MJAL
The modern journal of applied linguistics
Volume 1:2 March 2009
ISSN 0974 – 8741

Identifying sources of bias in EFL writing assessment through multiple trait scoring

By

Mohammad Ali Salmani-Nodoushan

Identifying sources of bias in EFL writing assessment through multiple trait scoring

Mohammad Ali Salmani-Nodoushan

The Author

Mohammad Ali Salmani-Nodoushan is an assistant professor of TEFL at the English Department of University of Zanjan, Iran. His research interests include language testing in general, and testing English for Specific Purposes, Computer Adaptive Testing, and Performance Assessment in particular.

Correspondence: salmani.nodoushan@yahoo.com or nodushan@znu.ac.ir
Address: English Department, College of Humanities, University of Zanjan, Iran.
Phone: 0098-241-5152664 (Office) or 0098-261-6467106 (Home)
Mobile: 0098-912-325-3829

Content

Abstract

For purposes of the present study, it was hypothesized that field (in) dependence would introduce systematic variance into EFL learners' performance on composition tests. 1743 freshman, sophomore, junior, and senior students all majoring in English at different Iranian universities and colleges took the Group Embedded Figures Test (GEFT). The resulting 582 Field-Independent (FI) and the 707 Field-Dependent (FD) students then took the 2000 version of the IELTS. Using SPSS commands for collapsing continuous variables into groups, and participants' IELTS scores (based on 25, 50, 75 percentiles), four proficiency groups were identified for each kind of cognitive styles. From each proficiency group, 36 FD and 36 FI individuals were selected through a matching process. The scores obtained by the resulting sample of 288 participants on the second writing task of the IELTS test were used as the data for this study. The results of data analysis revealed that individuals' cognitive styles resulted in a significant difference in their writing performance in proficient, semi-proficient, and fairly proficient groups, but not in the low-proficient group. The findings also indicated that cognitive style resulted in a significant difference in participants' performance on such aspects of EFL composition as content, structure, and language.

KEYWORDS: EFL writing; Multiple trait scoring; Writing assessment; Systematic variance; Test bias

1. Introduction

Bachman's (1990) model of multi-layered model of language ability (Communicative Language Ability or CLA) has shed some light at least on the areas where one should search for traces of possible factors that affect test scores in general and language test scores in particular (also see Alderson, 1991; Anivan, 1991; Canale and Swain, 1980; Hymes, 1974). Attempts at identifying factors that affect test scores have resulted in a taxonomy of factors. Such a taxonomy is neither exhaustive nor comprehensive. More research is needed to determine what other factors may influence the performance of test takers.

The test takers' cognitive styles is just one such potential area. The term 'cognitive style' refers to the link between personality and cognition that controls the way we learn things in

general and the particular approach learners adopt when dealing with problems. Cognitive styles are relatively stable indicators of how learners perceive, interact with, and respond to the learning environment (Keefe, 1979). In theory, there exist lots of cognitive styles. Nevertheless, only a few of the possible number of cognitive styles have received the attention of L2 researchers in recent years; one such area is "field independence" (FI) or "field dependence" (FD).

Field dependence (FD) refers to a cognitive style in which an individual tends to look at the whole of a learning task which contains many items. The FD individual has difficulty in studying a particular item when it occurs within a field of other items. The "field" may be perceptual or it may be abstract, such as a set of ideas, thoughts, or feelings. Field independence (FI), on the contrary, refers to a cognitive style in which an individual is able to identify or focus on particular items and is not discredited by other items in the background or context (Brown, 2000; Gollnick and Chein, 1994; Salmani-Nodoushan, 2006).

Due to the psychologists' hypothesized relationship of field-(in)dependence to cognitive and interpersonal abilities, it appears possible that language tests of today may favor learners with certain cognitive styles. The present study is an attempt at finding the possible effects of learners' cognitive styles on their performance on EFL writing tests.

2. Background

Over the past few years language testing specialists have called for performance assessment in EFL contexts. Advocates of performance assessments maintain that every task must have performance criteria for at least two reasons. On the one hand, the criteria define for students and others the type of behavior or attributes of a product which are expected. On the other hand, a well-defined scoring system allows the teacher, the students, and others to evaluate a performance or product as objectively as possible. If performance criteria are well defined, another person acting independently will award a student essentially the same score. Furthermore, well-written performance criteria will allow the teacher to be consistent in scoring over time. If a teacher fails to have a clear sense of the full dimensions of performance, ranging from poor or unacceptable to exemplary, he or she will not be able to teach students to perform at the highest levels or help students to evaluate their own performance.

In developing performance criteria, one must both define the attribute(s) being evaluated and also develop a performance continuum. For example, one attribute in the evaluation of writing might be writing mechanics, defined as the extent to which the student correctly uses proper grammar, punctuation, and spelling. As for the performance dimension, it can range from high quality (well-organized, good transitions with few errors) to low quality (so many errors that the paper is difficult to read and understand). Testers should keep in mind that the key to developing performance criteria is to place oneself in the hypothetical situation of having to give feedback to a student who has performed poorly on a task. Advocates of performance assessment suggest that a teacher should be able to tell the student exactly what must be done to receive a higher score. If performance criteria are well defined, the student then will understand what he or she must do to improve. It is possible, of course, to develop performance criteria for almost any of the characteristics or attributes of a performance or product. However, experts in developing performance criteria warn against evaluating those aspects of a performance or product which are easily measured. Ultimately, performances and products must be judged on those attributes which are most crucial.

