

Running Head: USING ATTRIBUTE HIERARCHY METHOD ON A READING TEST

Investigating the Cognitive Attributes Underlying Student Performance on a Foreign Language

Reading Test: An Application of the Attribute Hierarchy Method

Changjiang Wang, Mark J. Gierl, and Jacqueline P. Leighton

University of Alberta

Paper prepared for the graduate student poster session at the 2006 annual meeting of the National

Council on Measurement in Education, San Francisco, California

Abstract

Educational assessments are designed to facilitate teaching and learning. To achieve this purpose, the psychology underlying student performance must be well understood because most tests are based on cognitive problem-solving tasks. This calls for the integration of cognitive psychology with educational assessments. The present study illustrates the use of the attribute hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004) by applying it to a foreign language reading test. The AHM is a psychometric approach which integrates cognitive psychology with educational measurement. The results of the study indicate that, with the AHM, more cognitive diagnostic information could be produced, which can be used for guiding teaching and learning. The results also demonstrated how the AHM can be used to evaluate cognitive theories in a content domain.

Investigating the Cognitive Attributes Underlying Student Performance on a Foreign Language

Reading Test: An Application of the Attribute Hierarchy Method

Introduction

Educational assessments are designed to facilitate teaching and learning. However, due to the disjunction between cognitive psychology and educational measurement, most large-scale assessments typically yield very limited information regarding why some students perform poorly or how instructional conditions can be modified to improve teaching and learning (National Research Council, 2001). More research is now being devoted to the integration of cognitive psychology with educational assessment and several models towards this integration have been proposed. These models include, but are not limited to, the rule-space approach (Tatsuoka, 1995), the tree-based regression approach (Sheehan, 1997), the cognitive design system (Embretson, 1998), the evidence-centered design (Mislevy, Steinberg, & Almond, 2002), and the Attribute Hierarchy Method (AHM; Leighton et al., 2004). These models differ in the way cognitive information is used, but they all attempt to identify the cognitive components and processes that affect student performance on tests, resulting in cognitive diagnostic feedback that can be used to guide the teaching and learning process (Gorin, 2002). To date, the first four models have been empirically applied to the testing practice and promising results have been produced (Gorin, 2002). However, as a newly-proposed model with promising features, the AHM has not yet been applied to empirical studies. Thus, the present study illustrates the use of the AHM by using this approach to model and report on problem solving in a foreign language reading test.

The Attribute Hierarchy Method

The AHM (Leighton et al., 2004; See also Gierl, Leighton, & Hunka, 2000, in press) represents an important variation of Tatsuoka's rule-space approach because of the assumption that the cognitive attributes are hierarchically related in a model of task performance (Leighton & Gierl, 2006). A cognitive attribute is defined as a description of the procedural or declarative knowledge needed to perform a task in a specific domain (Leighton et al., 2004). In the present study, "attribute" is used as an umbrella term to refer to the cognitive processes and skills employed by students to correctly answer reading comprehension items. The assumption that cognitive attributes are hierarchically related better reflects the characteristics of human cognition because cognitive skills do not operate in isolation but belong to a network of interrelated competencies (Kuhn, 2001). Another strength of the AHM lies in its facility to guide test development. Once the hierarchy of attributes is identified for a domain, test developers can create items according to the hierarchical organization of the attributes. By doing so, the test developer achieves maximum control over the specific attributes each item measures. The AHM also offers a more convenient way of providing cognitive feedback to students. This feedback is achieved by mapping observed examinee response patterns onto expected examinee response patterns derived from the attribute hierarchy. A student with a certain observed response pattern is expected to have mastered the attributes implied by the corresponding expected response pattern. Similarly, a student may need more work on the attributes not yet mastered.

Method

Instrument

The data used in the present study were student responses to the multiple-choice items on the reading comprehension section of the 2002 administration of a provincial College English Test in China. Two random samples, each containing 1,500 examinees, were taken from the student response data for an initial analysis and then for a cross-validation analysis.

Procedure

The study was conducted in two stages. In the first stage, the *substantive* features of the items were analyzed. This stage involved specifying the attribute hierarchies and coding of the cognitive attributes measured by the test items. The second stage of the study was the *psychometric* analysis of the students' responses to the test items. This stage involved the generation of expected examinee response patterns, classification of observed response patterns, and evaluation of the attribute hierarchy.

