Towards Integrated Testing Approach: An Application of Cognitive Science and Deep Learning Principle

Tiantian Zhang[1], Quan Zhang[2]
(Deep Education Institute, Wisconsin, USA)

Abstract: The use of multiple-choice (MC) question types has been one of the most contentious issues in language testing. Much has been said and written about the use of MC over the years. However, no attempt has ever been made to introduce any innovation in test item types. The researchers proposed a jumbled words test item (JW) based on cognitive science and deep learning principles, and addressed the feasibility of replacing the type of multiple-choice (MC) question with JW to meet the ongoing rapid development of language testing practice. Two research questions were proposed ad hoc, focusing on the co-relationship between JW and MC scores. RASCH-GZ was used to perform item analyses (Rasch, 1960). The item difficulty parameters thus obtained were used to compare the two different test items. The sample data metric includes 40 Chinese participants. The findings revealed that correlation analysis revealed that the performance of the same group of subjects taking both JW and MC was not relevant (Pearson Corr = 0). This is primarily due to the total elimination of guessing factors inherent in test-takers during JW test performance. Three factors were specified for the design of the JW test: compute program, test difficulty, and score acceptability. These all have three dimensions. Data collected through questionnaires were analyzed using EFA in SPSS V.24.0. KMOs (=0.867) were found to be approximately one and significance at 0.000 (0.05), indicating that the construct of the questionnaire thus designed has better validity for factor analysis. Three important conclusions were obtained, the implications of which could provide impetus for our testing counterparts to practice more precisely and correctly, potentially reshaping our overall language testing practice. Limitations and recommendations for future research were also discussed.

[1]Zhang Tiantian Charles is a MA candidate in Education at Deep Education Institute, Wisconsin, USA.

[2]Prof. Zhang Quan, PhD is a supervisor of the program of education, Deep Education Institute, Wisconsin, USA. qzhang141@aliyun.com

Keywords: JW; MC; integrated testing; declarative knowledge; procedural knowledge, deep learning, Rasch-GZ

1. Research Background

The testing domain experienced two significant transitions in the 1980s: one from classical testing theory (CTT) to item response theory (IRT), indicating that the testing theory had matured; and the other from the use of computers for statistical analyses to the use of computer programming technology for testing administration, indicating that the testing method had evolved.

And in recent years, the transition from computer-based testing to Internet-based testing indicates not only that the testing method has been greatly updated, the testing media has been colorfully varied, but also that the testing distance has been enormously extended and stretched out. Despite these changes, the test item used remains largely unchanged, i.e., it is still a multiple-choice (MC) question type. Everyone knows that in China, the number of candidates taking language tests primarily consisting of MC is the highest in the world.

At the national level, there are several large-scale English language tests, including the Matriculation English Test (MET), College English Test Band 4 and Band 6 (CET-4 & 6), Test for English Majors Band 4 and Band 8 (TEM-4 & 8), and the Public English Testing System (PETS).

The total number of test-takers is estimated to be close to 15,000,000. However, in recent years, these tests have received little attention from their testing counterparts and researchers around the world (Wu & Zhang, 2018 and Gui, 2015). There have been ongoing debates about its measurement attributes, validity, and washback effects, and the test scores have yet to be approved by the testing counterparts in North America, Great Britain, Australia, New Zealand, and other countries where most Chinese students are likely to study.

Beginning in 2015, the Chinese University of Hong Kong (CUHK) declared that CET-4 and 6 scores would no longer be accepted (Wu & Zhang, 2018). Aside from that, reports of poor performance with MET scores in some Chinese provinces have garnered international attention in recent years.

Status Quo and Problems

It is true that MC question items have been a dominant and indispensable test item type, particularly in large-scale tests. In fact, it is a compromise (Gui, S. C,1990). Guided by the former, Zhang (2015, 2007, 2002) put forward the idea of using jumbled words test item (JW) to replace MC. However, at that time, due to no effective computer programming skills and facilities, JW could not be feasibly implemented in a real testing environment. In the authors' opinion, in doing scientific research, sometimes proposing a hypothesis is more significant than justifying it because the proposal would arouse more professionals to devote their energy and knowledge to do further research, explore, to seek the answer from the unknown. With cognitive considerations into language testing, supported by today's rapid development of computer and web technology, JW has come of age.

Global Trends

Nowadays, the language testing industry is looking for more and better large-scale assessments based on cognitive science, linguistics, and computer and Internet technology. In general, the trend is away from discrete testing and toward integrated testing.

The current research design is also consistent with the ten global trends presented by Professor Scott Paris (2019) and the paper presentation by Zhang (2019) at the recent ICEMEA. These ten major global trends, according to Prof. Scott Paris, are influencing the selection and training of people in education and the workforce.