Developing performance tasks or performance assessments seems reasonably straightforward, for the process consists of only three steps. The reality, however, is that quality performance tasks are difficult to develop. With this caveat in mind, the three steps include:

1. Listing the skills and knowledge the teacher wishes to have students learn as a result of completing a task. As tasks are designed, one should begin by identifying the types of knowledge and skills students are expected to learn and practice. These should be of high value, worth teaching to, and worth learning. In order to be authentic, they should be similar to those which are faced by adults in their daily lives and work;
2. Designing a performance task which requires the students to demonstrate these skills and knowledge. The performance tasks should motivate students. They also should be challenging, yet achievable. That is, they must be designed so that students are able to complete them successfully. In addition, one should seek to design tasks with sufficient depth and breadth so that valid generalizations about overall student competence can be made;
3. Developing explicit performance criteria which measure the extent to which students have mastered the skills and knowledge. It is recommended that there be a scoring

system for each performance task. The performance criteria consist of a set of score points which define in explicit terms the range of student performance. Well-defined performance criteria will indicate to students what sorts of processes and products are required to show mastery and also will provide the teacher with an "objective" scoring guide for evaluating student work. The performance criteria should be based on those attributes of a product or performance which are most critical to attaining mastery. It also is recommended that students be provided with examples of high quality work, so they can see what is expected of them.

3.1. Approaches to scoring writing

Scoring writing is a very delicate task. There is still a lot of controversy among teachers as to how students' writing assignments should be scored. Traditionally a student's writing performance was judged, in a norm-referenced approach, in comparison with the performance of others. Over the past few decades, however, this norm-referenced method has largely given way to criterion-referenced procedures. In a criterion-referenced approach to scoring writing, the quality of each essay is judged in its own right against such external criteria as coherence, grammatical accuracy, contextual appropriacy, and so on. Such an approach takes a variety of forms and falls into three main categories: (a) holistic, (b) analytic, and (c) trait-based. As Weigle (2002) claims, the holistic approach offers a general impression of a piece of writing; the analytic approach is based on separate scales of overall writing features; and the trait-based approach judges performance traits relative to a particular task.

3.1.1. Holistic scoring

An holistic scale is based on a single, integrated score of writing behavior. The aim of this method is to rate a writer's overall proficiency. To this end, an individual impression of the quality of a writing sample is made. Such a global approach to the student's writing tacitly reflects the idea that "writing is a single entity which is best captured by a single scale that integrates the inherent qualities of the writing" (Hyland, 2003, p. 227). The holistic approach contrasts with earlier assessment methods in which the rater tried to hunt errors in students, writing. As White (1994) says, the holistic approach emphasizes what the writer can do well rather than dwelling on his or her deficiencies. Although it is relatively easy of use, the holistic approach to scoring writing reduces writing to a single score. This means that teachers cannot gain diagnostic information which is crucial for their subsequent remedial teaching.

Moreover, raters must be carefully trained to respond in the same way to the same features in different students' writings because the holistic approach requires a response to the text as a whole. Cohen (1994, p. 317) summarizes the advantages and disadvantages of the holistic method as follows:

Advantages	Disadvantages
Global impression not a single ability	Provides no diagnostic information
Emphasis on achievement not deficiencies	Difficult to interpret composite score
Weight can be assigned to certain criteria	Smooths out different abilities in subskills
Encourages rater discussion and agreement	Raters may overlook subskills
	Penalizes attempts to use challenging forms
	Longer essays may get higher scores
	One score reduces reliability
	May confuse writing ability with language proficiency

The reliability of scores gained through the holistic approach improves when two or more trained raters score each paper. Without guidance, however, raters have trouble agreeing both on the specific features of good writing and on the relative quality of papers. Young teachers gradually gain the experience that will lead them to develop the confidence and skill to score consistently. However, scoring rubrics or guides can be used to help raters by providing bands of descriptions which correspond to particular proficiency or rhetorical criteria. Scoring rubrics are commonly designed to suit different contexts; They seek to reflect the goals of the course and what its teachers value as good writing. As such, scoring rubrics should be carefully written to avoid ambiguity.

It is possible for scoring rubrics to have nine- or ten-step scales. However, it is unlikely that scorers can reliably distinguish more than about nine bands. Most holistic rubrics found in the literature have between four to six bands. Examples of holistic rubrics can be found in Cohen (1994), Hamp-Lyons (1991), and White (1994). The following sample rubric for a holistically scored essay can be found in Hyland (2003, p. 228).

GRADE	CHARACTERISTICS
A	The main idea is stated clearly and the essay is well organized and coherent. Excellent choice of vocabulary and very few grammatical errors. Good spelling and punctuation.
B	The main idea is fairly clear and the essay is moderately well organized and relatively coherent.

	The vocabulary is good and only minor grammar errors. A few spelling and punctuation errors.
C	The main idea is indicated but not clearly. The essay is not very well organized and is somewhat lacking in coherence. Vocabulary is average. There are some major and minor grammatical errors together with a number of spelling and punctuation mistakes.
D	The main idea is hard to identify or unrelated to the development. The essay is poorly organized and relatively incoherent. The use of vocabulary is weak and grammatical errors appear frequently. There are also frequent spelling and punctuation errors.
E	The main idea is missing and the essay is poorly organized and generally incoherent. The use of vocabulary is very weak and grammatical errors appear very frequently. There are many spelling and punctuation errors.

A rubric for holistic scoring of an intermediate-level ESL essay (Adopted from Hyland (2003) with permission)

It is possible to devise more complex rubrics for complicated forms of writing. Such complex rubrics can be tailored to genre and topic. They can also take into account the fact that students may have to express and counter different viewpoints or draw on suitable interpersonal strategies. However, there is a dilemma here; while more delicate holistic rubrics are feasible, they are also more difficult to apply since the rater may encounter texts which simultaneously display characteristics from more than one category. As Hyland (2003, p. 228) puts it, even the above simple rubric may fail to provide an obvious basis for scoring "where, for instance, a text has a clear thesis statement and displays appropriate staging for the genre but contains numerous significant grammatical errors, so that features from B and C grades overlap." In such a situation raters may choose to make finer distinctions with + and – subdivisions (i.e., grading the problematic writing as a B – or C+.

3.1.2. Analytic scoring

In analytic scoring procedures, raters judge a text against a set of criteria which are important to good writing. Raters must give a score for each category. This helps ensure that features of good writing are not collapsed into one, and, as such, provides more information than a single holistic score. Analytic scoring procedures more clearly define the features to be assessed by separating, and sometimes weighting, individual components. Analytic scoring is, therefore, more effective in discriminating between weaker texts. Analytic scoring rubrics which are in wide use today have separate scales for content, organization, and grammar, with vocabulary and mechanics sometimes added separately. Each of these parts is assigned a numerical value.

