

Towards Cognitive Response Theory in Diagnostic Language Assessment

Quan Zhang

College of Foreign Studies,
Southern Medical University, P. R. China

This paper proposes that Cognitive Response Theory (CRT) be implemented in the form of computerized cognitive testing (CCT). It begins by contrasting key characteristics between traditional computer adaptive testing (CAT) and CCT. CCT is operationalized through a jumbled word (JW) test item, yielding two cognitive variables—response type and response time—to estimate ability level. The latent variables hypothesized to underlie test performance were tested against the data through the use of structural equation modeling (SEM). Results show promise for applying CRT to tests of English as a foreign language.

INTRODUCTION

The growing reliance on tests for making high stakes decisions and for improving educational outcomes has called attention to some serious limitations germane to theories guiding language testing practice.¹ In the author's view, scholars and experts of language testing should actively address such problems by refashioning assessments to meet current and future needs for quality information with the help of cognitive science as well as computer and multimedia technology. Take the multiple choice (MC) question format as an example. For many years, MC has been the dominant and indispensable test format in assessments ranging from teachers' informal quizzes to large-scale tests administered worldwide. Traditional computer-adaptive tests also use the MC question format to explore diagnostic assessment. Though the theoretical basis for the MC question format can be found, relevant criticisms have also been voiced.

This paper argues that with the advanced technology of computer programming and multimedia, a jumbled word (JW) item format is a promising alternative to the MC item format for language assessment. I begin by explaining some of the limiting characteristics of the MC item formats in contrast to those of the JW format. I then discuss some aspects of a pilot study that was conducted to evaluate an English grammar test in the JW format. The basis of the JW item test was based on Cognitive Response Theory (CRT) realized in the form of computerized cognitive testing (CCT). A CCT model is believed to better

The author would like to acknowledge with thanks Prof. Carol A. Chapelle and her organizing committee of the 5th annual TSSL conference, whose invaluable comments and suggestions for improvement helped refine the paper. The author's thanks also go to Eric Wu of Department of Psychology, UCLA, under whose guidance both CCT modeling and EQS running went well. The opinions expressed in this article are those of the author who remains solely responsible for any possible errors in the article. This research is supported by funding of Guangdong Provincial Project for New Century, PRC (2006-2008).

demonstrate the “Assessment Triangle”ⁱⁱ by explicitly connecting cognition, observation and interpretation. The research presented here, though preliminary, calls for feedback from the larger community of language testing. It is intended to demonstrate a tangible basis for further research towards diagnostic assessment via CCT.

THE TRADITIONAL COMPUTER ADAPTIVE TEST

Although the original idea of adaptivity can be traced back to the work of Binet (1909), Lord (1970), Birnbaum (1968), and others, only with the advent as well as availability of computers nowadays could the traditional computer adaptive testing (CAT) method become feasible for widespread operational research and implementation. “Adaptive” in this sense refers to a testing procedure that selects the next item to be presented to an examinee based on a test taker's performance on the previous one (Bunderson et al, 1989). Such a procedure is based on the idea that more information about a test taker's trait can be obtained from an item with a difficulty level fitting the test taker's ability. Therefore, an adaptive test requires a set of test items at various difficulty levels, nowadays mostly still in MC format, from an item bank. These test items are calibrated in advance so as to yield parameters that can be used by the selection procedure during test taking. CAT is widely acknowledged to have advantages over conventional paper-and-pencil tests in estimating test taker's ability; however, some characteristics of CAT limit the possibilities of measurement for language assessment. In light of cognitive theory, the following limitations are evident in current CAT practices: CAT is difficulty-bound, dichotomous-valued, MC-limited, time-neglected, and product-oriented (Zhang, 1993, 2002a).

Difficulty-bound refers to the fact that CAT relies on only one aspect of cognition, i.e., the adaptivity procedure is controlled by the item difficulty and a test taker's ability. In other words, ability is bound to item difficulty alone, each being interpreted only in the specific context of the other. The unidimensional ability-difficulty link is overly simplistic in view of the many cognitive variables that influence test performance to some extent. As a consequence, CAT cannot explicitly include substantially meaningful interpretation of what test performances should actually be inferred to mean.

CAT is dichotomous-valued in that it treats the adaptivity between item difficulty and a test taker's ability as binary-valued, i.e., yes-or-no type. Test takers who present wrong answers are all labeled as lacking certain knowledge in the tested domain. In this sense, CAT fails to distinguish test takers who have partial knowledge from those who don't have any in the process of problem-solving. This concerns very much test validity. In the view of cognitive science, human's cognitive ability is by no means dichotomous-valued. Instead, it is of more-or-less type. Over the past 40 years many researchers (Rasch, 1960; Bock, 1981; Hambleton, 1989; Hambleton & Swaminathan, 1985; Smith, 1987; Mislevy & Bock, 1984; Mislevy & Verhelst, 1987) have examined the hypothesis that dichotomous scoring does not capture the full information available in the responses concerning a person's cognitive ability. Most of these scholars have found that the degree of incorrectness of an answer can be quantified and used as an additional source of

information about the test taker's ability. To overcome this perceived deficiency of dichotomous scoring, a variety of techniques have been developed, such as response weighting, answer until correct, degree of confidence weighting, elimination scoring, and so forth (Smith, 1987).

The current CAT practices are mostly limited to a MC question format despite the fact that at least four problems with this item format have been found. First, good MC items are difficult to develop. The common practice is that each question stem and distractor, prior to its use, undergoes the process of item writers' moderation and pretesting. Second, tackling a MC question is a selective process rather than a precise one. It involves partial use of available minimal language cues selected from perceptual input on the basis of the test takers' expectation. As this partial information is processed, tentative decisions are made to be confirmed, rejected, or re-confirmed as coping with each item progresses (Snow & Lohman, 1989). Third, the attempts at the distractors made by test takers may reveal their cognitive level. In other words, the possible guessing behavior demonstrated in selecting the distractors may be taken as the indicator of test takers' cognitive ability. Ideally, such data should be utilized by test developers in post-test item analysis to gain insight into the test takers' cognitive ability. Finally, with further understanding of cognition, test users have also come to realize the importance of the guessing factors. Overall, it seems evident that MC question format should by no means be the only test form for measuring cognitive abilities, and language ability in particular.

CAT is also time-neglected as it does not record test takers' reaction time during test taking. The failure to collect these data prevents CAT from distinguishing among the abilities of test takers who obtain the same scores. This again concerns test validity because a single score on a test can be obtained in different ways by different examinees. In the view of cognitive science, solution time is an important cognitive variable particularly in measuring the procedural knowledge at command. It distinguishes experts' from novices' performances. Hence, without tracking test takers' solution time, CAT practice might in some way weaken the test validity.

