Received: June 26, 2011 *Accepted: September 9, 2011*

IQ and Test Format: A Study into Test Fairness

Reza Pishghadam[1], Maryam Sadat Tabataba'ian[2]

Abstract

The aim of this paper was to investigate the relationship between IQ and test format considering test fairness. This study took this relationship into account to see if people (examinees) with different levels of IQ performed differently on different test formats. To this end, 90 advanced learners of English from College of Ferdowsi University of Mashhad, Iran were asked to complete Wechsler's IQ test and a reading test which included four test formats (multiple choice, cloze test, c-test, summary writing) prepared by the researchers. The results of the correlational study indicate a significant relationship between IQ and its subscales and some test formats. The result of the t-test also demonstrated that the differences of the mean between high and low IQ groups were significant regarding certain test formats. The results of the regression equations also indicated that IQ and some of its subscales can also predict performance on certain test formats.

Keywords: *IQ, Reading, Test fairness, Test format*

1. Introduction

Teachers, test developers and researchers have always searched for more valid tests throughout history. One of the important aspects of test validity is test fairness (Bachman, 1990). Assessment systems were developed so that a fairer selection can be offered, based on which teaching and learning can be improved. Test fairness implies that tests should not be biased towards any testee in any form. For example, a test is not fair if its format is biased towards field dependents or extroverts, meaning that some groups of testees with these personality types will outperform other testees with different personality types on the test (Stobart, 2008). As Shohamy (2001) stated, tests that exclude certain groups of people are undemocratic and it is not fair to employ them in education.

As Bachman (1990) indicated cognitive factors, styles and personality types affect learning and performance on certain test formats. Scholars have paid attention to the need for examination of the effect of individual characteristics on test taking process as individual

[1] Department. of English Language and Literature, Ferdowsi University of Mashhad, Iran. E-mail: pishghadam@um.ac.ir

[2] Department of English Language and Literature, Ferdowsi University of Mashhad, Iran. E-mail: m.tabatabaeeyan@gmail.com

characteristics might have an effect on test performance; these differences might be a threat to the validity of the test (Bachman, 2000).

It seems that the relationship between IQ as an important aspect of cognition, and test format has not been examined to date; therefore, this study was set out to investigate the relationship between IQ and different test formats and to see whether IQ and its subscales could predict performance on different test formats. To shed more light on the relationship between test fairness, test format, and issues related to intelligence, we review these notions in the following sections.

2. Review of Literature

2.1 Test Fairness

Scholars have always searched for fairer assessment methods, although fairness cannot be fully guaranteed. Shohamy (2001) has correctly claimed that testing must be under careful examination because tests are mostly used for making high-stake decisions. As language tests are used for making decisions, they must be useful to individuals who live in the society (Fulcher & Davidson, 2007).

Validation must take account of test content, its method, and how test takers perform because tests play an influential role in educational and social decisions about individuals. According to Messick (1989), in discussing validity, attention must be paid to social consequences of a test. (cited in Bachman, 1990). As a result of this new, expanded notion of validity, test fairness matters have been highlighted (Kunnan, 1999). In 1985, the unitary concept of validity was proposed and construct validity was seen to be the most important validity. Messick (1989) attended to both construct validity and test consequences (cited in Chapelle, 1999). Messick's paper (1989) drew the attention of scholars to the consequences of the tests. The expanded notion of construct validity which Messick (1989) proposed includes both evidential and consequential bases for test validation (Kunnan, 1999). This framework is also believed to be a valuable basis for attending to issues of both validity and impact (Bachman, 2000).

According to this framework, tests cannot be validated themselves; rather the inferences regarding specific uses of a test are validated. The use and interpretation of test performance may not be equally valid for all abilities and in all contexts. Sources of bias, namely cultural background, background knowledge, cognitive characteristics and native language, ethnicity, age and gender affect test performance; therefore, they must be avoided (Bachman, 1990). Differential validity also takes test fairness into account. The aim of differential validity is to ensure that no testee will suffer because of the sources of bias and individual characteristics (Weir, 2005). Scholars must ensure that none of the sources of bias affect the measurement as tests are used for making high-stakes decisions (Weir, 2005; Shohamy, 2001). Effort must be made so that the tests are as fair as possible for the groups who want to take them (Stobart, 2005).