Analytic methods encourage teachers to pay close attention to specific features of writing quality. As such, they assist rater training, and give more detailed information, which means that they are also useful as diagnostic and teaching tools. It is recommended that raters, when devising an analytic rubric, use explicit and comprehensible descriptors that relate directly to what is taught. This allows teachers to target writing weaknesses precisely. It also provides a clear framework for feedback and revision. The criteria delineated in an analytic rubric can be introduced early in the writing course to show students how their writing will be assessed. They also point out to the students the properties that their teachers value in their writings. As Hyland (2003, p. 229) says, some critics of analytic scoring procedures, however, "point to the dangers of the halo effect; results in rating one scale may influence the rating of others, while the extent to which writing can be seen as a sum of different parts is controversial." Cohen (1994) and McNamara (1996) have identified the advantages and disadvantages of analytic rubrics as follows:

Advantages	Disadvantages
Encourages raters to address the same features	May divert attention from overall essay effect
Allows more diagnostic reporting	Rating one scale may influence others
Assists reliability as candidate gets several scores	Very time consuming compared with holistic method
Detailed criteria allow easier rater training	Writing is more than simply the sum of its parts
Prevents conflation of categories into one	Favors essays where scalable info easily extracted
Allows teachers to prioritize specific aspects	Descriptors may overlap or ambiguous

3.1.3. Trait-based scoring

Trait-based approaches are context-sensitive and, as such, differ from both holistic and analytic scoring methods. They do not presuppose that the quality of a text can be based on a priori views of good writing. Rather, as Hamp-Lyons (1991) claims, trait-based instruments are designed to clearly define the specific topic and genre features of the task being judged. The goal of trait-based scoring approaches is to create criteria for writing unique to each prompt and the writing produced in response to it. Trait-based approaches use either primary-trait or multiple-trait systems.

3.1.3.1. Primary trait scoring

In primary-trait scoring, criteria intended for holistic scoring as it involves rating a piece of writing are sharpened and narrowed by just one feature relevant to the writing task in

question. The primary trait is identified by the task designers (i.e., usually the writing teachers). It allows teachers and students to focus on a critical feature of the writing task (e.g. appropriate text staging, creative response, effective argument, reference to sources, audience design, etc.). Although primary-trait approaches recognize that it is not possible to respond to everything at once, in practice raters may find it hard to focus exclusively on the specified trait in focus; They may also inadvertently include other traits in their scoring. The primary-trait approach lacks generalizability. It requires a very detailed scoring guide for each specific writing task. This means that primary-trait scoring should be used in courses where teachers need to judge learners' command of specific writing skills rather than more general improvement.

3.1.3.2. Multiple trait scoring

Like analytic scoring, multiple-trait scoring requires raters to provide separate scores for different writing features. Unlike analytic scoring, multiple-trait scoring requires raters to ensure that the features being scored are relevant to the assessment task in question. As such, multiple-trait scoring is often regarded as an ideal scoring procedure. Multiple-trait scoring, as Hyland (2003, p. 230) puts it, "treats writing as a multifaceted construct which is situated in particular contexts and purposes, so scoring rubrics can address traits that do not occur in more general analytic scales." The examples Hyland (ibid) provides include the ability to "summarize a course text," "consider both sides of an argument," or "develop the move structure of an abstract."

Multiple-trait scoring is very flexible because each task can be related to its own scale with scoring adapted to the context, purpose, and genre of the elicited writing. It has benefits for raters, students, and course designers. Multiple-trait scoring encourages raters to attend to relative strengths and weaknesses in an essay. As for the students, it provides opportunities for them to have access to detailed feedback in relation to their writing performance. Multiple-trait scoring also assists washback into instruction directly.

Multiple-trait scoring, therefore, provides rich data which will inform decisions about remedial instruction and course content. One major disadvantage of multiple-trait scoring is that it requires enormous amounts of time to devise and administer. Another major disadvantage is that teachers may still fall back on traditional general categories in their scoring although traits are specific to the task (See Cohen, 1994: 323).

4. Statement of purpose

Over seventeen years of teaching experience has taught the researcher that compositions written by EFL learners, if scored through a robust scoring method, will show traces of the influence of test-irrelevant factors that bias test results by introducing systematic variance into the final scores. Therefore, the present study attempted to account for the probable effects of just one such factor (i.e., FD/FI cognitive style) on EFL learners' written performance. It was hypothesized that participants' FD/FI cognitive styles affected their writing performance in meaningful and significant ways. The study specifically addressed the following questions:

1. Is there a significant difference in the mean composition scores for FDs and FIs?
2. Is there a significant difference in the mean "content" scores for FDs and FIs?
3. Is there a significant difference in the mean "structure" scores for FDs and FIs?
4. Is there a significant difference in the mean "language" scores for FDs and FIs?

In all of the above questions, participants' proficiency levels were held constant. In other words, mean comparisons were done between FD and FI individuals within the same proficiency group.

5. METHOD

5.1. Participants and procedures

On the whole, 288 participants provided the sample for the present study. They were chosen in a systematic way to make the results of the study more dependable. In the first step of subject selection, 1743 freshman, sophomore, junior, and senior students all majoring in English in a number of Iranian universities and colleges took the Group Embedded Figures Test (GEFT). Their scores on the GEFT revealed that 582 of them were Field-Independent (FI), 707 were Field-Dependent (FD), and 454 were Mixed Field (MF) people. The 454 MF participants were discarded from the study. As such, the sample had two major subgroups: FD with 707 members, and FI with 582 members.

In the second step, both the FD and FI participant groups took the 2000 version (Test 4) of the IELTS (University of Cambridge Local Examinations Syndicate, 2000). The raw scores of these participants on the IELTS were used for classifying them into four proficiency groups. The method used for this step was the capabilities of SPSS for collapsing continuous variables into groups (See Pallant, 2001, pp. 81-84). The 25, 50, and 75 percentiles were calculated for

the IELTS scores of both FD and FI subgroups. As such, eight proficiency groups were identified: four for the FD participants—namely, low-proficient, semi-proficient, fairly-proficient, and proficient; and four for the FI participants—namely, low-proficient, semi-proficient, fairly-proficient, and proficient.