When it is said that CAT is product-oriented, it means that CAT does not assess test takers' problem-solving strategies or skills. What a score from CAT reflects is just the terminal answers, which are either right or wrong, and which offer no chance for test users to observe examinees' problem-solving procedures. In the view of cognitive science, the significant interpretation of test takers' real potentiality pertaining to problem-solving is evident from a display of their cognitive process rather than the terminal product. Therefore, the meaningfulness of inferences drawn from CAT assessment using MC questions may be compromised despite the fact that today's multimedia and web technologies make such observations feasible.

COMPUTERIZED COGNITIVE TESTING

Computerized Cognitive Testing (CCT) is a theoretical approach that is intended to provide an alternative to traditional CAT. It is therefore useful to examine its potential to

be applied in language assessment (Zhang 1993, 2007). Compared with CAT, CCT is unique in the following six aspects: CCT is cue-provided, polychotomous-valued, JW-adopted, time-recorded, process-oriented, and procedural-knowledge-based. Such features of CCT will be explained in detail in this section to show its advantages over traditional CAT.

CCT is cue-provided as it is capable of giving hints relevant to the solution to a problem in case that test takers fail to provide a correct answer at the first attempt. Evidence for the importance of providing such cues or hints comes from experiments showing that most students who failed to provide a correct answer at the first trial were not ignorant of or lacking in the relevant knowledge. Given a little help, they could quickly solve the problem. Then, why is the provision of hints better than difficulty-based adaptivity? In traditional CAT, an examinee will be given a different test item at a lower difficulty level if he or she previously fails in a relatively more difficult item. However, two items at different difficulty levels are usually of different content and may thus test different aspects of language ability. So the construct in such a test is inconsistent across items or defined in a way that encompasses the content of all the items. In contrast, CCT provides hints to lower the difficulty level of a test item but still keeps its content unchanged. In other words, both the difficult and the easier items assess the same language aspect, reinforcing the diagnostic assessment.

How to provide students with immediate and direct feedback on their test performance is one of the problems in education. In China, whether the test is a placement test, an achievement test, or a proficiency test, students usually do not receive feedback about their test performance until a long time after they have taken the test. Furthermore, what they receive is usually general feedback, such as information about what is the best choice for a multiple choice question. This is not only because we lack certain research methods for monitoring students' performance during a language test but also because it is not feasible to implement such a task in any traditional paper-and-pencil tests.

According to Anderson (1974, 1976, 1983, 1985), the retrieval process in information processing requires that certain cues be provided, either by the external stimulus or by the learner. Accordingly, developers of diagnostic assessment must consider providing such cues. This is based on a cognitive hypothesis that human knowledge is stored by means of propositional networks in the brain. The retrieval of the knowledge from the brain is achieved through the spread of activation. Thus, it appears that a well-organized knowledge structure in the brain is activated more quickly than otherwise. In some cases, the retrieval process may be baffled due to the lack of certain knowledge. Therefore, according to CCT a test taker failing to provide a correct answer for the first time may have the relevant knowledge but be unable to activate it due to the inefficient organization of the knowledge in the brain. In this case, a certain external stimulus is required to trigger the retrieval of that knowledge. In this sense, the significance of providing cues is twofold: A language test with cues is capable of (1) distinguishing test takers who have the relevant knowledge stored in the brain from those who are totally lacking in such knowledge and (2) further discriminating test takers who answer the same

number of questions correctly by revealing how they get the answers right. Hence, the provision of cues is potentially an important approach to diagnose a test taker's real language ability.

“Polychotomous-valued” means that CCT treats the 'cognition' between item difficulty and an examinee's ability as a continuum in terms of the degree of achievement. CCT presumes that test takers who fail on a test item for the first time may possess partial knowledge in the tested domain. In contrast to CAT, it offers hints to the answer so as to give the test taker more chances to succeed. If the test taker gets the right answer with the help of the hints, CCT also gives a partial credit. In this way, CCT can further distinguish test takers who have given the same number of correct answers based on the extent to which they seek help. The range of ability levels thus interpreted by CCT goes from the highest to the lowest with many intermediate scores in between. Such a detailed ability continuum, which maps observed responses to the strength of knowledge, best reflects the different levels of human cognitive ability, and thus proves to be another way to explore the real language ability of test takers.

The JW task form used by CCT has three advantages. First, the JW form allows for the assessment of integrative skills concerning both vocabulary and grammar knowledge. Second, the JW form demands dynamic performance. The third advantage of the JW form is that it prevents test takers from making a blind guess about the correct answer because the available language cues for guessing are minimized. In addition, the JW form is intended to assess test takers' procedural knowledge (knowing how) as well as declarative knowledge (knowing what) in cognition. The cognitive basis of the JW task design can be verified in the following three aspects:

- The JW form focuses more on the use of language than on the knowledge of language per se. In other words, it focuses more on procedural knowledge than declarative knowledge. In the view of cognition, procedural knowledge entails declarative knowledge. Thus, the integrative skills concerning both vocabulary and grammar knowledge assessed via JW test form is in fact test takers' procedural knowledge. In this sense, a test taker's quick, correct response to a JW item without the assistance of cues reflects his/her solid possession of both declarative and procedural knowledge, while a correct answer based on cues reveals that the test taker has the knowledge but that it has not been proceduralized.
- The JW test item demonstrates that the whole is more than the sum of its parts, an important Gestalt claim. In this sense, one's language ability is by no means merely the sum of one's vocabulary and grammar knowledge put together. This can be justified by the experiment and post-experiment interviews (see discussion of the experiment in the following section) conducted by the researcher. The interviews reveal that that some subjects who knew the meanings of individual jumbled words were still unable to put them into a logical sequence. Besides, we found that not all the subjects who knew the concept of tenses or

attributive clauses could fulfill the tasks well. In other cases, test takers could quickly get the correct answer when they used specific hints given by the test. Cognitively, this is interpreted as such that the test taker's declarative knowledge not being very well proceduralized.

- The JW test item, in contrast to the MC question form, allows for no random guessing. In cognitive science, the extent of guessing indicates a person's level of cognition. In other words, the guessing behavior demonstrated in coping with JW test items is considered as the indicator of test takers' language ability. Thus, a correct answer obtained through guessing indicates the test taker's knowledge concerning the subject-matter learning is incomplete. Similarly, failure to obtain a correct answer after consulting hints is interpreted as total lack of the relevant declarative knowledge being tested.