2 Research Purpose

In the authors' opinion, most testing practices worldwide to date have been concentrating on the reliability of test scores, the increased efficiency with which test scores are obtained, or methods or formats by which test items are delivered. The nature of the validity of what would be appropriate test items to use has been largely neglected. And so far, little has been addressed to any innovation in terms of test items to be used in computer-based or Internet-based tests. Here in particular, the authors are reminding our testing counterparts of what we should be aware of is more correct rather than more precise practice. It is in this sense that what the authors attempt to address is that, in view of cognitive science supported by the most updated computer programming technology today, MC should not be the only test item used in today's testing practice. JW could totally replace MC. The former attaches greater importance to cognitive considerations to bring revolutionary changes in test items and content via computers, and more significantly, to penetrate the human black box in terms of problem-solving and achieve a more valid and accurate interpretation of the ability parameter. Such is the aim of the present research. It provides a glimpse of the topics treated within the cognitive science, computer programming, and statistics practice being undertaken so far. This study is to validate the JW test item by addressing its cognitive basis, deep learning principles, research hypotheses, advantages of, issues of, and possible solutions to JW by conducting experiments, verifying hypotheses, and reporting the results to the language testing sector. It should be pointed out that the JW test item is presented herein as a basis for further discussion concerning the possible link between cognitive science and deep learning for language testing to bring about a cognitive innovation in terms of a feasible testing method as well as testing content via computers, and to further improve the estimation of ability, the discrimination, and the reliability of the JW test item.

3 Research Hypotheses

The authors argue for two important points as follows:

A test score based on an integrated test like JW reflects the test taker's ability better than that yielded from an MC test consisting of discrete test items and not vice versa. Furthermore, as integrated tests like JW require more than one skill, i.e., on collective thought processes, no guessing factor could be involved. Given the evidence mentioned above, it is possible to predict a kind of relationship between the two variables. The first null hypothesis goes as below:

H10: there exists a highly positive co-relationship between scores obtained from the integrated test and scores obtained from discrete tests;

H1a: there exists no such co-relationship between the two kinds of scores.

And the second one goes thus:

H20: there exists some guessing factor in the scores from the JW test;

H2a: there exists no guessing factor involved in the JW test.

These two hypotheses are to be justified by the test scores obtained from the experiments conducted by the authors.

4 Research Design and Implementation

Prior to further discussion, let's have a definition of what JW is referred to and how it works.

(1)JW is referred to as a test item that presents a group of words in random order to test takers who are required to make a grammatically correct sentence out of the group. To interpret in the sense of cognitive science, i.e., how to do things correctly and independently, JW is a subjective and integrated test in nature (Zhang, 2016).

(2)JW and its cognitive basis

JW is based on declarative knowledge and procedural knowledge. The concepts of declarative knowledge and procedural knowledge were initiated by Anderson (1983). JW was created in an attempt to integrate cognitive science and deep learning with testing practice in order to investigate the cognitive process revealed by test-takers during testing in order to evaluate real ability, promote an integrated testing approach, and further develop and refine such a formalization. Such a test item design, interpreted in the sense of cognitive science, actually refers to how declarative knowledge is observed to be proceduralized.

(3) In cognitive science, there are two important concepts: declarative knowledge and procedural knowledge. Declarative knowledge is referred to as unchanging, factual information stored in memory and known to be static, while procedural knowledge refers to the collective thought processes defined as how to do things known to be dynamic. Anderson (1983). In English testing, known grammatical concepts such as attributive clauses or subjunctive mood are examples of declarative knowledge. The learned set of complex grammatical rules, such as making a good sentence out of a random set of words, demonstrates an example of procedural knowledge. In the present study, what is meant by "procedural knowledge" can be best reflected in the process of coping with JW test items by test takers. As JW requires test takers to demonstrate how to make a grammatically correct sentence out of a group of words randomly presented to them, i.e., how to do things correctly and independently. In this sense, the relevant procedural is tested. Because of cognition, good procedural knowledge entails good declarative knowledge, hence proceduralized.

To elaborate, Anderson (1976, 1983, 1993) claims that recent advances in language testing research have clearly demonstrated that language proficiency is based on a variety of factors rather than the total arithmetic addition of listening, reading, and writing scores. To demonstrate the concept, consider a swimming test. Everyone understands that the passing score of a swimming test should not be simply the sum of the subskills: taking a breath and moving the arms and legs.

The goal here is to synthesize a process of cooperating, coordinating, and moving forward in the water in accordance with specific posture requirements and time constraints. So does the dancing test and the like. From the point of view of cognitive science, the total score in terms of arithmetic addition refers to declarative knowledge, which is knowledge about something or concepts and enables a test taker to describe a rule and perhaps apply it in a drill or a gap-fill, while procedural knowledge enables the test taker to apply that rule in real language use. From the point of view of language testing, declarative knowledge does not automatically cross over into communicative use of language. To be more precise, test takers may be able to memorize a grammar concept or to describe a grammatical rule and manipulate it through some controlled exercises such as cloze tests or gap-filling, and the listening and reading parts with MC questions. These are all discrete tests that only test the declarative knowledge of a language, and the simple arithmetic addition of the scores from the declarative knowledge cannot reflect the real underlying language proficiency of a test taker. Furthermore, declarative knowledge does not always imply procedural knowledge.

As a result, our students almost never apply the rule in real-life communication, whether spoken or written. In this sense, the authors' JW is the right case to make declarative knowledge proceduralized, to investigate test takers' understanding of declarative knowledge through JW test performance to determine how well declarative knowledge is proceduralized, and to realize the integrated testing approach. As a result, having the procedural knowledge tested successfully is critical.

(4) JW and its goal to achieve

Thus, the automatic production and output fluency demonstrated in test takers' real testing performance are specific indicators of successful learning without the use of random guessing. This necessitates a shift from declarative to procedural knowledge, leading to autonomy. In this sense, JW tests both parties' knowledge.