In the third step, participants from the same proficiency group but from different cognitive styles were matched on the basis of their IELTS raw scores. This was done to ensure maximum correspondence between the FD and FI participants in terms of language proficiency; for each IELTS score in the FD group, it was of vital importance to have a corresponding score in the FI group. As such, there was a one-to-one correspondence between IELTS scores in FD and FI groups. That is, each IELTS score in the FI group had a counterpart in the FD group; Individuals with scores which had no counterparts in either the FD or the FI groups were discarded from the study. For example, if a participant from the low-proficiency FD group had scored 13 on the IELTS but no one from the low-proficiency FI group had scored the same, that participant was discarded from the study.

In the last step, from each proficiency group in each cognitive style 36 participants were selected by means of the matching technique. For example, if one participant with a raw IELTS score of 13 from the low-proficiency FD group was chosen, one participant with a raw IELTS score of 13 from the low-proficiency FI group would also be chosen. For each proficiency group, 36 participants were selected in this way. Therefore, for each of the eight subgroups under study, there were 36 participants. As such, the final sample group of the study included 288 participants: 144 participants in the FI group (36 non-proficient, 36 semi-proficient, 36 fairly proficient, and 36 proficient), and 144 participants in the FD group (36 non-proficient, 36 semi-proficient, 36 fairly proficient, and 36 proficient).

The next step was to obtain a sample of EFL writing performance. To this end, the scores participants obtained on the "writing module" of the 2000 version of the IELTS were used as the data for the study. Five EFL teachers with an average of 15.3 years of teaching and assessing EFL writing were asked to use the "multiple trait scoring inventory" (see Appendix A) to assign scores to participants' compositions. As such, each participant received a score from each of these raters; the average of theses five scores was then used as that participants' writing score. The scores were then input to the "independent-samples t-test" statistic; since the totals for the overall test score and scores for language, content, and structure were not the

same, the scores were first converted into a scale of 100 and then were input into the t-statistic analysis.

5.2. Instruments

The instruments used for subject selection and data collection in this study included (a) The Group Embedded Figures Test (GEFT), (b) The 2000 version of IELTS.

5.2.1. The Group Embedded Figures Test (GEFT)

The Group Embedded Figures Test (GEFT) was used to identify participants' FD/FI cognitive styles. The GEFT instrument has been developed by Witkin, Raskin, and Oltman (1971). They reported a Spearman-Brown reliability coefficient of 0.82 for their instrument. The GEFT instrument contains three sections with 25 complex figures from which participants are asked to identify eight simple forms (labeled A to H). Section one of GEFT includes seven complex figures and sections two and three include nine complex figures each. The respondents are asked to find the simple forms (A to H) in the complex figures, and to trace them in pencil directly over the lines of the complex figures. The simple forms are present in the complex figures in the same size, the same proportions, and facing in the same direction as when they appear alone. In their study, Witkin, et al. (1971) reported a mean GEFT score of 12.0 for males (N=155) and a mean of 10.8 for females (N=242). The grand mean of participants in their study was 11.3. In 1980, Panek, Funk, and Nelson reanalyzed data from a previous investigation to determine the reliability and validity of the Group Embedded Figures Test (GEFT). They found that GEFT had adequate split-half reliability. They also noticed that estimates of internal consistency and construct validity for GEFT were adequate and satisfactory. Other studies that have reported adequate reliability and validity for GEFT include Cano, Garton, and Raven (1992), Brenner (1997), and Sexton and Raven (1999). For the purposes of this study, participants were identified as either field dependent (FD), mixed field (MF), or field independent (FI). Using the SPSS commands for collapsing a continuous variable into groups, I classified participants with GEFT scores below the 33.33 percentile into the FD group, those with GEFT scores above the 66.67 percentile into the FI group and those with GEFT scores in between into the MF group (See Pallant, 2001, pp. 81-84).

5.2.2. The IELTS

One of the steps of the present study was to assess the participants' level of proficiency. The instrument used to this end was the 2000 version of the IELTS. Based on their scores on the IELTS, the participants were classified into for proficiency groups: non-proficient, semi-proficient, fairly proficient, and proficient. Here again, the SPSS commands for collapsing a continuous variable into groups were used (See Pallant, 2001, pp. 81-84). This time, the SPSS was asked to afford four equal groups based on 25, 50, and 75 IELTS percentiles.

The writing module of this version of the IELTS includes consists of two tasks: (a) one based on an illustration that shows the figures for imprisonment in five countries between 1930 and 1980, and (b) a composition in which test takers agree/disagree with an opinion. The scores obtained by participants on this second task were used as the data for the present study (See Appendix B).

6. Results

One question addressed by the present study was whether there was a significant difference in the mean test scores for FD and FI individuals within the same proficiency group. Therefore, an independent-samples t-test was conducted to compare the composition scores for FD and FI individuals. The results revealed that, in the case of the low-proficient participants, there was no significant difference in scores for FD participants (M=32.29, SD=5.66), and FI participants [M=33.91, SD=04.66; t(70)=-01.325, p=0.195]. The magnitude of the differences in the means was small (Eta squared = 0.024). The guidelines (proposed by Cohen, 1988) for interpreting Eta squared values are: 0.01=small effect, 0.06=moderate effect, and 0.14=large effect. Expressed as a percentage, (Eta squared value multiplied by 100), only 2.40% of the variance in test performance was explained by cognitive style (see Pallant, 2001, p. 181). As for the semi-proficient group, the results revealed that there was a significant difference in scores for FD participants (M=55.38, SD=6.06), and FI participants [M=49.88, SD=4.07; t(70)=4.51, p=.0005]. The magnitude of the differences in the means was very large (Eta squared=.2895). 28.95% of the variance in test performance was explained by cognitive style. In the case of fairly-proficient participants, a significant difference was observed in scores for FD participants (M=69.56, SD=4.32), and FI participants [M=75.81, SD=4.11; t(70)=-6.285, p=.0005]. The magnitude of the differences in the means was very large (Eta squared=.3607). 36.07% of the variance in test performance was explained by cognitive style. Finally, in the

case of proficient individuals, too, the results revealed that there was a significant difference in scores for FD individuals (*M*=92.59, *SD*=5.92), and FI participants [*M*=89.46, *SD*=4.13; *t*(70)=2.596, *p*=.0125]. The magnitude of the differences in the means was almost large (Eta squared=.0878). 8.78% of the variance in test performance was explained by cognitive style (See Tables 1 and 2).