CCT is time-recorded. Another method of evaluating cognitive processing is to measure the amount of time test takers spend in problem solving (Klahr & Robinson, 1981; Anderson & Gluck, 2001)ⁱⁱⁱ. Data collected in this way can be highly informative. Here, time-recording refers to the reaction time or retrieval time spent by test takers in coping with each set of jumbled words during test taking. According to the speed at which a problem is solved or, in other words, certain knowledge is activated, CCT could produce a time parameter indicating test takers' levels of proficiency in six categories: Native User, Near Native User, Good User, Modest User, Average User and Poor User. As noted previously, procedural knowledge is executed rapidly and with minimal demands on attentional resources; therefore, assessment must take into consideration the solution time, which is one useful index of automaticity for many problem-solving tasks. This has been proven to be another "window on the mind" (Dillon, 1985; Just & Carpenter, 1992)^{iv} to observe the strategies test takers use in coping with JW test items.

When we say that CCT is process-oriented, we should first spotlight "process" in the sense of testing. For instance, in any tests of mathematics or geometry, test takers are usually required to write down the process to obtain a correct answer because each step in problem-solving indicates his/her relevant knowledge pertaining to the subject-matter learning. However, such an important cognitive assumption has not been paid much attention to in many language tests, particularly in the tests composed of multiple-choice questions. This is largely due to two reasons. On the one hand, test developers lack certain research methods in observing or measuring test takers' cognitive process during test taking. On the other hand, it is not feasible to do so in any traditional paper-and-pencil tests.

With the advances in computer technologies and the availability of computers, it is now feasible to make CCT process-oriented. Built on an information processing model, CCT is able to trace test takers' cognitive process of problem solving by recording their reaction time and remembering their use of the help options (i.e., cues). Thus, CCT is concerned more about how test takers get to the answer than whether the answer is right or wrong. In this sense, CCT can be considered an instrument to observe and trace what

is going on inside the test taker's brain while he or she is tackling each test item. This is what "process" means in the sense of cognitive testing, and the idea of assessing testing processes is in the spirit of "Assessment Triangle" as described in *Know What Students Know* (National Research Council, 2001).

Test takers' procedural knowledge can be best assessed by letting them arrange and re-arrange jumbled words into a logical sentence. This requires them to demonstrate how they sequence ideas in a second language. In this sense, the relevant procedural knowledge of sentence formation is tested. In view of cognition, good procedural knowledge entails good declarative knowledge. CCT implemented in this research

Table 1. Summary of Contrasts between CAT and CCT

Traditional Computer Adaptive Testing (CAT)	Computerized Cognitive Testing (CCT)
Difficulty-bound only Adaptivity is realized by adjusting item difficulty based on test takers' responses. Difficulty is reduced by providing easier items, which, makes the language aspect being tested inconsistent.	Cue provided Two factors are considered for adaptivity: (1) Response type (2) Response time Difficulty is reduced by providing relevant hints and keeping the language aspect being tested unchanged.
Dichotomous-valued Uses binary logistic model, i.e. yes-or-no type rating. All the incorrect answers are treated as wrong. Ability estimation is bound to item difficulty alone.	Polychotomous-valued Uses partial credit model, i.e. more-or-less type rating. All the correct answers are specified in different response types. Ability is estimated based on a continuum of degree of achievement.
MC-limited Requires separate skills only; Demands selective performance only; Offers chances for blind guessing.	JW-adopted Requires the integrative skills concerning both vocabulary and grammar knowledge. Demands dynamic performance. Offers no opportunity for blind guessing.
Time-neglected Test takers' reaction/solution time is not taken into consideration for ability estimation.	Time-recorded Response time is taken as important cognitive variables for ability estimation.
Product-oriented Test takers' strategies or skills demonstrated during problem-solving are not traced. Only terminal answers are scored.	Process-oriented Test takers' strategies or skills demonstrated during problem-solving are recorded for further diagnosis.
Declarative-Knowledge-based Reflects only declarative knowledge.	Procedural-Knowledge-based Procedural knowledge is tested. Best reflected in arranging and re-arranging a set of jumbled words into a logical sentence; Best demonstrates how integrated knowledge of language works independently; Good procedural knowledge entails good declarative knowledge.

generates a file for each test taker, recording the whole process of problem solving. This includes the test taker's response using or not using the hints, the item difficulty, the corresponding solution time, and the frequency of attempts. From such records, the test user is able to know exactly how the test taker solves each problem and how and why he or she fails. By tracking the test taker's behavior, CCT can find out his or her problems in information processing so as to provide individualized feedback for the follow-up instruction.

A Summary of the Differences between CAT and CCT

The six contrasting aspects between traditional CAT and CCT are summarized in Table 1. It highlights the important differences in measurement between these two testing approaches. Such differences shed light on the development of tests that can accurately measure examinees' knowledge and skills as well as on the development of tests designed for specific purposes such as diagnosis.

THE PILOT RESEARCH

The remaining part of the paper will report a pilot study which investigated the use of CCT for assessing examinees' English language ability. The study revealed the potential for CCT to be applied in language assessment. The methodology for the pilot study is built upon other research that the author has been conducting since 1993.

Participants

The original sample of participants consisted of approximately 200 vocational students in majors other than English at Southern Medical University, Guangzhou, China. The post-test data editing confirmed that 120 cases were valid data records.

Table 2. 15 Possible Response Types for CCT

Type	R/W	Hint-1	R/W	Hint-2	R/W	Hint-3	R/W
A	1						
B	0	N	1				
C	0	N	0	N	1		
D	0	N	0	N	0	N	1
E	0	Y	1				
F	0	N	0	Y	1		
G	0	N	0	N	0	Y	1
H	0	Y	0	N	1		
I	0	Y	0	Y	1		
J	0	Y	0	N	0	N	1
K	0	Y	0	Y	0	N	1
L	0	Y	0	Y	0	Y	1
M	0	Y	0	N	0	Y	1
N	0	N	0	Y	0	Y	1
O	0	N	0	Y	0	N	1
W							

Test Design

Ten JW test items (see Appendix A) were designed, each including 3 relevant hints. The average number of words used for each JW item was seven. Table 2 illustrates all the possible response types for the test. In the table, 1 and 0 indicate correct and incorrect answers while Y and N indicate whether or not a test taker used a specific hint. As each JW test item has three relevant hints, there are a total of 15 response types.