(5)Deep learning basis

Francois V. Tochon (2016) proposed two deep learning points that are important for JW design. The first is about knowledge. Knowledge is not tangible or demonstrable; the other is to emphasize process over product. This is a critical guideline for JW because only the process can be reflected, to a greater extent, in the knowledge embodied in test takers. The authors considered taking a more in-depth approach. It is appropriate for big data times, with the advent of online learning and testing. The in-depth method, also known as the in-depth education method, is a new interpretation of the process of foreign language education; therefore, testing should be included as well. Another thing worth mentioning is that the depth method not only advocates the teaching and learning of language as a whole holistic action but also reflects the language view of "language is a system" and "learning requires systematic learning." The deep approach respects learners' diverse views on learning strategies and the flexibility, variability, and modernity of teaching methods. It is an interpretation of the essence of foreign language learning with new ideas, complete content, multiple methods, demonstrations, and insights. Alternative modes of foreign language learning All this is quite in accordance with the design of JW. The research of the JW test item can be taken as one of the first contributions made to deep learning in the language testing field.

5 Research Methods

The authors concentrate on the methods used in the current study. A questionnaire was first designed to collect data about students' attitudes toward the JW test in preparation for JW test practice. Exploring factor analysis (EFA) was used to process the data in order to validate the questionnaire. A pilot study was carried out using a simple computer testing program. For the computer experiments, a small sample size was required. In this case, a mixed method of both qualitative and quantitative methods was used to analyze the research data. Both reliability and validity were considered. Rasch-GZ, the most recent Rasch-based computer software, was used to calibrate the test items. The pertinent findings, including Cronbach's alpha, will be discussed in detail in Part 6.

Research Questionnaire

For the JW program, the author specified three factors: computer operation, test difficulty, and score acceptance to be verified by the data to be collected from the questionnaires.

Figure 5-1 The EFA model for the JW test item type proposed by the authors

As shown in Figure 5-1, the JW test item consists of three factors, each containing three dimensions as illustrated above. The questionnaire is designed to collect information about test takers' attitudes towards the JW test item. We adopted herein a 5-Likert scale of "1 Strongly Disagree", "2 Disagree", "3 Neutral", "4 Agree" and "5 Strongly Agree" to check test-takers' attitudes towards the three aspects of JW as follows:

Items A1-A3 are thus designed to check the appropriateness regarding any computer-related aspects, such as whether their performance in JW is affected by their computer skills or not; whether the interface of JW is clear and explicit or not;

Items B4-B6 are thus designed to check the JW difficulties reflected in the number of words, etc.

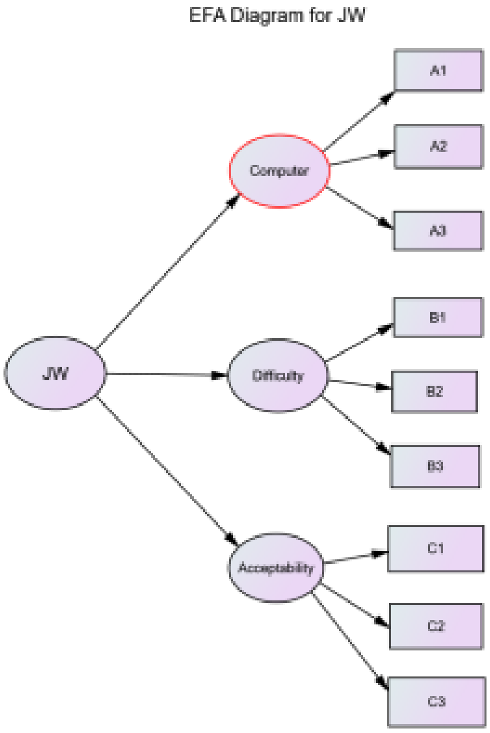

Figure 5-1 shows the hypothesis

Items C7-9 are thus designed to check the possible acceptability of the score obtained by the test takers. This would further help justify the feasibility of the JW test item type proposed by the author. For details, readers may refer to the questionnaire in Appendix I.

Research Experiments

Three experiments were conducted.

JW Experiment

This was a JW test consisting of 6 groups of words, first conducted in 2020. The authors supervised the whole test, and the scores were saved in the designated file folder of the computer for further analysis.

(A) Purpose

The purpose is two-fold: to assess the general English competence of our non-English majors and to observe whether the JW test item type could produce any guessing effect. The six JW test items with directions may be referred to in Appendix I.

(B) Subjects

Subjects are adult Chinese vocational students of non-English major (N=10) of a vocational college in Guangdong Province, China.

(C) The way to administer

The experiment was administered in a computer room on campus.

It takes four steps:

Step 1: The directions (in Chinese) for the JW tasks were read to the participants.

Step 2: The participants were encouraged to try more than once when they failed for the first time. Some questions and answers were taken till the authors made sure our subjects did understand everything about the JW test.

Step 3. The test began, with strict proctoring.

Step 4. The test takers reported that by raising their hand, they completed the task or gave up.

Step 5. Immediately after the test, all the participants were asked to complete the JW questionnaire (See Appendix II).

(D) The JW test results

The results obtained from the JW test are presented in Table 5-1 herein below.