Substantive Analysis

Specification of attribute hierarchy. The AHM starts with the specification of hierarchies according to cognitive theories in a content domain. In the present study, a number of cognitive attributes which are generally used by students in second language and foreign language (L2) reading comprehension and the relationship among these cognitive attributes were first identified through a review of literature in L2 reading comprehension (e.g., Farhady & Hessamy, 2005; Urquhart & Weir, 1998). Then, according to the relationship among the cognitive attributes, two attribute hierarchies were specified. The specification of two attribute hierarchies allowed us to

compare which of the attribute hierarchies reflected more accurately the cognitive attributes used by students to answer the test items.

Coding the cognitive attributes. Two expert raters who were experienced in teaching English as a foreign language and who were familiar with the test-taker population of the test in question were recruited. As an item can measure multiple cognitive attributes, the raters were instructed to code all cognitive attributes measured by the items. Items for which there was good agreement between the raters on the attributes were selected for AHM analysis.

Psychometric Analysis

Generation of expected examinee response patterns. After the attribute hierarchies were specified and the cognitive attributes of the items were coded, expected examinee response patterns were generated by matching the cognitive attributes required by the items with the possible attribute patterns of expected examinees. An expected examinee is a hypothetical examinee who correctly answers items that require only specific cognitive attributes that the examinee has mastered. If the attribute pattern of an expected examinee includes the cognitive attributes required by the item, then this examinee is expected to answer this item correctly. Conversely, if at least one of the cognitive attributes required by the item is missing in an expected examinee's attribute pattern, then the examinee may not answer this item correctly.

Classification of observed response patterns. After the expected response patterns were generated, the observed response patterns were classified. The classification was conducted in three steps. First, the parameters of each item were calibrated with 3-parameter logistic model using BILOG-MG (du Toit, 2003). Second, each observed response pattern was compared

against all expected response patterns and slips of both $0 \rightarrow 1$ and $1 \rightarrow 0$ were identified. Third, The product of the probabilities of each slip was calculated using the item parameters calibrated in the first step. This gave the likelihood that the observed response pattern was generated from each expected response pattern at the given ability level. The observed response pattern would then be classified as being generated from the expected response pattern with the largest likelihood.

Evaluation of the attribute hierarchies. To examine whether the specified attribute hierarchy truthfully reflect the cognitive attributes used by the students, the hierarchy consistency index (HCI) (Cui, Leighton, Gierl, & Hunka, 2006) was used. The values of the index range between -1 and $+1$. Low values of HCI indicate inconsistency between the observed response patterns and the expected response patterns specified by the attribute hierarchy, suggesting that the attribute hierarchy needs improvement. The mean and standard deviation of the HCI can be used as indicators of the overall model-data fit. A high mean and low standard deviation suggest the observed response patterns fit the AHM model well. In other words, the HCI results help validate the specified attribute hierarchies.

Results

Substantive Analysis

Specifying the Attribute Hierarchies

Based on a review of the literature in L2 reading comprehension, eight cognitive attributes involved in L2 reading and in responding to reading items were identified. These cognitive attributes include:

- (1) basic language knowledge (BA), such as word recognition and basic syntactic knowledge;
- (2) understanding the content, form, and function of sentences (US);
- (3) understanding the content, form, and function of larger sections of text (UT);
- (4) analyzing authors' purposes, goals, and strategies (PGS);
- (5) determining word meaning in context (WM);
- (6) making inferences based on background knowledge (INF);
- (7) understanding text with difficult vocabulary (VC); and
- (8) understanding text with complex syntactic structure (SY) (e.g., Alderson, 2000; Cain, Oakhill, & Lemmon, 2004; Perfetti, 1985, 1988; Urquhart & Weir, 1998).