Data Analysis

The data of test takers' responses were analyzed using PARSCALE4.1. (See Appendix B for the PARSCALE command file.) Technically, PARSCALE only accepts ordinal data; therefore, the interval data of test takers' response time were coded into six ordinal categories: (1) Native User, (2) Near Native User, (3) Good User, (4) Modest User, (5) Average User and (6) Poor User. In addition, test takers' response times on each item were also recorded. The ability scores and the categorical response data with response time assume the partial credit model with the standard scoring function. In sum, the assessment of test takers' ability took two cognitive variables into consideration: response type and response time.

Two examples provided here are test takers' responses on two test items in CCT. It is worth noticing that the response types labeled 'E', 'F', and 'I' (See Table 2) are typically syntactic-knowledge based, or rather procedural-knowledge based. Accordingly, these examples best illustrate how participants approached the jumbled word items with the help of hints provided.

Example One

Subjects were presented a set of jumbled words as follows:

terrible, Tom, described, the, service, sounds

They were unable to identify the hidden syntactical structure of the target sentence at the first trial. So the first attempts they made were sentences such as:

Terrible Tom described the sounds service



Tom described" the terrible sounds service



The terrible sounds service described Tom

Receiving such responses, a CAT system would presume the test takers are unable to cope with such an item and thus provide them with an easier one. As a result, the test item is changed and meanwhile, the language aspect being tested will probably be changed as well. In contrast, a CCT system would react differently in such a situation. It would provide a relevant hint germane to the item instead of providing a new test item. In this

example, the first hint the system provided is “*The sentence Begins with 'The service'*”; the second hint is “*This sentence contains an attribute clause*”. These hints make the item easier for test takers to tackle. With the hints provided, the subjects appeared to become aware of the existence of an imbedded attribute clause and made a complex sentence as follows:

The service Tom described sounds terrible.

Example Two

In another test item, test takers were presented the following jumbled words:

more, hormones, than, influence, adults, do

They were unable to identify the syntactical structure of the key without referring to hints either and thus made sentences mostly in random word orders. Some of their first attempts were:

More hormones than adults do influence

or

More adults do influence than hormones

and

Hormones do influence more than adults,

Once given the first hint, “*The sentence begins with 'Hormones,'*” test takers understood the sentence structure and made a meaningful sentence as follows:

Hormones do more than influence adults.

Here we should say it is not that the subjects know the sentence structure very well at the first time but that the subjects are believed to be better able to infer, with the hint(s) given, that the key of the JW test item must be a sentence in a complex structure. The examples, in some ways, justify the cognitive hypothesis: human knowledge is stored by means of propositional networks in the brain. The retrieval of the knowledge from the brain is achieved through the spread of activation. Thus, well-organized knowledge structure regarding attribute clause is activated more quickly with the help of hints and otherwise, more slowly. In case the retrieval is baffled, the test takers appear to lack the relevant grammar knowledge being tested. Hence, the provision of cues turns out to be an important way of diagnosing test takers’ real language ability. The first attempts made by test takers in the examples given above also demonstrate that JW items do not allow for test takers’ blind guessing.

RESULTS

The results of data analysis which examined item responses in addition to the latent factors underlying test performance are presented in this section.

Response Analysis

The present study obtained two ability curves, one indicating the curve based on response types and the other on both the response type and response time. As shown in Figure 1, response type and time values are highly correlated. The higher ability levels indicate response types A, B and E as these test takers spent less time in managing to get a correct arrangement of the jumbled words, while the low ability levels are those of type K, L and N. A further analysis of the responses showed that response time turned out to be a significant variable which was capable of distinguishing ability levels of the same response type.

Latent Factor Analysis

To verify the theoretical formalization described above, the present researcher applied Structural Equation Modeling (SEM) using EQS6.1 to investigate the concept of matching CCT traits with the expected model. Since the first application of SEM approach to language testing in 1981 (Bachman & Palmer, 1981), SEM has been used in a wide range of studies (e.g. Kunnan, 1995; Bae & Bachman, 1998). However, no studies in China have ever used a latent factor approach to address such fundamental issues about language abilities. Based on the above discussion, three latent factors have been specified which are believed in one way or another to influence the test taker's language ability. These three factors are (1) the test taker's mode of performance, (2) the test taker's condition and (3) the test item difficulty. According to SEM, these three latent factors are formed as three measurement models each containing four or five measured variables. Figure 2 shows the measurement model for test taker's mode of performance. The measured variables, V1 to V4, in the squared forms, indicate ST (Solution Time), HA (Hint-adopted), GG (Guessing) and CT (Cheating). These loaded on the latent variable Test Taker's Mode, i.e., the test taker's mode, showing how the test taker is coping with JW test items. Each single arrowed line expresses one variable affecting the other directly while each arrowed line pointing from E1, E2, and so on to the squared box indicates the un-interpretable parts of latent variables and can instead be understood as a kind of possible errors (Bentler, & Wu, 2002; Bentler, 2006; Byrne, 1994; Jöreskog, 1970,1977; Bachman, 1998; Kunnan 1998,1999; Purpura,1998; Rob, 2005).

Figure 3 shows the measurement model for the test taker's condition containing the measured variables, V5 to V9, indicating TR (Test Readiness), TF (Test Familiarity), ID (Individual Difference), TRF (Test Room Familiarity), CBF (Computer-based Familiarity), TIF (Test Item Familiarity). These loaded on the latent variable Test Taker's Condition, or F2.

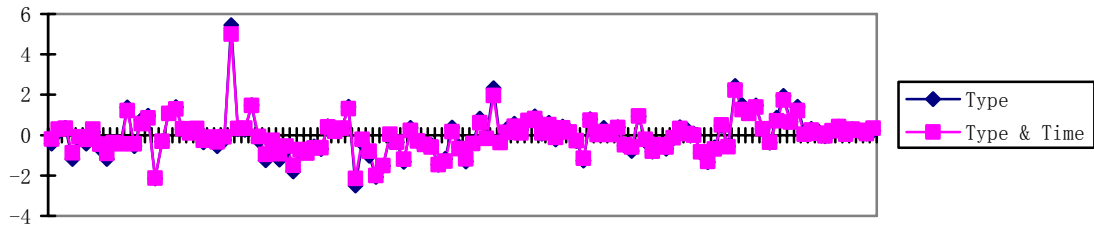


Figure 1. Ability curve based on the response type and response time (N=120).