Notice that the first section of the independent samples Test table in SPSS output provides the results of Levene's test for equality of variances; if the Sig. value for Levene's test is larger than 0.05, the first line in the output table should be used (i.e., Equal Variances Assumed). If this value is = 0.05 or smaller, the second line in the output table should be used (i.e., Equal Variances Not Assumed). This line of the table provides an alternative t-value which compensates for the fact that the variances for the two groups are not the same (see Pallant, 2001, p. 179). In my tables that report the results of the independent samples t-test, the F and t values for Levene's test are not reported. I have preferred to report only the appropriate lines from the t-test output tables (of SPSS). Also notice that Eta squared can range from 0 to 1 and represents the proportion of variance in the dependent variable that is explained by the independent (group) variable. SPSS does not provide Eta squared values for t-tests. The formula for Eta squared (Pallant, 2001, p. 180) is as follows:

$$\text{Eta squared} = \frac{t^2}{t^2 + (N1 + N2 - 2)}$$

Table 1

Group Statistics for Test Performance as the Dependent Variable

Proficiency	Cognitive Style	N	Mean	SD	Std. Error of Mean
Non-Proficient	FD	36	32.2917	5.66728	0.94455
	FI	36	33.9120	4.66143	0.77691
Semi-Proficient	FD	36	55.3819	6.06625	1.01104
	FI	36	49.8843	4.07471	0.67912
Fairly-Proficient	FD	36	69.5602	4.32089	0.72015
	FI	36	75.8102	4.11509	0.68585
Proficient	FD	36	92.5926	5.92288	0.98715
	FI	36	89.4676	4.13513	0.68919

Table 2

Independent Samples T-Test for Test Performance as the Dependent Variable

Proficiency	t	df	sig. (2-tailed)	Eta squared	Variance %
Non-Proficient	-1.325	70	0.190	0.0240	02.40
Semi-Proficient	4.514	70	0.000*	0.2895	28.95
Fairly Proficient	-6.285	70	0.000*	0.3607	36.07
Proficient	2.596	70	0.012*	0.0878	08.78

Another question under study was whether there was a significant difference in the mean "content" scores for FD and FI individuals. Therefore, an independent-samples t-test was conducted to compare the "content" scores for FD and FI individuals. The results indicated that there was a significant difference between FD and FI participants in all proficiency groups except for the 'proficient' subjects. In the case of the low-proficient individuals, there was a significant difference in scores for FD participants (M=31.94, SD=7.28), and FI participants [M=35.24, SD=5.62; t(70)=-2.151, p=.0355]. The magnitude of the differences in the means was medium (Eta squared=.0619). 6.19% of the variance in this case was explained by cognitive style. As for the semi-proficient group, the results revealed that there was a significant difference in scores for FD participants (M=56.25, SD=7.16), and FI participants [M=41.66, SD=5.78; t(70)=9.501, p=.0005]. The magnitude of the differences in the means was very large (Eta squared=.5632). 56.32% of the variance was explained by cognitive style. In the case of fairly-proficient individuals, a significant difference was observed in scores for FD participants (M=70.83, SD=4.95), and FI participants [M=73.95, SD=7.08; t(70)=-2.168, p=.034]. The magnitude of the differences in the means was medium (Eta squared=.0629). 6.29% of the variance was explained by cognitive style. Finally, in the case of proficient individuals, the results showed that there was no significant difference in scores for FD participants (M=92.36, SD=6.35), and FI participants [M=92.53, SD=4.68; t(70)=-.132, p=.8955]. The magnitude of the differences in the means was very small (Eta squared=.0002). .02% of the variance was explained by cognitive style (See Tables 3 and 4).

Table 3

Group Statistics for Content as the Dependent Variable

Proficiency	Cognitive Style	N	Mean	SD	Std. Error of Mean

Low-Proficient	FD	36	31.9444	7.28529	1.21421
	FI	36	35.2431	5.62059	0.93676
Semi-Proficient	FD	36	56.2500	7.16514	1.19419
	FI	36	41.6667	5.78638	0.96440
Fairly-Proficient	FD	36	70.8333	4.95516	0.82586
	FI	36	73.9583	7.08683	1.18114
Proficient	FD	36	92.3611	6.35819	1.05970
	FI	36	92.5347	4.68171	0.78028

Table 4

Independent Samples T-Test for Content as the Dependent Variable

Proficiency	t	df	sig. (2-tailed)	Eta squared	Variance %
Low-Proficient	-2.151	70	0.035*	0.0619	06.19
Semi-Proficient	9.501	70	0.000*	0.5632	56.32
Fairly Proficient	-2.168	70	0.034*	0.0629	06.29
Proficient	-0.132	70	0.895	0.0002	00.02

The third question addressed by the present research was whether there was a significant difference in the mean "structure" scores for FD and FI individuals. Therefore, another independent-samples t-test was conducted to compare the "structure" scores for FD and FI individuals. The results indicated that there was a significant difference between FD and FI individuals in all proficiency groups. In the case of the low-proficient participants, there was a significant difference in scores for FD participants (M=29.51, SD=5.09), and FI participants [M=32.81, SD=7.52; t(70)=-2.178, p=.0335]. The magnitude of the differences in the means was medium (Eta squared=.0634). 6.34% of the variance in this case was explained by cognitive style. As for the semi-proficient group, the results revealed that there was a significant difference in scores for FD participants (M=55.72, SD=12.79), and FI participants [M=61.45, SD=10.61; t(70)=-2.067, p=.0425]. The magnitude of the differences in the means was very close to medium (Eta squared=.0575). 57.5% of the variance was explained by cognitive style. In the case of fairly-proficient individuals, a significant difference was observed in scores for FD participants (M=69.27, SD=7.52), and FI participants [M=75.34,