2.2 Test Format

One of the many factors that affect test performance is test format. Bachman (1990) proposed a framework for test methods and revised it in Bachman and Palmer (1996). He (1996) stated that

test performance is affected by test method. Test method effect is considered important because it is not known whether test performance is due to the test takers' knowledge or their ability to answer certain formats (Baker, 1989).

Bachman and Palmer (1996) divided method facets – or test task characteristics – into 5 categories:

1. Characteristics of setting;
2. Characteristics of the test rubrics;
3. Characteristics of input;
4. Characteristics of the expected response; and
5. Relationship between input and response.

This study focuses on the expected response which can be of three types: selected response, limited production response, or extended production response (Bachman & Palmer, 1996). Language assessment is also classified into three broad categories based on the response the assessor expects test takers to write:

1. Selected-response assessment
2. Constructed-response assessment
3. Personal-response assessment

Selected-response assessment, which includes true-false, matching, and multiple-choice items, provides the language material for the student and requires the test taker choose the correct answer from the available options. In contrast, in constructed-response assessment, which contains fill-in, short-answer, and performance assessment, test takers are expected to produce language while doing the test. Finally, personal assessment, which includes conference, portfolio, and self- or peer-assessment, unlike the other types of assessment, asks test takers to perform and actually produce language. The answers the students provide can be completely different, and in this type of assessment the students communicate what they really want to communicate. In fact, they are free to express their views (Brown & Hudson, 1998).

To date, the relationship between field independence and language test performance (Hansen & Stanfield, 1981 & 1983), cognitive variables and language proficiency test performance (Chapelle & Roberts, 1986), test response format and text organization in reading comprehension tests (Kobayashi, 2002 & 2004), vocabulary test format and age (Bowels & Salthouse, 2008) and multiple-choice and open-ended test formats in L1 and L2 reading and L2 listening (In'nami & Koizumi, 2009) have been attended to in research regarding different test formats.

2.3 Intelligence

Intelligence is traditionally considered as a real, single, measurable, inborn and unchangeable entity. It was traditionally believed to determine our material success (Jarvis, 2005). Illeris (2008) mentioned that although there have been many definitions of intelligence and these definitions have changed a lot since intelligence first appeared, it is yet hard to define intelligence. He defined intelligence as the ability to learn and think, to understand and solve problems. Dornyei (2005) also defined intelligence as the ability to learn. Binet (1905) was the first person who developed a true IQ test with Simon. Intelligence has always been an important

and controversial issue in education and has long attracted scholars' and teachers' attention (cited in Illeris, 2008).

Resistance to the idea of the existence of one and only one intelligence continued for decades, and different theories were proposed that opposed this tradition (Stobart, 2008). Intelligence was divided into more specific intelligences as the g factor did not account for the total ability of individuals. It only showed the overall intellectual ability of individuals. Different scholars have tried to divide the general intelligence into specific intelligences (e.g., Cattell, 1963; Gardner, 1983, 1993; Guildford, 1967; Sternberg, 1988; Thorndike, 1920; Wechsler, 1987). These different intelligences were all similar to general intelligence but distinct from it (cited in Mayer & Geher, 1996).

Gardner (1983) considers there are different kinds of intelligence. According to Jarvis (2005), Gardner paid attention to a complete range of learner's mental abilities. He questioned the usefulness of general intelligence and proposed a modular approach. He identified seven interdependent intelligences: Linguistic Intelligence (the ability to use language in written and oral forms), Logical/ Mathematical Intelligence (the ability to manipulate numbers and reason), Spatial Intelligence (the ability to form mental models of the world), Kinesthetic/ Bodily Intelligence (having a well-coordinated body and understanding the world through body), Musical Intelligence (the ability to recognize and produce music), Interpersonal Intelligence (the ability to work well with people), and Intrapersonal Intelligence (the ability to know oneself and be able to act adaptively based on this self-knowledge) (Armstrong, 2000; Nolen, 2003). He later (1999) added a number of other intelligences to these seven intelligences (Naturalist Intelligence, Existentialist Intelligence, Spiritual Intelligence) (Sternberg, 2004). He stated that there might also be other intelligences that might be added to the list and some of these identified intelligences may no longer be qualified to be called an intelligence.