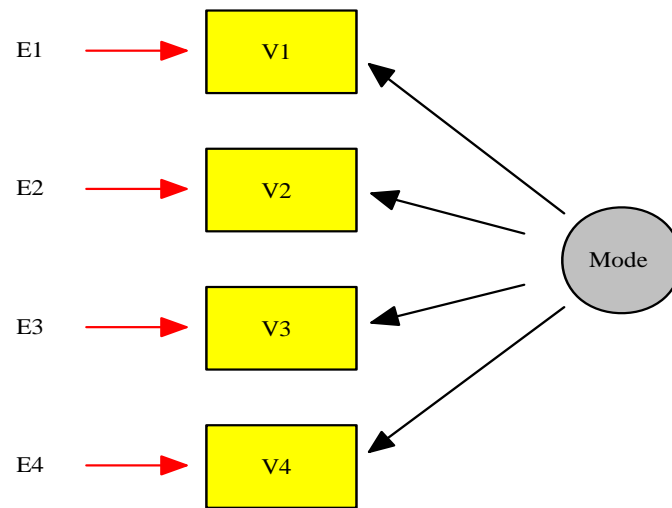


Figure 2. Measurement Model for Test Taker's Mode

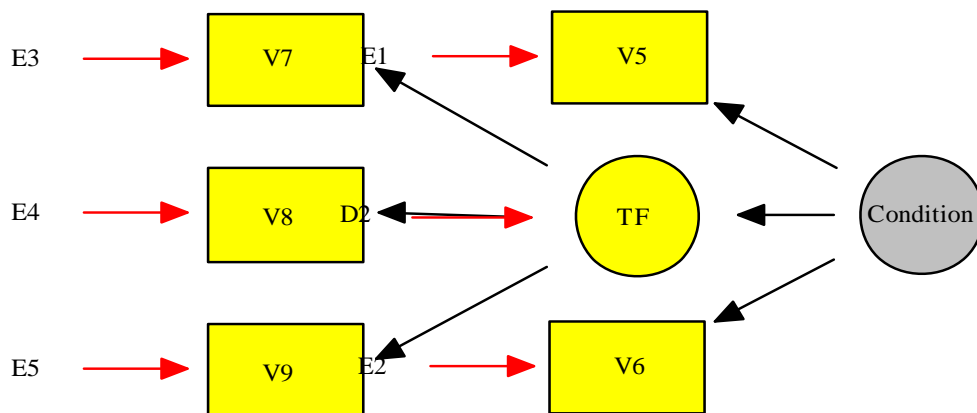


Figure 3. Measurement Model for Test Taker's Condition

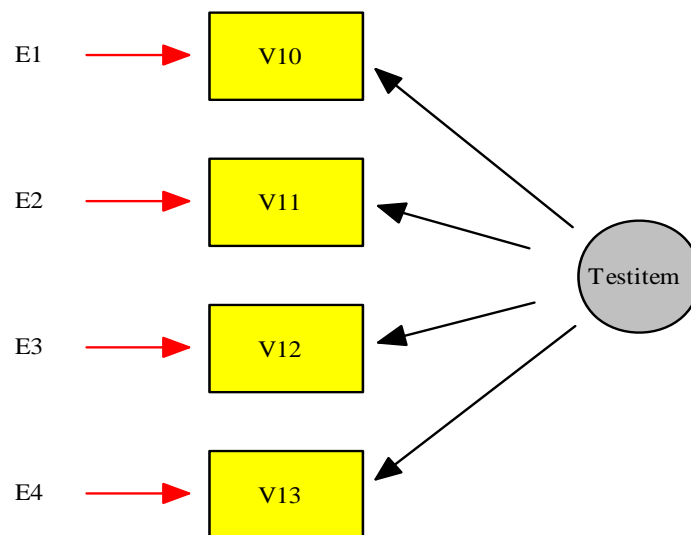


Figure 4. Measurement Model for JW Test Item Difficulty

Figure 4 presents the measured variables, V10 to V13, indicating WN (Word Number), SS (Sentence Structure), VOC (Vocabulary) and BG (Background Knowledge), which loaded on the latent variable JW Item Difficulty, or F3.

According to SEM principles, these measurement models are formulated in the confirmatory mode and are based on prior experimental results conducted during the researcher's doctoral studies. Parts of the data are the raw data collected from the test takers of PRETCO^v administered from 2002-2005 in Guangdong Province, PR China. Figure 5 presents the second-order model^{vi} using Test Taker's Mode, Test Taker's Condition and Test Item Difficulty linking both the independent and dependent variables and their associated measured variables and errors. According to SEM, such a model is based on the hypothesis that these three latent variables are structured as illustrated to represent the construct of language ability measured by this test.

As a second-order model, parameter estimates for measured variables and correlations among the latent variables are all calculated with EQS6.1. Figure 5 shows the output containing the goodness-of-fit statistics.^{vii} The comparative fit index (CFI)^{viii} = .931 indicates that the model is reasonably acceptable. However, as it is .931 rather than .95 or above, we may presume that there exist some other factor(s) that influence the measured language ability. Figure 6 shows some minor inappropriateness regarding the z-scores of variables like Test Room, Computer and Individual Difference as shown in Figure 6.

Figure 7 shows measurement equations with standard errors and test statistics after 198 cycles. As shown in the figure, both guessing and solution time are significant. But each of these z-scores for TRF (Test Room Familiarity), CBF (Computer-based Familiarity) and ID (Individual Difference) turns out to be negative, much smaller than the abstract