	1	2	3	4	5	6
JW001	W	W	W	W	W	W
JW002	W	W	W	W	W	W
JW003	W	W	W	W	W	W
JW004	W	W	W	W	W	W
JW005	W	W	W	W	W	W
JW006	W	W	W	W	W	W
JW007	W	W	W	W	W	W
JW008	W	W	W	W	W	W
JW009	W	W	W	W	W	W
JW010	W	W	W	W	W	W

Table 5-1. The 10-subject-6 JW matrix

Table 5-1 presents the raw data of the 10-subject-to 6-JW test item matrix wherein all response types are W, showing all the subjects failed to give a grammatically correct sentence for each group of words randomly displayed to them.

MC Experiment

MC test consisted of 6 MC questions ad hoc administered to the same subjects who took the six JW test items. The six MC questions were actually revised MC questions based on the six JW test items previously taken by the subjects.

(A) Purpose

The purpose is to obtain the scores from both JW and MC and compare the scores obtained from the two tests of different test items.

(B) Experiment design

The experiment design here is actually a common subject equating with the JW test and at the same time, to justify the 'error' involved in True Score Theory. As test equating is big topics (Xie, 2017; Lord,1980; Kolen & Brennan,1995; Bachman, 1996, 1990; Gui, 1990, 2004, 2005, 2015; Zhang, 2019, 2016, 2004), here in our experiment, we only adopted the single-group design under common-subject equating design (Xie, 2017). The idea of single-group design in our case is shown in Figure 5-2 below.

(C) Subjects

Subjects are the same adult Chinese vocational students of non-English major (N=10) who took the six groups of JW test items in the JW Experiment conducted previously.

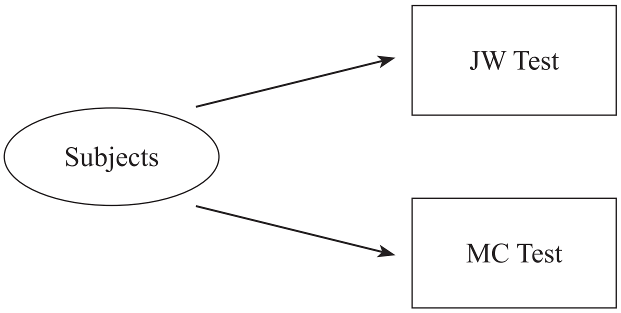

Figure 5-2 The single-group design under common-subject equating design

(D) The way to administer

The experiment was administered just in a classroom. A paper-based test with directions (as seen in Appendix II with answer keys).

Guessing Experiment

(A) Purpose

This was a JW test consisting of 6 groups of words in French ad hoc conducted in 2021[See Appendix III with answer keys.]. The purpose is to justify our hypothesis that no guessing factor was involved in the JW test item type even if the French words to be formed are very simple sentences. Therefore, the scores obtained from the JW test can best reflect the true or real language ability of our test takers.

(B) Subjects

Subjects are the same adult Chinese vocational students of non-English majors (N=10) and none of them had any knowledge of French.

(C) The way to administer

The way to administer this guessing experiment was the same as the one to conduct the JW experiment (as seen in 5.2.1).

Subjects and Data Collection

What characterizes the subjects used in the study can be described in three aspects as follows:

Homogenous. All the subjects are of non-English major. The average age is 17. And 5 are male and 5, female;

Randomly selected. All the subjects were randomly selected out of a total of 105;

Highly motivated. They are willing to participate in such a research activity because their performance will be put into consideration for final academic evaluation. Therefore, the data are fully reliable. All the data were collected in time. And no missing case.

Rasch-GZ

Rasch-GZ, the first Chinese version of the Rasch-based item analysis and test equating system, was used to process all the data. And the results will be discussed in Part 6.

6 Results and Discussion

This section provides a detailed discussion of four aspects: the questionnaire results; JW and MC test results; guessing experiment results; and the research methods used for the analysis of the study.

Questionnaire Result and Discussion

To ensure the reliability and feasibility of the research, the author designed a questionnaire with a 5-Likert scale adopted and conducted EFA on the three factors possibly inherent in JW as shown in Figure 5-1. The data thus collected via the questionnaire is shown in Table 6-1 below.

KMO and Bartlett’s Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.867
Barteltt’s Test of Sphericity	Approx. Chi Square	198.499
	df	36
	sig	.000

Table 6-1. KMO and Bartlett’s Test

As we know, the sphericity test is used to test whether something is suitable for factor analysis. KMO checks whether the partial correlation between variables is small; Bartlett's spherical test is to test whether the correlation matrix is a unit matrix. The closer the KMO value tends to be to 1, the more suitable for factor analysis. It is universally accepted that if the KMO value is higher than 0.8, it is very suitable for factor analysis. As indicated in the above table, our KMO value is 0.867, which means that the questionnaire designed by the authors is highly suitable for factor analysis. One thing worth mentioning is that before performing EFA, it is not necessary to know how many factors to use and the relationship between each factor and the observed variables. When conducting EFA, since there is no prior theory, the factor structure of the data can only be inferred by perception through factor loading. In research, it is difficult to obtain scientific results if one is only starting from data, which may even contradict existing theories or experience. Therefore, EFA is more suitable for the tentative analysis of data without theoretical support.

One more thing about the DF, i.e., degree of freedom, is that it is 36, showing the total sampled subjects is 40. Around 90 questionnaires were distributed. These 40 out of 90 were fully answered questionnaires and therefore valid. Another important point is the scree plot obtained from the questionnaire. Figure 6-1 shows the idea.