Among the eight attributes, BA is assumed to have been mastered by all the examinees and is the prerequisite for all other attributes. US is a sentence-level attribute and is the prerequisite for UT. PGS, WM, and INF all require the understanding of contextual information and therefore require UT. Based on the interrelationship among these six attributes, Hierarchy 1 was specified in Figure 1a. To achieve finer granularity, Hierarchy 2 (Figure 1b) was specified with VC and SY included. These two attributes were included because, when difficult vocabulary or sentences were encountered, the reader would have to rely on morphological clues or contextual cues (e.g., Perfetti, 1985). In other words, understanding difficult vocabulary and sentences represents different cognitive attributes from understanding easy vocabulary and sentences.

Coding the Attributes

There were 20 items in the reading section of the test. However, four items were not used

for coding because they had negative discrimination. Also, as mentioned previously, BA was assumed to have been mastered by all examinees, and thus was not coded by the raters. For the attributes US, UT, PGS, INF, and WM, the raters first had a discussion about the meaning of these attributes. Once they achieved the same understanding of these attributes, they were instructed to code the cognitive attributes measured by the items. For the attributes of VC and SY, the raters were first instructed to identify the parts of text students needed to read to correctly answer the items. Then they were asked to rate the difficulty level of the key vocabulary and the syntactic complexity of these texts. If the key vocabulary was considered difficult for the target test-taker population, then VC was considered to be measured by the item. Similarly, if the syntactic structure of the text was considered complex, then SY was considered to be measured by the item. In the end, the raters achieved agreement on nine of the remaining 16 items, which were then used in the subsequent analyses. The coding results of the nine items were presented in Table 1.

Psychometric Analysis

Generation of Expected Examinee Response Patterns

The expected examinee response patterns, the corresponding ability levels, and expected examinee attribute patterns generated from the two attribute hierarchies are respectively shown in Table 2a and Table 2b. Hierarchy 1, with six cognitive attributes, generated 10 unique examinee attribute patterns and expected examinee response patterns. With two more cognitive attributes, Hierarchy 2 generated 37 unique attribute patterns but only 25 unique expected examinee response patterns. This result occurred because some different attribute patterns may

generate the same expected examinee response patterns due to the small number of items (e.g., the attribute patterns of both “11100000” and “11100100” produced the same expected examinee response pattern of “110000000” because none of the nine items measures the attributes of BA, US, UT, and WM.).

As an illustration, in Table 2a, Row 3 should be interpreted as follows: An examinee who has attributes of BA, US, and UT is expected to answer items 1, 2, 5, 7, and 8 correctly, producing the expected examinee response pattern (110010110). However, Row 4 in Table 2b indicates that an examinee with only attributes of BA, US, and UT is expected to answer only items 1 and 2 correctly but not items 5, 7, and 8, because items 5, 7, and 8 measure the attributes of VC or SY in addition of BA, US, and UT. In other words, Hierarchy 2 has finer granularity than Hierarchy 1.

Classification of Observed Response Patterns

To classify the observed response patterns, each observed response pattern is compared against all expected examinee response patterns where slips of the form $0 \rightarrow 1$ and $1 \rightarrow 0$ are identified. The product of the probabilities of each slip was calculated to give the likelihood that the observed response pattern was generated from an expected examinee response pattern for a given ability level. An examinee with a certain observed response pattern is likely to have mastered the attributes in the expected examinee response pattern which had the largest likelihood.

Summaries of the classification results from Hierarchies 1 and 2 are, respectively, displayed in Tables 3a and 3b. Tables 3a and 3b show that different numbers of examinees have

been classified into different ability levels. Moreover, the numbers of examinees classified into each ability level are comparable across the two samples. In addition to classifying the examinees, the tables also indicate the attributes the examinees have and have not mastered. For example, Table 3a indicates that 495 examinees in Sample 1 have been classified into the ability level of 0.529, and 31 examinees have been classified into the ability level of 0.544. Although the ability levels of these two groups of examinees are similar, the attributes that they have mastered are not the same. The examinees at the ability level of 0.529 have mastered the attributes of BA, US, UT and PGS, while the examinees with an ability level of 0.544 have mastered the attributes of BA, US, UT, INF, and WM. In other words, with the AHM, the examinees not only get an overall ability estimate, but they also get the cognitive diagnostic information as to which attributes they have likely mastered and which attributes they still need to acquire or strengthen.