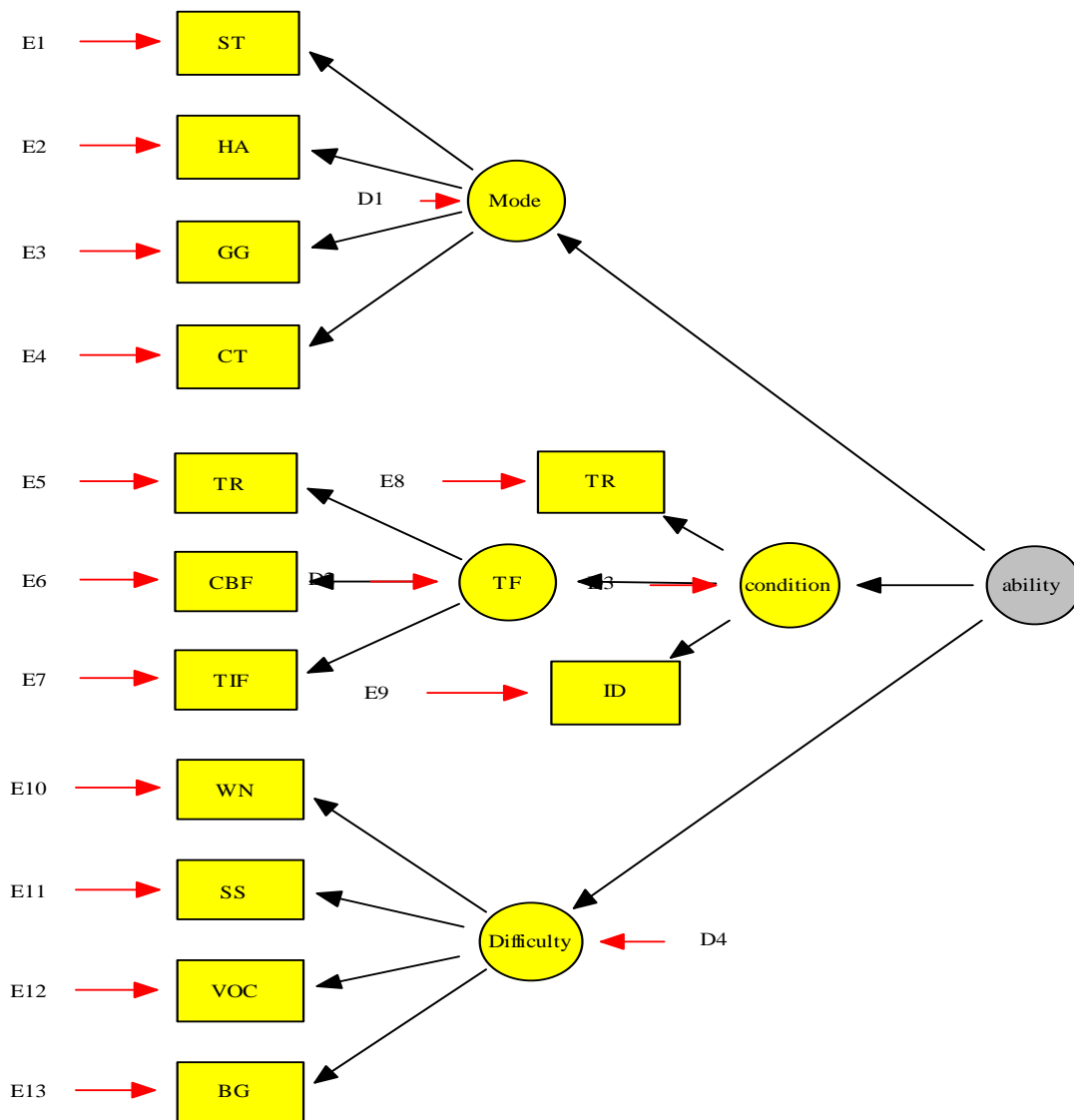


Figure 5. The Second-order Model for Test Taker’s Mode, Test Taker’s Condition and JW Test Item Difficulty

chi-square = 67.712 (df = 51)
 probability value for the chi-square statistic = .05859
 the normal theory rls chi-square for this ml solution = 63.756.
 bentler-bonett normed fit index = .780
 bentler-bonett non-normed fit index = .910
 comparative fit index (cfi) = .931
 root mean-square error of approximation (rmsea) = .052
 90% confidence interval of rmsea (.000, .083)

Figure 6. Goodness of fit indices for structural model 1 (N=120)

value, 1.96, of 95% significance, suggesting that the measurement model for the test taker's condition (diagrammed previously in Figure 3) was not reasonably designed or that the data collection from questionnaires was problematic, or both. To be more exact, this can also be interpreted as an indication that the three potential construct-irrelevant variables of examinees' familiarity to the test room, examinees' familiarity to computers, and examinees' individual condition did not influence the language ability as measured.

Based on these results, the model was revised into a more parsimonious structural model as shown in Figure 8. The new model includes a bidirectional relationship between the latent variable Test Taker's Mode and the latent variable JW Item Difficulty. In this model the two correlated latent variables are each associated with four measured variables. This

GUESS = V2	=	.153*F1	+	1.000 E2
		.022		
		7.093@		
TIME= V3	=	112.327*F1	+	1.000 E3
		10.230		
		10.980@		
CLASSRM=V6	=	-.760*F2	+	1.000 E6
		1.191		
		-.638		
COMPUTER=V7	=	-.103*F2	+	1.000 E7
		.188		
		-.546		
INDIVIDAL=V9	=	-.371*F3	+	1.000 E9
		.303		
		-1.223		

Statistics significant at the 5% level are marked with @.

Figure 7. Measurement equations with standard errors and test statistics (Iteration = 198)

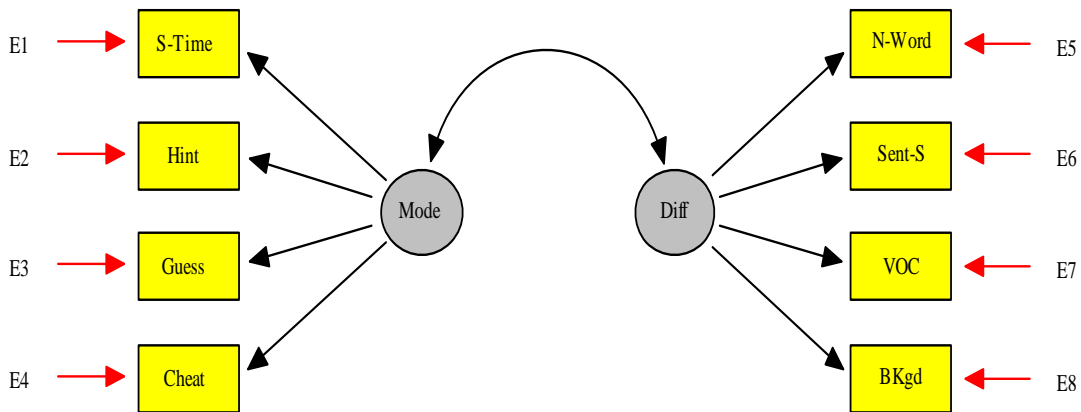


Figure 8. The revised structural model of CCT

model will need to be tested in subsequent research. Hopefully such research using SEM building under the guidance of cognitive science will help to test the CCT model and ultimately lead to significant diagnosis and estimation of language ability.

CONCLUSIONS

With the advent of the World Wide Web and the growth of the Internet, there is an increasing interest in expanding the availability of psychometric assessment services via the Internet. The traditional CAT provides some innovations over traditional linear testing that can be used for this purpose. However, a more significant expansion of assessment possibilities rests on the application of more sophisticated testing theory. The key aspects of assessment have been articulated as an “Assessment Triangle” consisting of cognition, observation and interpretation. The idea of Assessment Triangle was elaborated in the executive summary of “Knowing What Students Know: The Science and Design of Educational Assessment” compiled by The National Research Council (2001): “a model of cognition and learning, or a description of how people represent knowledge and develop competence in a subject domain, is a cornerstone of the assessment development enterprise. Unfortunately, the model of learning is not made explicit in many assessment development efforts, is not empirically derived, and/or is impoverished relative to what it could be” (p. 176).