SD=4.71; *t*(70)=-4.106, *p*=.0005]. The magnitude of the differences in the means was large (Eta squared=.1940). 19.40% of the variance was explained by cognitive style. Finally, in the case of proficient individuals, too, the results showed that there was a significant difference in scores for FD participants (*M*=92.88, *SD*=8.98), and FI individuals [*M*=88.71, *SD*=8.29; *t*(70)=2.045, *p*=.0455]. The magnitude of the differences in the means was very close to medium (Eta squared=.0563). 5.63% of the variance was explained by cognitive style (See Tables 5 and 6).

Table 5 ***5.2.1. The Group Embedded Figures Test (GEFT)***

5.2.1. The Group Embedded Figures Test (GEFT)

Group Statistics for Structure as the Dependent Variable

Proficiency	Cognitive Style	N	Mean	SD	Std. Error of Mean
Low-Proficient	FD	36	29.5139	5.09094	0.84849
	FI	36	32.8125	7.52600	1.25433
Semi-Proficient	FD	36	55.7292	12.7979	2.13297
	FI	36	61.4583	10.6171	1.76952
Fairly-Proficient	FD	36	69.2708	7.52600	1.25433
	FI	36	75.3472	4.71141	0.78524
Proficient	FD	36	92.8819	8.98322	1.49720
	FI	36	88.7153	8.29418	1.38236

Table 6

Independent Samples T-Test for Structure as the Dependent Variable

Proficiency	**t**	**df**	**sig. (2-tailed)**	**Eta squared**	**Variance %**
Low-Proficient	-2.178	70	0.033*	0.0634	06.34
Semi-Proficient	-2.067	70	0.042*	0.0575	05.75
Fairly Proficient	-4.106	70	0.000*	0.1940	19.40
Proficient	2.045	70	0.045*	0.0563	05.63

The last question addressed by the present research was whether there was a significant difference in the mean "language" scores for FD and FI participants. Therefore, another independent-samples t-test was conducted to compare the "language" scores for FD and FI participants. The results indicated that there was a significant difference between FD and FI individuals in all but the 'low-proficient' proficiency groups. In the case of the low-proficient participants, there was no significant difference in scores for FD participants (*M*=35.41, *SD*=9.79), and FI participants [*M*=33.68, *SD*=6.72; *t*(70)=.876, *p*=.3845]. The magnitude of the differences in the means was small (Eta squared=.0108). 1.08% of the variance in this case was explained by cognitive style. As for the semi-proficient group, the results revealed that there was a significant difference in scores for FD individuals (*M*=54.16, *SD*=7.47), and FI participants [*M*=46.52, *SD*=8.24; *t*(70)=4.12, *p*=.0005]. The magnitude of the differences in the means was large (Eta squared=.1951). 19.51% of the variance was explained by cognitive style. In the case of fairly-proficient individuals, a significant difference was observed in scores for FD participants (*M*=68.57, *SD*=6.24), and FI participants [*M*=78.12, *SD*=5.066; *t*(70)=-7.123, *p*=.0005]. The magnitude of the differences in the means was very large (Eta squared=.4202). 42.02% of the variance was explained by cognitive style. Finally, in the case of proficient individuals, too, the results showed that there was a significant difference in scores for FD participants (*M*=92.53, *SD*=7.44), and FI individuals [*M*=87.15, *SD*=3.93; *t*(70)=3.835, *p*=.0005]. The magnitude of the differences in the means was large (Eta squared=.1736). 17.36% of the variance was explained by cognitive style (See Tables 7 and 8).

Table 7

Group Statistics for Language as the Dependent Variable

Proficiency	Cognitive Style	N	Mean	SD	Std. Error of Mean
Low-Proficient	FD	36	35.4167	9.79705	1.63284
	FI	36	33.6806	6.72777	1.12130
Semi-Proficient	FD	36	54.1667	7.47018	1.24503
	FI	36	46.5278	8.24356	1.37393
Fairly-Proficient	FD	36	68.5764	6.24752	1.04125
	FI	36	78.1250	5.06652	0.84442

Proficient	FD	36	92.5347	7.44315	1.24053
	FI	36	87.1528	3.93713	0.65619

Table 8

Independent Samples T-Test for Language as the Dependent Variable

Proficiency	t	df	sig. (2-tailed)	Eta squared	Variance %
Low-Proficient	0.876	70	0.384	0.0108	01.08
Semi-Proficient	4.120	70	0.000*	0.1951	19.51
Fairly Proficient	-7.123	70	0.000*	0.4202	42.02
Proficient	3.835	70	0.000*	0.1736	17.36

7. Discussion

A close look at the results reported in tables 1 through 8 suggests that test takers' cognitive styles result in statistically significant differences in test performance. Although non-proficient individuals' cognitive styles accounted for 2.4% of the variance observed in their composition scores, the effect was not large enough to result in a statistically significant difference between FD and FI participants' test performance (p=.1905). The difference for other proficiency groups was statistically significant. As for individuals' overall writing performance, FD/FI affected semi-proficient and fairly-proficient participants' composition scores more than either proficient or low-proficient participants.