In EFA analysis, the scree plot shows the

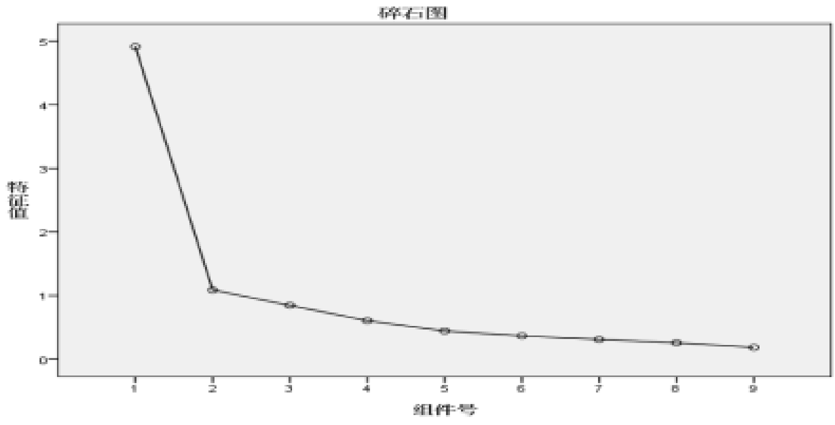

Figure 6-1 Scree plot for the factors possibly inherent in the questionnaire

possible number of factors that might be involved in the questionnaire design. As shown in the figure above, starting from the third factor downward on the abscissa, the curve tends to be decreasingly smooth. Therefore, we retain the first three factors. Or, it can be interpreted that the first three factors we conceived are well accepted. In the actual research process, it is often necessary to observe the reflected phenomena from multiple angles. Therefore, researchers often design multiple observed variables and collect a large amount of data from multiple variables to analyze and find the regularity. In the authors' opinion, although a large multivariate sample will provide us with rich information for our research, it will also increase the difficulty of data collection and processing. More importantly, there exists a certain correlation between or among many variables, which leads to the overlapping of information, thus increasing the complexity of problem analysis.

To our understanding, factor analysis is to merge and synthesize much relevant and overlapping information in the research work and turn the original multiple variables and indicators into a few comprehensive variables and comprehensive indicators to facilitate analysis and judgment. Therefore, in this sense, we prefer to use fewer factors to analyze the various types of information that exist in each variable. Or, we use a few factors to describe the relationship between many indicators and use a few factors to respond to the original data to obtain most of the information (Sapnas & Zeller, 2004).

JW and MC Test Results and Discussion

Relevant to the previous discussion (6.1), all the answers obtained from the JW test are wrong. And the post-test interview with each test taker revealed that the test was very difficult. However, when the same JW test was revised and administered in the form of an MC question type to the same test-takers, many of them could get the right answers. This gives us at least three significant enlightenments, as follows.

Three Significant Enlightenments

First and foremost, the JW test scores reflect the 'true scores,' which equal the observed scores of the test takers, and there is no 'error' involved. The scores obtained from the MC questions type included both observed scores and 'errors,' which could refer to 'guessing factors.'

Second, in light of the deep learning principle, the JW test is process-oriented because the computer program could record each step the test takers took in dealing with the randomly displayed words, trials, and failure, whereas the MC question type is product-oriented. The reason is straightforward. They could get a score as long as they chose an option that turned out to be the correct answer (even if guessing is involved). Table 6-2 shows both the item difficulty of the MC test items and the ability of test-takers. The parameters were calibrated via Rasch-GZ (5.4). Finally, the reliability between JW scores is (r=0), showing H10: "There exists a positive co-relationship between scores obtained from JW (the integrated test in nature) and scores obtained from MC (the discrete test in nature) is rejected. And H1a: "There exists no such co-relationship between the two kinds of scores" is confirmed.

The Item Difficulty and Subject Ability

The item difficulties of the MC test and the corresponding ability parameters were calibrated via Gasch-GZ. Table 6-2 shows the results.

The difficulty is expressed in the logit unit with +3 as the most difficult and -3 as the least difficult. Therefore, we can see MC questions Nos. 1 and 2 are easy items because the language point being tested in No. 1 is the attributive clause, and in the second one it is just a choice of word. No. 3 is a very easy item because the idea is very simple. No.6 is mediate. Some test-takers did not know the meaning of the word "cultivated." The sentence structure is not complicated. Nos. 4 and 5 are difficult to test because most test-takers lack background knowledge of the

ITEM ANALYSIS

RASCH MODEL, 1 PARAMETER LOGISTIC NORMAL METRIC (D=1.7)

ITEM DIFFICULTIES FROM 40 PERSONS ALL POSSIBLE SCORES ON THE TEST

==

Item No.	Difficulty	Stand Error	Corr. No.	Score	Stand Error
1	-0.893	0.376	1	-1.899	1.147
2	-0.893	0.376	2	-0.842	0.946
3	-1.038	0.384	3	-0.002	0.910
4	1.083	0.391	4	0.807	0.946
5	1.083	0.391	5	1.805	1.123
6	0.658	0.367			

Table 6-2 The item difficulty and subject ability based on the Rasch model

words. Even though some got the right answer, they admitted in the post-test interview that they got it by random guessing. What's more, when asked to put the sentence into Chinese, they were unable to.