This paper described an effort to move beyond an inexplicit assessment model to one that takes into account the three points of the triangle. From the perspective of cognitive science, the JW test item and its cognitive basis were elaborated; points of contrast between current CAT practice and CCT designs were discussed, and a pilot study of examinees’ performance on such a test was conducted. It is believed that once CCT can be put into use, it will contribute to the evolution of practice in computer-based language testing. The present paper contributes to this evolution through SEM-based research to support CRT. One thing worth mentioning is that, although SEM approach to language testing via EQS has found a wider application and computer software like PARSCALE has been used for quite some time internationally, they have rarely been utilized in language assessment before this study. Hopefully, the research presented in this paper can be held as a good starting point for further investigation of diagnostic language assessment via computerized cognitive testing.

REFERENCES

- Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 60, 451-474
- Anderson, J. R. (1976). *Language, memory and thought*. Hillsdale NJ: Lawrence Erlbaum Associates
- Anderson, J. R. (1983). *The Architecture of cognition*. Cambridge, MA: Harvard

University Press.

- Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: W.H. Freeman and Company.
- Anderson, J. R., & Gluck, K. (2001). What role do cognitive architectures play in intelligent tutoring system? In S. M. Carver & D. Klahr (Eds.), *Cognition & instruction: Twenty-five years of progress* (pp. 227-261). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bachman, L. F. & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language learning, 31*, 67-86.
- Bachman, L. F. (1998). Modern language testing at the turn of the century: assuring that what we count counts. *Newsletter of the American Association for Applied Linguistics, 21*(2), 11-13.
- Bachman, L. F. (2006). Assessment Use Argument (AUA). Report presented at International Conference for English Teaching, Shantou University, Guangdong Province, PRC.
- Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing, 15*(3), 380-414.
- Bentler, P. M. & Wu, E.J.C. (2002). *EQS6 for Windows user's guide*. Encino, CA: Multivariate Software, Inc.
- Bentler, P. M. (2006). *EQS6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.
- Binet, A. (1909). *Les idées morderne sure les enfants*. Paris: Ernest Flammarion
- Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Bock, R. D., & Atkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*, 443-469.
- Bunderson, C. V., Inonye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational Measurment* (3rd ed.) (pp. 367-408). New York: American Council on Education/MacMillan Publishing Company.
- Byrne, B. M. (1994). *Structural Equation Modeling with EQS and EQS Windows*. Thousand Oaks, CA: Sage.

- Dillon, R. F. (1985). Predicting academic achievement with models based on eye movement data. *Journal of Psychoeducational Assessment*, 3, 157-165.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn, (Ed.) *Educational measurement* (3rd ed.) (pp. 147-200). New York: American Council on Education/MacMillan Publishing Company.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.
- Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp.141-144). Washington DC: American Psychological Association.
- Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. In Arthur S. Goldberger & O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp.85-112). New York/London: Seminar Press.
- Jöreskog, K. G. (1977). Structural Equation Models in the Social Sciences: Specification estimation and testing. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265-287). Amsterdam: North Holland.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 90, 122-149.
- Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.
- Klahr, D., & Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. *Cognitive Psychology*, 13, 113-148.
- Kunnan, A. J. (1995). *Test takers characteristics and test performance: a structural modeling approach*. Cambridge: Cambridge University Press.
- Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing*, 15(3), 295-332.
- Kunnan, A. J. (1999). Recent Developments in Language Testing. *Annual Review of Applied Linguistics*, 19, 235-253.
- Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed), *Computer-assisted instruction, testing and guidance* (pp. 139-183). New York: Harper & Row.
- Mislevy, R. J., & Verhelst, N. (1987). Modeling item responses when different subjects

- employ different solution strategies. Technical Report RR-87-47-ONR, Educational Testing Service, Princeton, NJ.
- National Research Council. (2001). *Knowing what students know*. Washington DC: National Academy Press.
- Parshall, C. G., & Kromrey, J. D. (1993). Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect. Paper presented at the annual meeting of the American Education Research Association, Atlanta.
- Purpura, J. (1996). Investigating the relationships between selected cognitive characteristics of test takers and performance on language tests. Unpublished doctoral dissertation, University of California, Los Angeles.
- Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high-and-low-ability test takers; a structural equation modeling approach. *Language Testing*, 15(3), 333-379.
- Reckase, M. D. (1973). An interactive computer program for tailored testing based on the one-parameter logistic model. Paper presented at the National Conference on the Use of On-line Computers in Psychology, St. Louis Mo.
- Reese, C. (1992). Development of a computer-based test for the GRE general test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Revuelta, J., & Ponsada, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 311-327.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Education Research.
- Schoonen, R. (2005). Generalizability of Writing Scores: An Application of Structural Equation Modeling. *Language Testing*, 22(1), 1-30.
- Smith, R. M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement*. 24(23), 217-231.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn, (Ed.), *Educational measurement* (3rd ed.) (pp. 263-331). New York: American Council on Education/MacMillan Publishing Company.
- Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 14, 181-196.

- Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.) *Computerized adaptive testing: A primer* (pp.65-102). Hillsdale, NJ: Erlbaum.
- Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp.233-286). Hillsdale, NJ: Erlbaum.
- Weiss, D. J., & Betz, N. E. (1973). Ability measurement: Conventional or adaptive? (Research Report 73-1), Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, (NTIS No. AD757788).
- Weiss, D. J. (1974). Strategies of adaptive ability measurement (Research Report 74-5), Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, (pp.104-930).
- Zhang, Q. (1993). Computerized cognitive testing: Theory, method and practice. Unpublished doctoral dissertation, Guangzhou Institute of Foreign Languages, Guangzhou, RPC.
- Zhang, Q. (2002a). Computerized cognitive testing: A general introduction. Presentation made at ETS, Princeton, NJ.
- Zhang, Q. (2002b). BILOG and Parscale: Different but Alike. Paper presented at Language Testing and Teaching, International Conference, Shanghai Jiaotong University.

