as construct irrelevant variance leading to construct under-representation.

What our belief pursues is that without taking account of IELTS instructors' experience and test takers' introspective evidence- those working practically and being involved in the field, tests will stand at risk in terms of validity, use and score interpretation. The point is well-argued by Mathew (2014) and Wall (2000), viewing the fact that involvement of stakeholders in the validation-relevant and evidence-needed issues is crucially in requirement since the test takers and test users as well as the whole stakeholders (Wall, 2000) are affected by the test results.

On the ground of whatever has been addressed so far, caveat in order lies with the very nature of IAWTO design, demanding further triangulated investigation: It should be believed that IELTS instructors and test takers' cognitive way of thinking and perception towards IAWTO can provide the testers with evidence which is not provided otherwise; since piloting the items in the corner of the simulated context without calling the attention of the teachers and learners into work is different from just designing some items without considering the social consequences.

REFERENCES

[1] Bachman, L. F. (1995). Fundamental Consideration in language testing. Oxford: Oxford University Press.

[2] Bachman, L. F. & Palmer, A.S. (2000). Language testing in practice. Oxford: Oxford University Press.

[3] Borsboom, D. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement. 50 (1),* 110–114.

[4] Brennan,R.T. (2013). Commentary on “Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50(1),* 74–83.

[5] Brown, H. D. (2004). Language assessment: principles and classroom practices. San Francisco: Francisco San State University.

[6] Cambridge IELTS. (2011). Official examination papers from University of Cambridge: ESOL Examinations. Cambridge: Cambridge Publications.

[7] Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. *Language Testing, 29(1),* 19–27.

[8] Davies, A. (2011). Kane, Validity and Soundness. *Language Testing, 26(1), 37*- 42.

[9] Davies, A. (2010). Test Fairness: a response. *Language Testing, 27(2), 171*-176.

[10] Davies, A. (2014). Remembering 1980. *Language Assessment Quarterly, 11*, 129–135.

[11] Dornyei,Z. (2010). Questionnaires in Second Language Research (2nd ed.). London: Routledge Publication

[12] Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing 2008 25 (2),* 155–185.

[13] Eckes, T. (2012).Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly, 9,* 270-292.

[14] Fulcher,D. & Davidson, F. (2007). Language Testing and Assessment. London: Routledge

[15] Fulcher, G. (2012). Practical Language Testing. London: Hodder Education.

[16] Fulcher, G., (2009). Test Use and Political philosophy. *Annual Review of AppliedLinguistics.29*, 3-20.

[17] Gipps, C.V. (1994). Beyond Testing: towards a theory of educational assessment. London: Palmer Press.

[18] Green, A. (1997). Studies in language testing: Verbal protocol analysis in language testing research. Cambridge: Cambridge University Press

[19] Haertel, E. (2013). Getting the Help We Need. *Journal of Educational Measurement,* 50 (1), 84–90

[20] Harsch.C. & Rupp. A.A. (2011). Designing and Scaling Level-Specific Writing Tasks in Alignment with the CEFR: A Test-Centered Approach. *Language Assessment Quarterly, 8,* 1–33.

[21] Henning, G. (1987). A guide to language testing: Development, evaluation, research, Rowley: Newbury House.

[22] Huhta,A., Alanen,R., Tarnanen,M., Martin,M., & Hirvelä,T.(2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31(3),* 307– 328.

[23] IELTS Handbook. (2007). Retrieved on February 13, 2014, from: www.ielts.org.

[24] Jonson, J.S., & Lim, G.S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26(4),* 485-505

[25] Kane, M.T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50(1),* 1-73

[26] Kane, M. T. (2013). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement. 50(1),* 115–122.

[27] Kane.M. T. (2010). Validity and fairness. *Language Testing, 27(2),* 177-182.

[28] Mathew, R. (2004). Stakeholder involvement in language assessment: Does it improve ethicality? *Language Assessment Quarterly*, *1*, 123–135.