We can draw two important conclusions from the preceding discussion:

To begin, the MC question type only tests declarative knowledge in the cognitive science sense, and the knowledge has NOT been proceduralized. That is, the test takers were unable to communicate in a real-life situation. Following that, the observed scores from the MC question type are not true scores but rather scores with errors. This is supported further by the ability parameters, which show the standard error for each level.

Guessing Factor

To further justify our hypothesis that JW is guessing-free, we conducted another JW test in French. Though both the idea and the sentence structure are simple, none of the test takers got a grammatically correct arrangement of the words. All the participants are Chinese local residents and know nothing about the French language. Our discussion will be mainly concentrated on the following three aspects:

(1) The issue of guessing in language testing

In the research of language testing, the issues of guessing have been debated throughout the short history ever since MC was widely used in tests on a large scale. Prior to extending our discussion, it is necessary for us to define "guessing" in the sense of language testing. According to Thorndike (1904), "guessing" is referred to as an array of behaviors that occur when a test taker responds to an alternative choice question to which he does not "know" the answer. Here, Thorndike was referring to the guessing behavior for MC question types. Item analysis based on CTT holds that any correct answer to an MC question (with four options) contains 25% of the guessing factor. The three-parameter model of IRT argues that the parameter C is the lower asymptote of the item characteristic curve and represents the probability of the test takers with a low ability to correctly answer an item (Hambleton, 1985). However, the authors prefer Thorndike (1904) that "the happiest solution to the guessing problem lies NOT in correcting for guessing but in preventing it." This is the very right point that the JW test item could achieve. The JW test item designed in this way has successfully prevented test takers from obtaining any correct answer by simply random guessing.

(2) Guessing and inferencing

When discussing guessing in relation to inferencing, we should first distinguish between inferencing in general and inferencing in the context of testing. In general, there are two types of inferencing: inductive (from specific facts to a general conclusion) and deductive (from specific facts to a general conclusion) (from a general conclusion to specific facts). In both cases, the relevant knowledge must be complete from the standpoint of cognition. This can be seen in any geometry or physics problem (See Anderson, l983, Chapter 9) as well as in any detective story.

In contrast, inference is highlighted in this context of testing.

Instead, the relevant knowledge involved is usually insufficient. Inferencing, from the standpoint of cognition, refers to the process of determining whether two (or more) linguistic units are related in semantic memory, syntactic structure, or both. If they are, what might the nature of their relationship be? In most cases, incomplete knowledge supports the conclusion that the two units are somehow related. In such a case, the attempt is very similar to guessing. In this sense, the author contends that guessing and inferencing are closely related, particularly when dealing with test items such as JW.

(3) Random guessing invalid for JW

Related to the previous discussion, either guessing or inferencing must be based on some relevant yet incomplete knowledge. In the guessing experiment using French words, our interpretation, in the light of cognitive science, goes thus: it is unlikely for people who totally lack certain knowledge to do a relevant thing correctly out of random guessing. In our case, it would be unlikely for test takers who totally lack knowledge of the French language to put the randomly arranged French words into a grammatically correct sentence out of pure guessing, even if the ideas and the sentence structure to be formed were simple. Guessing does not work! Based on this, we can reject our null hypothesis H20 (2.3) and take the alternative one that there is no guessing factor involved in the JW test.

Features of JW Test item

Up to this point, we can summarize the characteristics inherent in the JW in three aspects as follows:

(1) No guessing involved

This is supported by the results of the experiments we designed. And, while the JW test results were observed, they can also be considered true scores. According to cognitive science, if some knowledge is not well organized in the human brain, it can be activated by some external stimulus such as hints; if no relevant knowledge is present, no matter what kind of stimulus is used, nothing can be activated.

As a result, most participants gave up trying again because they lacked the language points being tested and the ideas for new words to be formed.

(2)Process-oriented

We achieved process-oriented testing to some extent by designing the JW test because the JW test computer system could record each step the test takers took during their test performance. The test administrator could determine whether the test takers understood or not based on the recorded process, i.e., whether they possessed or lacked the relevant language skills being tested.

(3) Integrated approach

Based on the above points, the author is very confident in stating that through JW test practice, we have in some way taken one big yet significant step towards the realization of the integrated testing approach.

7.Conclusions and Significances

There are at least three important conclusions drawn from this research.

JW, Generally Accepted Test item

Based on the above points, the author is very confident in stating that through JW test practice, we have in some way taken one big yet significant step towards the realization of the integrated testing approach.

JW, a Challenging Test item

JW, in the opinion of testing professionals, provides a better solution to the problem of guessing inherent in other test items. This is very much in the spirit of Thorndike's (1904) idea about guessing, i.e., preventing guessing in language testing, which is the best solution to the guessing problem. This can be regarded as a significant contribution to the elimination of potential "errors" discussed in the true score theory. Despite this, the most difficult aspects of language testing are the following: writing, moderating, and pre-testing of such test items. All of this necessitates more professional but stringent training. In this regard, JW is the most difficult test item.

Five Advantages

There are at least five obvious advantages over traditional paper-based tests. They go as follows:

Process-oriented

The first advantage is that JW is process-oriented; thus, in the future, it will not be a question of whether to use JW, but of how to use it well. In general, computerized JW increases the flexibility of test management significantly. Individually paced adaptive JW tests would eliminate the need for test takers to wait for others to complete them before moving on to the next group. In general, these benefits would help to alleviate the practical concerns of language teachers, students, and testing professionals.