APPENDIX A

TEN JUMBLED WORD TEST ITEM USED FOR CCT

Jumbled Word Test Item	Key to JW Test Item
hinders, too, calcium, growth, children's, much Hint 1: Begin with 'Too'; Hint 2: The word 'hinders' used as verb; Hint 3: This is a simple sentence structure.	Too much calcium hinders children's growth.
biologists, cultivated, oysters, to, spawn, induce Hint 1: Begin with 'Biologists' Hint 2: The word 'induce' used as verb; Hint 3: This is a simple sentence structure	Biologists induce cultivated oysters to spawn.
terrible, Tom, described, the, service, sounds, that Hint 1: Begin with 'The' Hint 2: The word 'that' used as relative pronoun Hint 3: This sentence contains an imbedded attribute clause.	The service that Tom described sounds terrible.
more, hormones, than, influence, adults, do, Hint 1: Begin with 'Hormones' Hint 2: The word 'do' used as verb; Hint 3: 'more than' used as collocation.	Hormones do more than influence adults.
Awhile, glaciers, float, and melt, about Hint 1: Begin with 'Glaciers' Hint 2: The word 'float' used as verb; Hint 3: This is a simple sentence structure.	Glaciers float about awhile and melt.
what, is, their most computers, matters Hint 1: Begin with 'What'; Hint 2: the word 'is' used a verb; Hint 3: This sentence contains a subject clause.	What matters most is their computers.
they, do, left, make, with, margarine Hint 1: Begin with 'They'; Hint 2: The word 'left' used a post-modifier; Hint 3: This is a simple sentence structure.	They make do with margarine left.
complain, beaver, dams, fishing, enthusiasts, about Hint 1: Begin with 'Fishing'; Hint 2: The word 'complain used as verb;' Hint 3: This is a simple sentence structure.	Fishing enthusiasts complain about beaver dams.
would, further, delay, us, greater, cause, losses Hint 1: Begin with 'Further'; Hint 2: The word 'cause used as verb;' Hint 3: This is a simple sentence structure.	Further delay would cause us greater losses.
A, reelection, win, cartoon, helped, him Hint 1: Begin with 'A'; Hint 2: The word 'helped' used a verb; Hint 3: This is a simple sentence structure.	A cartoon helped him win reelection.

APPENDIX B

PARSCALE COMMAND FILE FOR PARTIAL CREDIT MODEL

```

CCTJW01.PSL  TOWARDS COGNITIVE RESPONSE THEORY (JUMBLED WORD DATA)
              GENERALIZED PARTIAL CREDIT MODEL - EAP SCALE SCORES

>COMMENTS
  This example scores and calibrates the data of categorical response type with response time assuming the partial credit model with standard scoring function.
  To illustrate the situation where 10 jumbled word items are involved, each with 3 relevant hints provided. Totally, 16 categories for the response type are specified.
  The standard score function assumes 16 is the high category, so response modification is required in BLOCK1.
  Thus, for response to each item produced by a test taker, there are two records: response type and response time.
  As PARSCALE accepts ordinal data, the real-valued response time presented by test takers is converted into six categories coded: Native User, Near Native User, Good User, Modest User, Average User and Poor User.
  The items are analyzed in two subtests. The first subtest consists of 10 response types and the second, of 10 response time codes.
  The data file contains the test taker ID, followed by the 10 response type and time code

>FILES  DFNAME='CCTWJ03.DAT', SAVE;
>SAVE   SCORE='CCTWJ03.SCO', COMBINE='CCTWJ03.CMB';
>INPUT  NIDW=9, NTOTAL=20, NTEST=2, LENGTH = (10,10), COMBINE=2;
        (9A1, 1X, 20A1)
>TEST1  TNAME='TYPE', ITE = (1(1)10), NBLOCK=1, SLOPES=(1.0(0)10), THRESHOLDS=(0.0(0)10);
>BLOCK1  BNAME='BLK-TYPE', NIT=10, NCAT=15, ORIGINAL=(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O),
        MODIFIED=(15,14,13,12,11,10,9,8,7,6,5,4,3,2,1);
>CALIB  PARTIAL, LOGISTIC, NQPTS=31, CYCLE = 100, NEWTON=2, CRIT=0.001, SCALE=1.7, SPRIOR;
>SCORE  MLE, SMEAN=0.0, SSD=1.0, NAME='PCM_MLE', PFQ=5;
>TEST2  TNAME='TIME', ITE = (11(1)20), NBLOCK=1, SLOPES=(1.0(0)10), THRESHOLDS=(0.0(0)10);
>BLOCK2  BNAME='BLK-TIME', NIT=10, NCAT=6, ORIGINAL=(A,B,C,D,E,F),
        MODIFIED=(6,5,4,3,2,1);
>CALIB  PARTIAL, LOGISTIC, NQPTS=31, CYCLE = 100, NEWTON=2, CRIT=0.001, SCALE=1.7, SPRIOR;
>SCORE  MLE, SMEAN=0.0, SSD=1.0, NAME='PCM_MLE', PFQ=5;
>COMBINE NAME=STRAIGHT, WEIGHTS=(0.5,0.5);
>COMBINE NAME=STRAIGHT, WEIGHTS=(0.9,0.1);

```

ⁱ For detailed limitations of current assessment, interested readers may refer to pp.26-29 in *Know what students know: The science and design of educational assessment*. National Academy Press. Washington, DC. 2001.

ⁱⁱ For details about “Assessment Triangle”, see p.2, *Know what students know: The science and design of educational assessment*. National Academy Press. Washington, DC. 2001.

ⁱⁱⁱ For detail, see reaction-time studies ,p. 98. National Research Council. (2001). *Knowing what students know*. Washington DC: National Academy Press. USA

^{iv} For detail, see reaction-time studies ,p. 99. National Research Council. (2001). *Knowing What Students Know*. Washington DC: National Academy Press. USA

^v RETCO is abbreviated from Practical English Test for Colleges administered twice a year in technical and vocational institutes and colleges in China with the total number of candidates reaching over a million a time. The author has been the chief examiner of PRETCO at Guangdong Provincial level since 1998.

^{vi} The relevant structural equation model is currently under moderation based on the information given from LM Test.

^{vii} Briefly, goodness-of-fit yielded via EQS6.1 is actually referred to the kind of matching or approximation parameter regarding the observed data to the expected model after certain designated iterations (In our case, 300 cycles were set). It can be also understood as function of the data measuring the distance between the hypothesis and the data and the probability of obtaining data. The most common tests for goodness-of-fit are the chi-square test, Kolmogorov test, and Cramer-Smirnov-Von-Mises test. EQS6.1 uses chi-square test.

^{viii} According to Kunnan (1998), generally, if any of these indices are above .90, the thumb is that there is recommendation from the indices that there is a model fit, pending examination of the Chi square statistic and model interpretability.