[29] McCarter, S., & Ash, J. (2003). IELTS Test builder. Macmillan: Macmillan Education

[30] McNamara, T. (2006). Validity in Language Testing: The Challenge of Samuel Messick's Legacy. *Language Assessment Quarterly*, 3(1), 31-51

[31] Messick, S. (1995). Validity of Psychological Assessment: validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50(9),* 741-749.

[32] Moss, A.P. (2013). Validity in Action: Lessons From Studies of Data Use. *Journal of Educational Measurement. 50(1),* 91–98.

[33] Newton,P. & Shaw,S. (2013). Book Announcement: Validity in Educational and psychological assessment. *Research Notes*, 54(2), 33-36.

[34] O'Loughlin. K. (2011). The Interpretation and Use of Proficiency Test Scores in University Selection: How Valid and Ethical Are They? *Language Assessment Quarterly*, 8, 146–160

[35] Reath. A. (2004). Language Analysis in the Context of the Asylum Process: Procedures, Validity, and Consequences. *Language assessment quarterly. 1(4),* 209–233.

[36] Schouwstra, S.J. (2000). On Testing Plausible Threats to Contruct Validity): The Institutional repository of the University of Amsterdam (UvA). Retrieved On January 2014, from: http://dare.uva.nl/document/56520.

[37] Shaw, S., & Imam, H. (2013). Assessment of International Students Through the Medium of English: Ensuring Validity and Fairness in Content-Based Examinations. *Language Assessment Quarterly, 10*,452–475.

[38] Shohamy, E. (1997). Testing methods, testing consequences: are they ethical? are they fair? *Language Testing. 14(3),* 340-349.

[39] Sireci, S.G. (2013). Agreeing on Validity Arguments. *Journal of Educational Measurement.* 50 (1), 99–104.

[40] Spaan,.M. (2007). Evolution of a Test Item. *Language Assessment Quarterly*, *4*(3), 279–293.

[41] Stevenson, D.S. (1985). Authenticity, validity and a tea party. *Language Testing,*2 (41), 41-47

[42] Tineke, B. (2014). A lifetime of language testing: an interview with J. Charles Alderson. *Language Assessment Quarterly, 11(1),* 103-119.

[43] Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28,* 499–509.

[44] Weigle,S.C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 1,* 197-223.

[45] Weigle,S.C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263-287.

[46] Wiegle, S.C.(1999).Investigating rater/prompt interactions in writing assessment Quantitative and qualitative approaches. *Assessing Writing, 6,*145-178.

[47] Weir, C. J. (1990). Communicative language testing. London: Prentice-Hall, Inc.

[48] Xi, X. (2010). Aspects of performance on line graph description tasks: influenced by graph familiarity and different task features. *Language Testing, 27(1),* 73–100.

[49] Xi, X.(2010). How do we go about investigating test fairness? *Language Testing, 27(2*), 147-170.

[50] Xie,Q. (2013). Does Test Preparation Work? Implications for Score Validity. *Language Assessment Quarterly, 10,* 196–218

[51] Yu,G., Rea-Dicken,P., & Kiely,R. (2007). The cognitive processes of taking Academic Writing Task 1. *IELTS Research Reports Volume 11.*Retrieved On February 2014, from http:// www.ielts.org/researchers/volume 11.aspx.

Seyyed Mohammad Alavi is full professor at the University of Tehran, Iran. He has taken his PhD in TEFL from England and has published a good number of articles in the field of language assessment. He has been Vice- Chancellor of the University of Tehran and has also been and is in charge of assessment department at the University of Tehran. Furthermore, he has supervised a good number of PhD dissertations and M.A. theses.

Ali Panahi Masjedlou, PhD candidate in TEFL at the University of Tehran, has published and presented some papers in the field of language assessment. He has two Diplomas, i.e., Diploma in TESOL and Diploma in Teacher Training from England. Also, he has an International TESOL Certificate from the USA.. At present, he is working on his dissertation titled: "On the Validity Argument of IELTS Listening" supervised by Professor Seyyed Mohammad Alavi and advised by Dr. Shiva Kaivanpana.