Sternberg (1985) also redefined intelligence and proposed a triarchic theory of intelligence (Williams & Burden, 1997). According to Jarvis (2005), Sternberg proposed three types of smartness. First, componential intelligence which includes three components: knowledge acquisition component which involves curiosity and affects our learning strategy, performance component which determines our actual ability and metacomponents which are the conscious higher mental processes. Second, experiential intelligence which is concerned with the effect of experience on our intelligence. Third, contextual intelligence which regards intelligence within its cultural context. This type regards what we actually use intelligence for.

Sternberg (2004) claimed that Successful Intelligence is more important than traditional notions of intelligence. Deary and Smith (2004) claimed that Sternberg's model of component function explained the differences between individuals in performing reasoning tasks well. Sternberg (1985) drew the attention of scholars to creative and practical intelligences by proposing this new theory of intelligence (Salovey, Mayer & Caruso, 2002).

It must be stated that there is no agreement regarding the definition of intelligence, and the multicomponential nature of mental abilities shows that we can expect some mental variation within the individuals with regard to their specific mental abilities (Dornyei, 2005).

3. Research Questions

As tests play an important role in teaching and as test fairness is one of the considerations of test developers, this study aimed at seeking the relationship between IQ and test format. In this study,

the aim was to focus on test formats used more often in different tests (Cloze test, C-test, Summary, and M.C.). Therefore, this study was set out to answer the following three questions:

Q1: Is there a significant relationship between IQ and performance of advanced language learners on different test formats?

Q2: Is there a significant difference between the means of high and low IQ groups, regarding performance of advanced language learners on different test formats?

Q3: Do IQ scores predict performance of advanced language learners on different test formats significantly?

4. Methodology

4.1 Participants

One hundred and fifty English learners from College of Ferdowsi University of Mashhad took part in this study. Only 90 learners were at the same level of reading proficiency after they were homogenized. As the participants were expected to be able to speak and write English with a good command of grammatical structures and adequate vocabulary, upper-intermediate and advanced students were chosen. All participants were familiar with the four test formats which were the concern of this study because they were studying English in one institute and all these formats could be seen in the tests the students took. To obtain more accurate results, the participants' reading comprehension proficiency was homogenized using the reading part of standard paper-based TOEFL test administered in 2003.

Out of the 90 participants, 50 were female and 40 were male. The participants' age ranged from 18 to 47, majoring in different fields of study at university. The participants of the study had agreed with anonymous publication of data before they took part in the study.

4.2 Instrumentation

Researchers used two instruments to collect the data: Wechsler's Adult Intelligence Scale (WAIS) III and a reading test.

To measure IQ of the subjects, Wechsler's Adult Intelligence Scale (WAIS) III (1981) was used. The test consists of two scales: verbal and performance. The verbal scale is composed of Information (29 items), Digit Span (14 items), Vocabulary Knowledge (40 items), Arithmetic (14 items), Comprehension (14 items) and Similarities (13 items) subscales. The Verbal scale of WAIS is used to check the Crystallized intelligence (knowledge and skills related to education and experience) of individuals. The performance part, which measures Fluid intelligence (the ability to see relationships, as in analogies and letter and number series), consists of Picture Comprehension, Picture Arrangement, Block Design, Digit Symbol and Object Assembly.

In the Information section of the test, a testee is presented with one question at a time and asked to respond to it. The testees' responses receive 1 point for correct responses and 0 for incorrect responses, allowing a range of scores from 0 to 29. In the Digit Span section of the test, testees listen to sets of numbers, and are asked to repeat them on the spot. The testees' responses receive 1 point for correct responses and 0 for incorrect responses and the range of scores is from 0 to 17. In the Arithmetic subtest of the test, testees are provided with some questions one at a

time, and are required to respond to them in the allotted time. The testees' responses receive 1 point for correct responses and 0 for incorrect responses allowing a range of scores from 0 to 18 as the four last questions receive 0, 1 or 2. In the Comprehension part of the test, testees are provided with some questions one at a time and are asked to answer them. The examinees' responses receive 0, 1, or 2 points, depending on how well they may answer the questions, allowing a range of scores from 0 to 28. In the Similarities section of the test, examinees are asked to find the similarities between some words. The examinees' responses receive 0, 1, or 2 points depending on how exact the responses are, allowing a range of scores from 0 to 26.