Evaluation of the Two Attribute Hierarchies

Two attribute hierarchies with different granularity were used in the present study. To evaluate the model-data fit of the two attribute hierarchies, the means and standard deviations of the HCI (Cui et al., 2006) of the two attribute hierarchies were calculated and the results are presented in Table 4. As there are no objective criteria for the HCI values at the current stage, the evaluation of the two hierarchies can only be assessed in comparative terms. As can be seen from Table 4, comparable values of mean and standard deviation of the HCI were produced across the two samples for each hierarchy. Moreover, for both samples, higher means and lower standard deviations of the HCI were produced from Hierarchy 2 than from Hierarchy 1. These results

indicate that Hierarchy 2, as a more fine-grained hierarchy than Hierarchy 1, reflected greater accuracy about the cognitive attributes used by the examinees as the model provides a better fit to the observed response data.

An Illustration

To illustrate how the AHM can be used to make diagnostic inferences, we show the classification results of two observed response patterns, “110010101” (henceforth called ORP1) and “110110100” (henceforth called ORP2) in Table 5a and Table 5b. Table 5a represents the classification results by Hierarchy 1 and Table 5b represents those by Hierarchy 2.

As can be seen from Tables 5a and 5b, although five items were correctly answered in both ORP1 and ORP2, they were classified into different ability levels with different attribute patterns. In Table 5a, Hierarchy 1 classified ORP1 into the ability level of 0.450 because the expected examinee response pattern “110010111,” which was associated with this ability level, had the largest likelihood among all expected examinee response patterns. The corresponding attribute pattern of this expected examinee response pattern is “111001,” indicating that students with ORP1 have mastered the attributes of BA, US, UT, and WM. On the other hand, ORP2 was classified into the ability level of -0.130 because the expected examinee response pattern “110010110,” which was associated with this ability level, had the largest likelihood among all expected examinee response patterns. The corresponding attribute pattern of this expected examinee response pattern is “111000,” indicating that students with ORP2 have mastered the attributes of only BA, US, and UT, but not WM, and thus got a lower ability estimate.

An inspection of Table 5a and Table 5b also indicates that Hierarchy 2, with its finer

granularity, provided diagnostic information on two more attributes, VC and SY, than Hierarchy 1. For example, Hierarchy 1 only indicated that students with ORP1 have mastered the attributes of BA, US, UT, and WM, and that they need more work on the attributes of PGS and INF. However, Hierarchy 1 could not provide any information regarding the attributes of VC and SY because its granularity is not fine enough to include these two attributes. On the other hand, Hierarchy 2 indicated that the same students have mastered the attributes of BA, US, UT, WM, and VC, and they need more work on the attributes of PGS, INF, and SY, thus provided more diagnostic information than Hierarchy 1. These results also indicate the importance of selecting an appropriate cognitive model in a content domain when using the AHM.

Discussion

The present study was designed to illustrate the use of the AHM by applying this method to a L2 English reading test. The AHM represents an integration of cognitive psychology and psychometrics. It starts with cognitive theories of a content domain and then involves substantive and psychometric analyses. As a result, the present study was conducted in two stages, with substantive analyses followed by psychometric analyses on the test. In the substantive analyses, eight cognitive attributes involved in L2 reading comprehension were identified and two attribute hierarchies with different granularity were specified from a review in the literature of L2 reading comprehension (Figure 1a and Figure 1b). The purpose of specifying two hierarchies is to compare which of the attribute hierarchies reflected more accurately the cognitive attributes used by students in answering the test items. Then, two expert raters were asked to code the cognitive attributes measured by the test items. Nine items for which good agreement was achieved on the

cognitive attributes were selected for the psychometric analyses (Table 1).

The psychometric analyses include the generation of expected examinee response patterns, classification of observed response patterns, and evaluation of attribute hierarchies. In the present study, the two attribute hierarchies generated different number of expected examinee response pattern (Tables 2a and 2b). Based on the two attribute hierarchies and the expected examinee response patterns, the examinees' observed response patterns were classified into different ability levels with different attribute patterns (Tables 3a and 3b). With the classification, the AHM provides not only an overall ability estimate, but also important diagnostic information to the examinees. Thus, the examinees would be able to know which attributes have been mastered and which attributes require additional work. Finally, the performance of the two attribute hierarchies was evaluated using the HCI. The results of HCI analysis indicated that Hierarchy 2 more accurately reflected the cognitive model underlying the examinee data on the test, as indicated by the higher values of mean and lower values of standard deviation. Moreover, with its finer granularity, Hierarchy 2 provided richer diagnostic information to the examinees.