No Guessing Involved

Another advantage claimed by JW is that there is no guessing involved because JW records and traces the test performance revealed by the test takers. This provides uniformly precise scores for test-takers. By contrast, standard fixed-MC paper-based tests always generate the score with guessing factors involved for students of average ability.

Test Security Enhanced

One practical advantage of JW is that test security is greatly improved. The reason is straightforward. It is unlikely that two test-takers will be shown the same set of words at random in the same order. It is unlikely that two people will have the same set of words in the same order. Such tests can be widely used, on a large scale, and with high security when face recognition technology (FRT) is used.

Cheating Put under Control

As a result of the preceding advantage, test takers sitting nearby, next to each other, using any electronic means to cheat become completely out of the question. As a result, cheating can be completely eliminated.

Test Duration Shortened

Time-saving can be considered as the fifth advantage of JW, which can usually be reduced by 50% or more and still maintain higher accuracy than the fixed version of MC. Test-takers will not waste time trying some items or groups of words that might be too difficult or too easy. In addition, the testing administrator or authority can also save time.

Based on the descriptions of the advantages addressed so far, JW can probably be considered as one big yet significant step towards the realization of the integrated testing approach. Its perspective indicates the future general tendency of language testing: computerized, adaptive, and integrated testing (CAIT).

8 Limitations and Follow-up Studies

The integrated testing approach and JW test item type formalized in the present paper is an interdisciplinary-oriented yet exploratory approach to language testing and ability estimation. While the authors are confident about the research work, flaws and infelicities are by no means avoidable. This final section is reserved to discuss the major limitations inherent in JW and to propose suggestions for follow-up studies.

The Limitations

There are at least three specific limitations associated with the JW study, which fall into three aspects: tough test item writing, smaller sample size, and possible adaptive to be included.

Tough test item writing

As previously stated, the features of JW are fantastic and would reshape the ongoing testing practice; however, when it comes to the writing of specific JW test items, we have quickly realized that producing such a JW test item for large-scale tests is more difficult than writing MC questions. It also costs more in terms of money, time, and energy to conduct a pre-test of JW test items. Additionally, stricter training for test item producers is required.

Smaller Sample Size

Although the experiments went smoothly, it should be noted that the number of subjects (N = 10) in some experiments is small, making the parameter estimates less stable. Despite this, the authors argue that, while it is frequently necessary for researchers to design multiple observation variables and collect a large amount of data from multiple variables in order to analyze and find the regularity, it also increases the difficulty of data collection and processing. Statistically, there is always a certain correlation between or among many variables, which leads to information overlapping, increasing the complexity of problem analysis. As a result, the sample size in some experiments is smaller.

Adaptiveness to be Included

The JW test item type can be delivered adaptively if necessary, and it can be used in a computer-based or Internet-based testing environment. This is the insight provided by our subjects following the experiments. Most people complained about the new words in the test, the ideas, or the number of words.

As a result, the future JW test should be adaptive in terms of both word count and idea formation, beginning with the most basic sentence structure from six words on and gradually increasing the word count, while the ideas to be formed become increasingly complex. Everyone knows that in traditional tests, which are typically paper-based, all test takers receive the same test items in the same order. Mismatches between the test taker's ability to item difficulty result in wasted testing time.

These issues would be largely solved if each test taker chose a subset of the total set of items that were appropriately difficult for his or her ability level. Adaptive means attempting to solve the problem by using a test taker's previous responses to select subsequent items of appropriate difficulty.

As a result, it is also known as "tailored testing." In other words, it should be a type of computer-administered test in which the next item or set of items chosen to be presented to test takers is determined by the accuracy of the test taker's responses to the most recent items administered. This is a common practice in computerized adaptive testing (CAT), which selects questions sequentially to maximize exam precision based on what is known about the test taker from previous questions (Weiss & Kingsbury, 1984). The difficulty of the exam appears to suit the test taker's ability level. Consider the JW test. If a test taker excels at a moderately difficult task, such as correctly combining a group of six randomly displayed words into a grammatically correct sentence, the next group of words will be more difficult, or a group of seven words will be presented, or the sentence structure will be more complicated. If they do not perform well, they will be faced with a group of fewer words or a simpler sentence structure.

In this sense, JW would be a computerized cognitive measuring instrument, delivering an interactive task via computer. Its goal is to select a group of words and randomly display them for each test taker so that they can simultaneously and most effectively assess their abilities based on their ability characteristics. This necessitates the knowledge being proceduralized.

Follow-up Studies

In the authors' opinion, there are many aspects to be improved, among which two aspects are obvious: face recognition technology (FRT) and multimedia technology to be included in the follow-up studies.

Face Recognition Technology(FRT) to be used

The year 2014 was a turning point in FRT, which moved from theory to application. In 2018, it is an important time node for the full application of face recognition technology, and the era of "brushing face" has officially arrived. At present, from the perspective of FRT application, it is mainly concentrated in three areas: attendance and access control, security, and finance. At present, "face-shaping" has gradually become a new trend. With the continuous expansion of the commercial application of FRT, "face-brushing" services are becoming more common. Face recognition has the advantage of being natural and imperceptible. This identification method is difficult for test-takers to resist because it is undetected. The authors believe that FRT will eventually become the dominant recognition technology.