The single most frequently used test to establish a level of verbal intellectual (VI) functioning is the Vocabulary subtest of the WAIS-III. The WAIS Vocabulary subsection consists of 40 words. An examinee is presented with 1 word at a time and asked to define each word's meaning. The examinee's responses receive 0, 1, or 2 points, depending on how well he or she defines the word, allowing a range of scores from 0 to 80. The Vocabulary subtest is quick to administer, correlates highly (.91-.95) with the Verbal Scale of the WAIS-III, and comes with extensive normative data. The reliability coefficient is .97 for the Verbal IQ.

In this study, we used the translated version of the WAIS-III, which was prepared by Azmoon Padid Institute (1993) in Tehran, Iran to measure participants' IQ. The verbal part of Wechsler's intelligence test was chosen.

IQ might be related to different skills; therefore, the researchers chose reading so that only the test format affects the results of the research, not the skill being tested. Two reading texts were chosen from paper-based TOEFL test which was administered in 2004. The two readings had topics which seemed to be familiar to all students in order for the test results not to be affected by some participants' topic familiarity. The response formats tested were also related to the same text so that different content does not affect test performance (Chen, 2004). Two tests were designed which tested four test formats: multiple-choice questions, cloze test, c-test and summary writing. Multiple-choice questions were chosen as the representative of selected response format as these questions have a fixed answer. The answer to c-test and cloze test is neither too fixed, nor too open-ended. Summarizing was also chosen as the most open-ended type of question with extended productive response.

The multiple-choice questions used in the test were taken from the TOEFL test. Each of the two readings had 10 multiple-choice questions so there were 20 multiple-choice questions to be answered (Cronbach's alpha = 0.523). The cloze test was prepared from the first half of the same reading texts. Every seventh word in the text was deleted (Farhady, Ja'farpur & Birjandi, 1994). The participants were expected to answer 50 cloze test questions (Cronbach's alpha = 0.568). The c-test was prepared from the second half of the same reading texts. Half of every other word was deleted (Raatz & Klein-Braley, 2002). The participants were expected to answer 50 c-test questions (Cronbach's alpha = 0.709). The students were also asked to write a summary of the reading texts in their own language. Summary writing was chosen instead of open-ended questions because as Kobayashi (2002) also states, summary writing measures overall comprehension better than open-ended questions. Like in Kobayashi's study (2002), the participants were asked to write the summary in L1 so that the raters were not affected by their writing proficiency in English. Writing the summary in L1 also helped eliminate undesirable factors from performance on summary writing. Test takers could also refer back to the text so that the recall factor is eliminated. The summaries were rated by two raters and the interrater reliability was calculated to be 0.87. The test was piloted with 6 participants whose proficiency

level was the same as the target population before it was given to the participants to make sure that it was suitable for the students' level of proficiency.

4.3 Procedure

The data collection started in August 2009, and it ended in October, 2009. One hundred and fifty language learners, who had once been homogenized by the institute as they were studying at certain levels, took part in the reading part of an actual paper-based TOEFL test (2003) to ensure the homogeneity of the participants regarding their reading proficiency. There were five readings on the test and each had 10 questions. On the whole, the participants answered 50 multiple-choice reading questions, and it took about 90 minutes. The normal curve was drawn for the obtained grades and the students whose grades were within one standard deviation below and above the mean were chosen (N = 90 learners).

The reading test was then given to the chosen participants. As the reading tests were long, they were administered in two sessions. Each of them took about 70 minutes. First, the cloze test and the c-test parts were given to participants so that their memory did not affect test results. After that, they were asked to take the multiple-choice test and write the summary of the text. The IQ test was also given to the participants so that their IQ was obtained. The WIAS-III was held as an interview, and it took about 50 minutes to administer it. They were asked to attend the institute and then they were interviewed.

The collected data were entered into and processed with SPSS 17. The results gained from the tests taken by the participants fell within the interval data so the Pearson Product moment formula was used to calculate the correlation between each test format and IQ scores. According to the results of the participants' performance on IQ, two groups (high (N=45) and low (N=45)) were formed and the t-test was run to see whether the difference between the means of the high and low IQ groups was significant. Multiple regression analysis was also used to see which subscales of IQ were better predictors of performance on each test format.

5. Results

The first question of this study was whether there was a relationship between test format and IQ. The following table shows the result of the correlational analysis.