The present study is the first that applies the AHM using empirical testing data. As a cognitively-based psychometric approach, the AHM has obvious advantages over traditional psychometric scoring methods for three reasons. First, in addition to providing an overall ability estimate, the AHM provided more detailed cognitive diagnostic information to the examinees, with which the examinees could make focused efforts on attributes that they have not yet mastered. In other words, the AHM can be instrumental in improving student learning. Second, the AHM also provides useful information for construct validation. As an indispensable

component of the AHM, the substantive analysis yields information as to what the items are intended to measure. Then, if HCI results indicate good fit between the attribute hierarchies and the observed response data, then positive evidence about the construct validity of the test can be obtained. Third, the AHM has the mechanism of feeding back to the cognitive theories of a content domain. For example, in the current study, Hierarchy 2 was found to reflect more accurately the cognitive attributes students use on the test. Such results indicate that VC and SY are two important cognitive attributes that should not be ignored in a cognitive model of L2 reading comprehension.

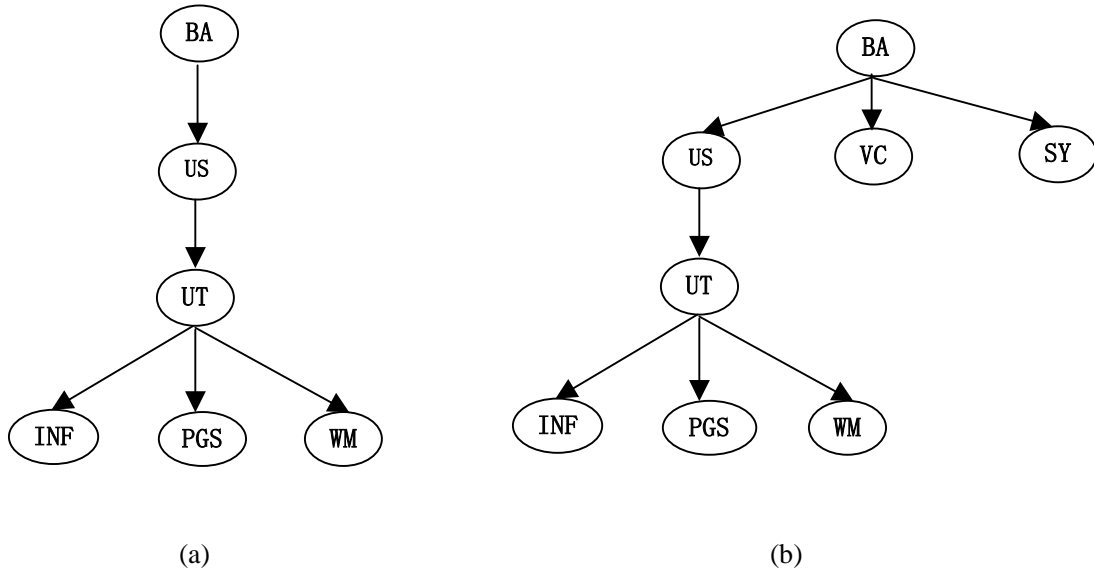


Figure 1: The Two Hierarchies of L2 Reading Comprehension Specified in the Present Study.

Table 1

The cognitive attributes measured by the test items (“1” indicates the cognitive attribute is measured; “0” indicates the cognitive attribute is not measured)

	Item 1	Item 2	Item 3	Item 4	Item 5	Item6	Item 7	Item 8	Item 9
BA	1	1	1	1	1	1	1	1	1
US	1	1	1	1	1	1	1	1	1
UT	1	0	1	1	0	1	0	0	1
INF	0	0	1	0	0	0	0	0	0
PGS	0	0	0	1	0	1	0	0	0
WM	0	0	0	0	0	0	0	0	1
VC	0	0	0	0	1	0	1	0	1
SY	0	0	0	0	0	1	0	1	1

Table 2a

Expected Examinee Response Patterns, Examinee Attributes, and Ability Levels Based on Hierarchy 1