When compared to traditional biometric methods such as fingerprint recognition and iris recognition, the benefits are primarily focused on four points: non-contact, non-intrusive, complete hardware foundation, fast and convenient collection, and good scalability. In a complex environment, it is expected that once the issue of FR accuracy is resolved, FR will quickly replace fingerprint recognition and become the mainstream recognition technology for use in large-scale examinations. Therefore, the authors highly suggest that applying FRT to the language testing field can quickly popularize security at the level of both computer-based and Internet-based testing applications, which will become the main trend of computer-based testing development in the future.

Application of Multimedia to be Used

It is known to all that multimedia technology has been widely used in various fields of education and teaching due to its combination of graphics, text, audio, and video. This makes it more suitable for the practice of an integrated language testing approach. Therefore, the authors here are mainly focusing on the application in the sense of incorporating listening and speaking skills into JW. This would be considered compulsory and feasible. Therefore, the authors suggest that further study be focused on the incorporation of both listening and speaking parts and technologies of multimedia into JW. Once appropriately corporated, JW would be the qualified test item type to be made for tests on a large scale. An inevitable trend in the development of an integrated testing approach As for the specific methods in terms of application form, at least the following three points should be addressed in some detail.

(1) Demonstrating examples with voice directions. Before the test begins, multimedia technology can be used to demonstrate and explain how to do a JW test item, such as the specific requirements and grammatically correct arrangement of JW. The test administrator may use a combination of multimedia and projectors to display the test content, pictures, moving images, or self-made animation, which is conducive to the good understanding of the test takers to achieve the best results.

(2) Interactive testing. Here, "interactive" actually refers to "communicative". As multimedia and network technology provide comprehensive simulation with pictures, text, and sound, it is conducive to the creation and maintenance of the situation. With a friendly interface, the image is intuitive, and the subject knowledge is organized and managed in the form of hypertext and hyperlinks. In such a testing environment, test takers can immediately get feedback, accept their test results, and adjust their learning methods. Such a kind of interactive test is conducive to stimulating the test takers' interest in

learning and in playing its role as a cognitive subject. Again, from the point of view of cognitive science, if the test is interactive, it implies that the knowledge has been proceduralized.

(3) Modern remote testing. Modern remote testing refers to a new type of testing mode in which test administrators and test-takers are relatively separated in time and space, using modern information technologies such as network technology and multimedia video technology to transmit the test to off-campus in real-time or non-real-time.

Finally, the authors believe that with the development and popularization of computer multimedia technology and the Internet, the future trend of language testing will make more use of multimedia technology to fully develop integrated testing.

Aligning JW to Chinese Standards of English Language Ability (CSE)

Since CSE was officially released, it was supposed to act as one of the compulsory measures to promote language testing and assessment in China. Therefore, aligning the present research with CSE in the future will be good for the development of language testing in China. Therefore, the authors propose that follow-up studies focus on the CSE to JW, too. The conference proceedings of ICLTA may serve as good references.

References

Anderson, J. R. (1976). Language, memory, and thought. Psychology Press.

Anderson, J. R. (1983). The architecture of cognition. Harvard University Press.

Anderson, J. R. (1993). Rules of the mind. Lawrence Erlbaum Associates Inc.

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

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

Gui, S. C. (1990). Item Banking 2. *Modern Foreign Languages,* 4, 66–72.

Gui, S. C. (2004). Rethinking English Education in China - Chapter Practice. *Modern Foreign Languages,* 5, 687–704.

Gui, S. C. (2005). On Current Foreign Language Teaching. *Chinese Foreign Language,* 1, 5–8.

Gui, S. C. (2015). International language testing research trend. Overseas Annual Report on the Development of Humanities and Social Science. Wuhan University Press.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principle and application. Kluwer Nijhoff Publishing.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. Springer-Vertag.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche.

Sapnas, K. G., & Zeller, R. A. (2004). Minimizing sample size when using exploratory factor analysis for measurement. Journal of Nursing Measurement, 10(2), 95–96. https://www.ncbi.nlm.nih.gov/pubmed/12619534

Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. Teachers College, Columbia University.

Tochon, F. V. (2016). Help them learn a language deeply Deep approach to world languages and cultures (Chinese Edition). Deep University Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement,* 21(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wu, J. Y., & Zhang, Q. (2018). Computerized Adaptive Integrated Writing System (CAIWT) and Its Validity. Pacific-Rim Objective Measurement Symposium (PROMS), Fudan University, Shanghai, China.

Xie, X. Q. (2017). Test Equating Handout.

Zhang, Q. (2002). Computerized cognitive testing: Theory, method and practice. In Gwyn Boodoo (Chair), report presented at Educational Testing Service (ETS). Princeton University, USA.

Zhang, Q. (2004). Item analysis and test equating in language testing: Research and application. Higher Education Press.

Zhang, Q. (2007). Jumbled word test, a promising alternative for MC format: An assumption based on cognitive science. In Bachman, L. F.(Chair). SCALAR, UCLA, USA.

Zhang, Q. (Ed.) (2015). Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings. Springer Singapore. https://doi.org/10.1007/978-981-10-1687-5

Zhang, Q. (Ed.) (2016). Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings. Springer Singapore. https://doi.org/10.1007/978-981-10-8138-5

Zhang, Q. (2019). Rasch model: Research and practice in China. In Myint Swe Khine. (Ed.). International trends in educational assessment: Emerging issues and practices. Retrieved from http://catalog.loc.gov