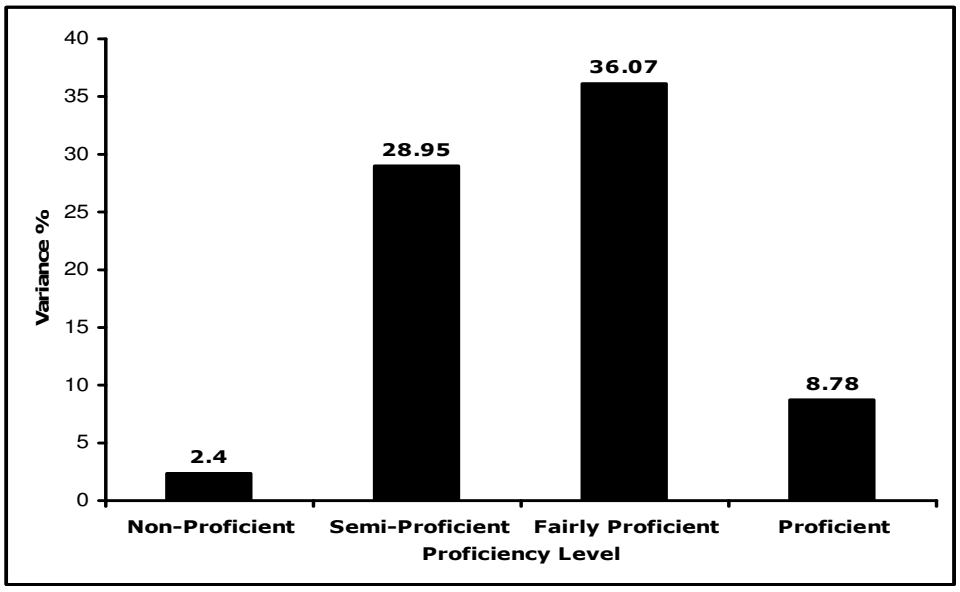

Figure 1. Percentages of variance explained by cognitive style across different proficiency levels for participants' overall writing performance.

Figure 1 compares the percentages of variance that are accounted for by cognitive style across different proficiency levels. In fact, a continuum or cline can be suggested for the effect of cognitive style on individuals' writing performance with minimum effect at the non-proficient end of the continuum and maximum effect at the fairly-proficient end with proficient and semi-proficient people falling in between. The reason why low-proficient participants were not affected by cognitive style could be that they had not reached a threshold level of proficiency that allows room for extraneous factors to apply. The reason why proficient subjects were affected less that fairly- and semi-proficient subjects could be that their proficiency is high enough to disallow such an effect.

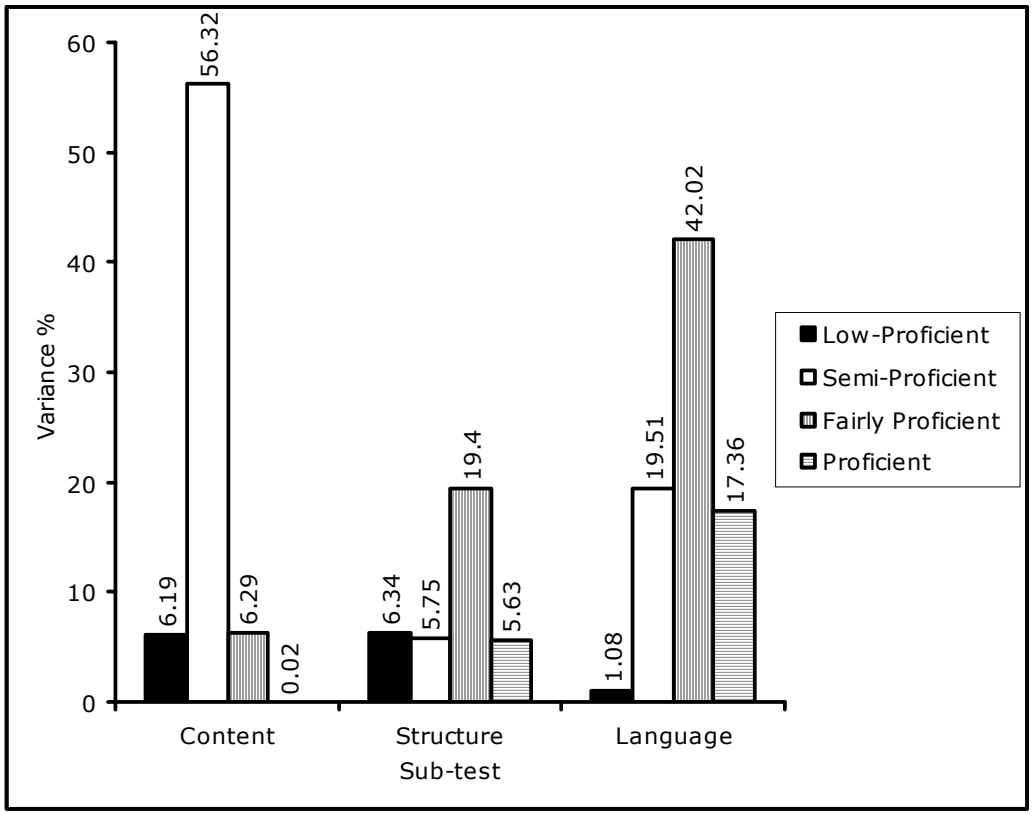

Figure 2. Percentages of variance explained by cognitive style across different proficiency levels.

Figure 2 compares different proficiency groups in relation to content, structure and language aspects of EFL composition. As shown in figure 2, the results revealed that FD/FI was a

factor that affected participants' performance on the different aspects of composition (i.e., structure, content, and language). As for content, the semi-proficient participants appeared to be influenced by cognitive style more the other groups. The reason might be that subjects in this group have just reached a threshold of proficiency that causes a heavy reliance on monitoring. This shows itself in their attention to single sentences at the cost of overall composition texture. In other words, FI semi-proficient subjects, as the raters' evaluation showed, were more focused on sentences than on overall organization. This resulted in their writing of good isolated sentences that failed to stick together to form a unified holistic composition. In earlier studies, this tendency had been referred to as cognitive tunnel vision (Brown, 2000). When it came to structure and language, fairly proficient participants were affected more than the other groups. Here again, the difference lies in the amount of attention that is given to field. An examination of the descriptions presented for both language and structure in Appendix A shows reveals that they require different levels of attention to details. While analytical FI people are experts at attending to isolated parts of a whole, holistic FD people can attend to the overall organization of a field more than its composing parts.