Expected Examinee	Examinee Attributes	Expected Examinee Response Pattern	Ability Level
1	100000	000000000	-2.239
2	110000	010010110	-0.823
3	111000	110010110	-0.130
4	111100	111010110	-0.088
5	111010	110111110	0.529
6	111110	111111110	0.659
7	111001	110010111	0.450
8	111101	111010111	0.544
9	111011	110111111	1.243
10	111111	111111111	1.517

Table 2b

Expected Examinee Response Patterns, Examinee Attributes, and Ability Levels Based on Hierarchy 2

Expected Examinee	Examinee Attributes	Expected Examinee Response Pattern	Ability Level
1	10000000	00000000	-2.226
2	11000000	01000000	-1.582
3	11000001	010000010	-1.495
4	11100000 or 11100100	110000000	-1.164
5	11110000 or 11110100	111000000	-1.162
6	11100001 or 11100101	110000010	-1.022
7	11110001 or 11110101	111000010	-1.019
8	11101000 or 11101100	110100000	-1.016
9	11111000 or 11111100	111100000	-1.014
10	11000010	010010100	-0.993
11	11000011	010010110	-0.798
12	11101001 or 11101101	110101010	-0.674
13	11111001 or 11111101	111101010	-0.667
14	11100010 or 11100110	110010100	-0.453
15	11110010 or 11110110	111010100	-0.446
16	11101010 or 11101110	110110100	-0.202
17	11111010 or 11111110	111110100	-0.19
18	11100011	110010110	-0.142
19	11110011	111010110	-0.126
20	11100111	110010111	0.45
21	11110111	111010111	0.491
22	11101011	110111110	0.53
23	11111011	111111110	0.589
24	11101111	110111111	1.269
25	11111111	111111111	1.429

Table 3a

Classification Results Based on Hierarchy 1

Examinee Attributes	Expected Examinee Response Pattern	Ability Level	No. of Examinees (Sample 1)	No. of Examinees (Sample 2)
100000	000000000	-2.239	138	173
110000	010010110	-0.823	216	246
111000	110010110	-0.130	195	192
111100	111010110	-0.088	43	41
111010	110111110	0.529	495	422
111110	111111110	0.659	230	235
111001	110010111	0.450	72	63
111101	111010111	0.544	31	43
111011	110111111	1.243	73	66
111111	111111111	1.517	7	19

Table 3b

Classification Results Based on Hierarchy 2

Examinee Attributes	Expected Examinee Response Pattern	Ability Level	No. of Examinees (Sample 1)	No. of Examinees (Sample 2)
10000000	00000000	-2.226	38	42
11000000	01000000	-1.582	26	41
11000001	01000010	-1.495	13	28
11100000 or 11100100	11000000	-1.164	83	69
11110000 or 11110100	11100000	-1.162	33	56
11100001 or 11100101	11000010	-1.022	34	35
11110001 or 11110101	11100010	-1.019	5	3
11101000 or 11101100	11010000	-1.016	231	193
11111000 or 11111100	11110000	-1.014	40	39
11000010	010010100	-0.993	83	95
11000011	010010110	-0.798	13	17
11101001 or 11101101	110101010	-0.674	73	66
11111001 or 11111101	111101010	-0.667	7	19
11100010 or 11100110	110010100	-0.453	65	57
11110010 or 11110110	111010100	-0.446	5	10
11101010 or 11101110	110110100	-0.202	22	35
11111010 or 11111110	111110100	-0.19	8	2
11100011	110010110	-0.142	47	52
11110011	111010110	-0.126	6	9
11100111	110010111	0.45	333	348
11110111	111010111	0.491	35	51
11101011	110111110	0.53	134	113
11111011	111111110	0.589	19	16
11101111	110111111	1.269	126	92
11111111	111111111	1.429	21	12

Table 4

HCI Results for Hierarchy 1 and Hierarchy 2

	Hierarchy 1		Hierarchy 2	
	Sample 1	Sample 2	Sample 1	Sample 2
Mean of HCI	0.2883	0.2691	0.5539	0.5411
SD of HCI	0.4589	0.4911	0.3931	0.4190