Table 1. The Results of Correlational Analyses between Test Format and IQ Subscales

	Information	Comprehension	Arithmetic	Similarity	Digit	Vocabulary	Total IQ
M.C.	-.011	.193	.322**	.100	.021	.084	.178
Cloze	.227*	.331**	.502**	.196	.107	.112	.378**
C-Test	.183	.341**	.342**	.171	.193	.186	.375**
Summary	.103	.151	.373**	.201	-.011	.307**	.343**

* p < .05, ** p < .01

As Table 1 presents, there is a moderate correlation between IQ and cloze test (r = .378, p < .05), IQ and c-test (r = .375, p < .05) and finally IQ and summary writing (r = .343, p. < 05).

Out of the six subscales of IQ, Information correlated moderately with performance on cloze test ($r = .227$, $p < .05$). Comprehension also correlated moderately with performance on cloze test ($r = .331$, $p < .05$) and c-test ($r = .341$, $p < .05$). We find it interesting that Arithmetic IQ correlated with all test formats: Arithmetic IQ and multiple-choice ($r = .322$, $p < .05$), Arithmetic IQ and cloze test ($r = .502$, $p < .05$), Arithmetic IQ and c-test ($r = .342$, $p < .05$) and arithmetic IQ and summary writing ($r = .373$, $p < .05$). Finally, Vocabulary Knowledge (Verbal IQ) correlated only with summary writing ($r = .307$, $p < .05$).

To answer the second question, t-test was also run to see if there is any significant difference between the means of the high and low IQ groups. The results of the t-test study are shown in the following table.

Table 2. Comparisons of Performance on Different Test Formats Based on Performance on IQ Test

Variables	High IQ Group (n=45) Mean	Low IQ Group (N=45) Mean	T
M. C.	31.1667	29.3889	1.209
Cloze	22.4889	19.7333	2.778
C-Test	27.1111	23.7333	2.747
Summary	40.2611	34.6222	3.889

As the results show, the difference between the means of the two high and low IQ groups is not significant regarding the performance on multiple-choice questions ($t = 1.209$, $p > .05$). However, the differences between means are significant regarding the performance on all other test formats.

Finally, to answer the third research question, multiple regression analysis was run using IQ as the predictor of performance on the four test formats.

Table 3 Multiple Regression Analyses Predicting Performance on Multiple-Choice by IQ

Predictors	R	R^2	Adjusted R^2	F	P	B
Total IQ	.974	.950	.949	1676.396	.000	.219
Arithmetic	.322	.104	.094	10.201	.002	.955

Table 3 shows that Total IQ accounts for about 95% of the total variance in performance on multiple-choice test format ($R^2 = .950$, $p < .05$) and out of the subscales of IQ, Arithmetic IQ accounts for about 10% of the total variance in performance on multiple-choice test format ($R^2 = .104$, $p < .05$). Therefore, having a high IQ and a high arithmetic IQ were the best predictors of performance on multiple-choice questions.

Table 4. Multiple Regression Analyses Predicting Performance on Cloze Test by IQ

Predictors	R	R^2	Adjusted R^2	F	P	B
Total IQ	.978	.957	.957	1981.141	.000	.153
Comprehension	.982	.964	.963	1174.312	.000	.392
Arithmetic						.998

Table 4 indicates that Total IQ accounts for about 96% of the variance in performance on cloze test format (R^2 = .957, p < .05) and out of the six subscales of IQ, Comprehension and Arithmetic IQ also account for about 96% of the variance in performance on cloze test format. Having high IQ, Comprehension and Arithmetic IQ were the best predictors of performance on cloze test questions.

Table 5. Multiple Regression Analyses Predicting Performance on C-Test by IQ

Predictors	R	R^2	Adjusted R^2	F	P	B
Total IQ	.977	.955	.954	1867.425	.000	.185
Digit	.978	.957	.956	651.694	.000	.045
Comprehension						.003
Arithmetic						.009

As shown in Table 5 scores on IQ can predict about 95% of the total variance in performance on c-test questions (R^2 = .955, p < .05). It is also shown that Digit, Comprehension and Arithmetic subscales of IQ account for about 96% of the total variance in performance on c-test format (R^2 = .957, p < .05). Having high IQ and high Digit, Comprehension and Arithmetic IQ were the best predictors of performance on c-test questions.