8. Conclusion

The present study attempted to account for the probable effects of FD/FI cognitive style on participants' scores on EFL writing tests. The results showed that cognitive styles imposed their strongest effects on test performance when test takers were fairly proficient. Maybe, fairly proficient test-takers are subconsciously led towards less reliance on monitoring their linguistic performance. More research is required to see if this claim holds true. The study also revealed that the holistic or analytic nature of composition correlated with FD/FI cognitive style. Holistic aspects of composition (e.g. organization) correlated positively with FD style and negatively with FI style; analytic aspects (e.g. sentence-level grammaticality), by way of contrast, correlated positively with FI style and negatively with FD style.

In brief, the results of the study showed that factors other than proficiency are sources of systematic variance in test scores. This finding has implications for test developers; a well-designed test is expected to minimize, if not eradicate, the effects of extraneous factors on test results.

Acknowledgements

I am greatly indebted to Dr Vahid Ghahraman for providing me with a copy of Group Embedded Figures Test (GEFT). I also acknowledge the cooperation and assistance of the many participants of the study who made it possible for me to collect the data for the present study possible. I am also grateful to the anonymous reviewers for the comments and suggestions they made for improving the present paper. To all, many thanks again for their assistance and encouragement.

References

Alderson, J. C. (1991). Language testing in the 1990s: How far have we come? How much further do we have to go? In Anivan, 1991.

Anivan, S. (Ed.). (1991). *Current developments in language testing*. Singapore: SEAMEO Regional Language Center.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brenner, J. (1997). An analysis of students' cognitive styles in asynchronous distance education courses. *Inquiry*, *1*(1), 37-44.

Brown, H. D. (2000). *Principles of language learning and teaching* (4th ed.). White Plains, NY.: Addison Wesley Longman, Inc.

Canale, M., & Swain, M. (1980). Theoretical bases of communication approaches to second language teaching and testing. *Applied Linguistics*. 1(1), 47.

Cano, J., Garton, B. L., & Raven, M. R. (1992). Learning styles, teaching styles and personality styles of preservice teachers of agricultural education. *Journal of Agricultural Education*, Spring 1992, 46-52.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle and Heinle.

Gollnick, D. M., & Chein, P. C. (1994). *Multicultural education in pluralistic society*. New York: Macmillan.

Hamp-Lyons, L. (ed). (1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

Hyland, K. (2003). *Second language writing*. Cambridge: Cambridge University Press.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Pallant, J. (2001). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows (Version 10)*. Philadelphia, PA: Open University Press.

Panek, P. E., Funk, L. G., & Nelson, P. K. (1980). Reliability and validity of the Group Embedded Figures Test across the life span. *Percept Motor Skills, 50*(3), 171-174.

Salmani-Nodoushan, M. A. (2006). Is field dependence or independence a predictor of EFL reading performance? *TESL Canada Journal, 24* (2), 82-108.

Sexton, J., & Raven, M. (1999). The relationship between thinking styles, field dependence and independence, and student performance on selected thinking exercises in an undergraduate agriculture course. *NAERC '99: Research Fresh From Florida*. 561-573. (Proceedings of the 26th Annual National Agricultural Education Research Conference).

University of Cambridge Local Examinations Syndicate. (2000). *Cambridge IELTS 2: Examination papers from the University of Cambridge Local Examinations Syndicate.* Cambridge: Cambridge University Press.

Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

White, E. (1994). *Teaching and assessing writing* (2nd ed.). San Francisco: Jossey-Bass.

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1971). *Group embedded figures test manual*. Palo Alto, CA: Consulting Psychologist Press.

Appendix A: Multiple trait scoring inventory for scoring students' writing

		SCORE	
CONTENT	Explicitness of events	1	Events not stated
		2	Events only sketchy
		3	Events fairly clearly stated
		4	Event explicitly stated
	Documentation of events	1	No recognizable events
		2	Clearly documents events
		3	Includes most events
		4	Clearly documents events
	Evaluation of the significance of events	1	None or confused evaluation
		2	Little or weak evaluation
		3	Some evaluation of events
		4	Full evaluation of events
	Providing personal comment	1	No or weak personal comment
		2	Inadequate personal comment
		3	Some personal comment
		4	Personal comment on events
STRUCTURE	Orientation of the writing assignment	1	Missing or weak orientation
		2	Orientation gives some information
		3	Fairly well-developed orientation
		4	Orientation gives all essential information
	Providing background	1	No background provided
		2	Some necessary background omitted
		3	Most actors and events mentioned
		4	All necessary background provided
	Sequencing	1	Haphazard and incoherent sequencing
		2	Account partly coherent
		3	Largely chronological and coherent
		4	Account in chronological/other order
	Provision of reorientation	1	No reorientation or includes new matter
		2	Some attempt to provide reorientation
		3	Reorientation largely "rounds off" sequence
		4	Reorientation "rounds off" sequence
LANGUAGE	Control of language	1	Little language control
		2	Inconsistent language control
		3	Good control of language
		4	Excellent control of language
	Use of vocabulary	1	Reader seriously distracted
		2	Lacks variety and is verbose
		3	Adequate vocabulary choice
		4	Excellent use of vocabulary
	Choice of grammar	1	Reader seriously distracted
		2	Lacks variety and richness
		3	Adequate grammar choice

		4	Excellent use of grammar
	Appropriateness of tone and style	1	Poor tone and style
		2	Inconsistent tone and style
		3	Mainly appropriate tone and style
		4	Appropriate tone and style

Appendix B: Writing Task 2 from IELTS version 2000 (Test 4)

WRITING TASK 2

You should spend about 40 minutes on this task.

Present a written argument or case to an educated reader with no specialist knowledge of the following topic.

The position of women in society has changed markedly in the last twenty years. Many of the problems young people now experience, such as juvenile delinquency, arise from the fact that many married women now work and are not at home to care for their children.

To what extent do you agree or disagree with this opinion?

You should write at least 250 words.

You should use your own ideas, knowledge and experience and support your arguments with examples and relevant evidence.