Table 5a

Classification of Observed Response Patterns (110110100) and (110010101) Using Hierarchy 1

Examinee Attributes	Expected Examinee Response Pattern	Ability Level	Observed Response Pattern 110010101		Observed Response Pattern 110110100	
			Likelihood	Slips	Likelihood	Slips
			100000	000000000	-2.239	0.012
110000	010010110	-0.823	0.322	3	0.246	3
111000	110010110	-0.130	0.465	2	0.336	2
111100	111010110	-0.088	0.416	3	0.300	3
111010	110111110	0.529	0.085	4	0.262	2
111110	111111110	0.659	0.066	5	0.211	3
111001	110010111	0.450	0.551	1	0.063	3
111101	111010111	0.544	0.469	2	0.051	4
111011	110111111	1.243	0.038	3	0.013	3
111111	111111111	1.517	0.020	4	0.006	4

Table 5b

Classification of Observed Response Patterns (110110100) and (110010101) Using Hierarchy 2

Examinee Attributes	Expected Examinee Response Pattern	Ability Level	Observed Response Pattern 110010101		Observed Response Pattern 110110100	
			Likelihood	Slips	Likelihood	Slips
10000000	00000000	-2.226	0.012	5	0.012	5
11000000	01000000	-1.582	0.050	4	0.044	4
11000001	010000010	-1.495	0.044	5	0.039	5
11100000 or 11100100	110000000	-1.164	0.127	3	0.103	3
11110000 or 11110100	111000000	-1.162	0.114	4	0.092	4
11100001 or 11100101	110000010	-1.022	0.110	4	0.087	4
11110001 or 11110101	111000010	-1.019	0.099	5	0.078	5
11101000 or 11101100	110100000	-1.016	0.086	4	0.283	2
11111000 or 11111100	111100000	-1.014	0.077	5	0.253	3
11000010	010010100	-0.993	0.385	2	0.303	2
11000011	010010110	-0.798	0.326	3	0.249	3
11101001 or 11101101	110101010	-0.674	0.048	6	0.147	4
11111001 or 11111101	111101010	-0.667	0.043	7	0.131	5
11100010 or 11100110	110010100	-0.453	0.633	1	0.464	1
11110010 or 11110110	111010100	-0.446	0.567	2	0.415	2
11101010 or 11101110	110110100	-0.202	0.346	2	1.000	0
11111010 or 11111110	111110100	-0.19	0.309	3	0.893	1
11100011	110010110	-0.142	0.465	2	0.335	2
11110011	111010110	-0.126	0.416	3	0.300	3
11100111	110010111	0.45	0.551	1	0.063	3
11110111	111010111	0.491	0.481	2	0.054	4
11101011	110111110	0.53	0.085	4	0.262	2
11111011	111111110	0.589	0.071	5	0.223	3
11101111	110111111	1.269	0.037	3	0.012	3
11111111	111111111	1.429	0.024	4	0.008	4

References

- Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.
- Cain, K., Oakhill, J. V., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96*, 671-681.
- Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2006, April). *The Hierarchical Consistency Index: A person-fit statistic for the Attribute Hierarchy Method*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- du Toit, M. (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Lincolnwood, IL: Scientific Software International.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396.
- Farhady, H., & Hessamy, G. (2005, July). *An empirical investigation of the L2 reading comprehension skills*. Paper presented at the annual meeting of the Language Testing Research Colloquium, Ottawa, Canada.
- Gorin, J. S. (2002). *Cognitive and psychometric modeling of text-based reading-comprehension GRE-V items*. Unpublished doctoral dissertation, University of Kansas, Kansas.
- Kuhn, D. (2001). Why development does (and does not occur) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 221-249). Hillsdale, NJ: Erlbaum.
- Leighton, J. P., & Gierl, M. J. (2006). Identifying and evaluating cognitive models in educational measurement. *Educational Measurement: Issues and Practice*, Manuscript submitted for publication.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205-236.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477-496.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Perfetti, C. A. (1985). *Reading ability*. London: Oxford University Press.
- Perfetti, C. A. (1988). Verbal efficiency in reading ability. In M. Daneman, G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 109-144). New York: Academic Press, Inc.
- Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333-352.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.
- Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. London: Longman.