Table 6. Multiple Regression Analyses Predicting Performance on Summary Writing by IQ

Predictors	R	R^2	Adjusted R^2	F	P	B
Total IQ	.983	.967	.966	2595.672	.000	.271
Vocabulary	.985	.970	.969	1402.806	.000	.375
Arithmetic						1.055

As Table 6 indicates IQ scores can predict about 97% of the total variance in performance on summary writing (R^2 = .967, p < .05) and Vocabulary and Arithmetic subscales of IQ account for about 97% of the total variance in performance on summary writing (R^2 = .970, p < .05). Having high IQ and high Vocabulary and Arithmetic IQ were the best predictor of performance on summary writing.

6. Discussion

The aim of the present study was first to investigate the relationship between IQ and test format, second to see if a significant difference between the means of high and low IQ groups existed, and finally to see how much IQ and its subscales predicted performance on different test formats.

In the present study, the correlational analysis showed that performance on multiple-choice questions is not related to IQ, while performance on cloze test, c-test, and summary writing are all related to IQ. Out of the six subscales of IQ, there was a significant relationship between Information and cloze test, Comprehension and cloze test and c-test, and Vocabulary knowledge and summary writing. Arithmetic section of IQ also correlated with all test formats. These results are justifiable if we look into the nature of these tests and modules of IQ. Cloze tests are holistic tests which require general background knowledge to be taken more effectively. That is why; Information and Comprehension correlate with them. In a similar vein, due to the nature of c-tests, Comprehension is of vital importance to take the incomplete words. Moreover,

it is fair to state that making a summary requires full knowledge of Vocabulary, because when one summarizes a text, they must have the required words at hand to shorten it without copying from the text. As we know, there is a relationship between language and mathematics (Farhady et al., 1994), and therefore, it is quite natural that we observe Arithmetic is influential in all test formats. And finally, multiple choice questions were found not to be correlated with all modules of IQ except for Arithmetic. Since this type of test is decontextualized, it is fair to say that the only module of IQ contributing to taking it more successfully is the Arithmetic part, which deals with logic.

The results of the t-test analysis showed that those with a higher IQ have performed better than those with a lower IQ in all test formats except for multiple-choice. It was also shown that individuals with higher Vocabulary knowledge have performed better than those with lower Vocabulary knowledge on summary writing test format.

The results of the regression equations done on these four test formats also indicated that IQ and Arithmetic IQ can significantly predict success in performance on multiple-choice questions. Performance on cloze test could also be predicted by IQ, Comprehension, and Arithmetic IQ. Of the two subscales, Arithmetic IQ is a better predictor. C-test was the third test format that was studied in this study. Performance on c-test could be significantly predicted by IQ, Comprehension, Arithmetic IQ, and Digit Span. The obtained results regarding summary writing, which was the last test format to be analyzed, showed that IQ, Arithmetic IQ, and Vocabulary knowledge were good predictors of success in this test format. Arithmetic IQ seemed to be the subscale of IQ that can predict performance on all these four test formats.

The results of the present study were not in line with the study done by Raatz (2002) who had investigated the influence of intelligence on c-test performance. In his study, Vocabulary knowledge was found to be highly related to performance on c-test while in the present study no relationship was found between performance on c-test and Vocabulary knowledge. The difference can be because this study is concerned with foreign language test performance while Raatz had focused on first language test performance.

The results of the present study confirmed Alderson's ideas, (2000) who pointed to the fact that employing only one method for measuring the understanding of the text is not adequate and objective and subjective methods of evaluation must be utilized side by side. According to him, good reading tests are the ones that use different techniques for assessing reading comprehension skills (cited in Weir, 2005).

Moreover, our findings support the claims made by Stobart (2008) and Brown (2004). They believe that standardized tests do not guarantee test fairness, and therefore, these tests must be scrutinized to guarantee fairness for all participants. If one attends to the results of this and other similar studies, the designed tests will be more valid as the consequential validity of the tests, which is an important measure of validity and has been ignored until recently, will increase. And as Bachman (2000) stated, now that we have the methodological, theoretical and technological resources, plans must be made to ensure validity in practice and high quality tests must be developed.

As tests with one format disadvantage a group and there might be other sources of bias, it will be best to use a mixture of different test types. Employing different test types will result in a more complete picture of the students' ability (Brown & Hudson, 1998). As Hamp-Lyons (1997) stated, testers must accept responsibility for all the consequences they are aware of (cited in Hamp-Lyons, 2000).

Finally, researchers are recommended to carry out research that aims at gaining a more thorough picture of the factors that affect performance on different test formats. The same study with more participants and more reliable tests can be replicated to ensure that the observed results are reliable and valid. Different genders might also perform differently on different tests; therefore, it is worth considering its effect. Skills other than reading comprehension might also correlate differently with IQ, so the effect of IQ on them can also be taken into account. Individuals with different styles of learning and different personality types might also perform differently on different tests and styles might affect individuals' test performance so it is worth taking their effect into account.

References

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

Armstrong, Th. (2000). *Multiple intelligences in the classroom.* Alexandria: Association for Supervision.

Azmoon Padid Institute. (1993). *The standardization of Wechsler's Adult Intelligence Scale III*. Tehran, Iran.

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Baker, D. (1989). *Language testing: A critical survey and practical guide.* London: Hodder & Stoughton Limited.

Binet, A., & Simon, T. (1905). New methods for the diagnostic of the intellectual level of abnormal persons. *L'Annee Psychologique, 11(2),* 191-244.

Bowles, R. P., & Salthouse, T. A. (2008). Vocabulary test format and differential relations to age. *Psychology and Aging, 23*(2), 366-376.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* New York: Longman.

Brown, J. D., & Hudson, Th. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1-22.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*(1), 254-272.

Chapelle, C., & Roberts, C. (1986). Ambiguity tolerance and field dependence as predictors of proficiency in English as a second language. *Language Learning, 36*(1), 27-45.

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: A reply to Kobayashi, 2002. *Language Testing, 21*(2), 228-234.

Deary, I. J., & Smith, P. (2004). Intelligence research and assessment in the United Kingdom. In R. J. Sternberg (Ed.), *International handbook of intelligence.* Cambridge: Cambridge University Press.

Dornyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Mahwah: Lawrence Erlbaum Associates, Inc.

Farhady, H., Jafarpur, A., & Birjandi, P. (1994). *Testing language skills: From theory to practice.* Tehran: The Organization for Researching and Composing University Textbooks for Humanities, SAMT.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* New York: Routledge.

Gardner, H. (1983). *Frames of mind.* New York: Basic Books.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, H. (1999). *Intelligence reframed. Multiple intelligences for the 21st century.* New York: Basic Books.

Guildford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical Concerns. *Language Testing, 14*(3), 295-303.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*(4), 579-591.

Hansen, J., & Stansfield, C. (1981). The relationship between field independent-dependent cognitive styles and foreign language achievement. *Language Learning, 31(4),* 349-367.

Illeris, K. (2008). *How we learn: Learning and non-learning in school and beyond.* New York: Routledge.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244.

Jarvis, M. (2005). *The psychology of effective learning and teaching.* Cheltenham: Nelson Thornes Ltd.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220.

Kobayashi, K. (2004). *The effects of test methods on reading test scores of Chinese students learning Japanese as a foreign language.* Unpublished master's thesis. Ochanomizu University, Japan.

Kunnan, A. J. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics, 19(2),* 235-253.

Mayer, J. D., & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence, 22*(3), 89-113.

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

Nolen, J. L. (2003). Multiple intelligences in the classroom. *Education, 124*(1), 115-119.

Raatz, U. (2002). C-tests and intelligence. In J. A. Coleman, R. Grotjahn & U. Raatz (Eds.), *University language testing and the c-test.* Bochum: AKS-Verlag.

Raatz, U., & Klein-Braley, Ch. (2002). Introduction to language testing and to c-tests. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the c-test.* Bochum: AKS-Verlag.

Salovey, P., Mayer, J. D., & Caruso, D. R. (2002). The positive psychology of emotional intelligence. In C. S. Synder & Sh. J. Lopez (Eds.), *Handbook of positive psychology.* New York: Oxford University Press.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391.

Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly*, *17*(1), 29-38.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence.* New York: Viking Penguin.

Sternberg, R. J. (2004). North American approach to intelligence. In R. J. Sternberg (Ed.), *International handbook of intelligence.* Cambridge: Cambridge University Press.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education, 12*(3), 275-287.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* Abingdon: Routledge.

Thorndike, R. L. (1920). Intelligence and its uses. *Harpers' Magazine, 140*(4), 227-235.

Wechsler, D. (1987). *Wechsler memory scale- Revised.* New York: Psychological Corporation.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* London: Palgrave.

Williams, M., & Burden, R. L. (1997). *Psychology for language teachers: A social constructivist approach.* Cambridge: Cambridge University Press.