International Journal of Learning, Teaching and Educational Research
Vol. 12, No. 3, pp. 61-79, July 2015

Application of Conditional Means for Diagnostic Scoring

Hollis Lai[1]
Mark J. Gierl[2]
Oksana Babenko[3]

[1] School of Dentistry, Faculty of Medicine and Dentistry
[2]Centre for Research in Applied Measurement and Evaluation
[3]Department of Family Medicine, Faculty of Medicine and Dentistry
University of Alberta, Canada

Abstract. In educational assessment, demand for diagnostic information from test results has prompted the development of model-based diagnostic assessments. To determine student mastery of specific skills, a number of scoring approaches, including subscore reporting and probabilistic scoring solutions, have been developed to score diagnostic assessments. Although each approach has a unique set of limitations, these approaches are, nevertheless, often used in diagnostic scoring, whereas an alternative approach, Complex Sum Scores (CSS), has not received much attention yet. With the process of developing model-based diagnostic assessments becoming increasingly complex, we revisit the CSS and demonstrate two applications of the CSS in the development of diagnostic assessments. Two applications include: (a) illustrating and validating skills within the model, and (b) partial mastery scoring using model-based distractors. By demonstrating the two applications, we aim to show how model-based diagnostic assessments can be developed and scored using the CSS scoring approach, the results of which can be used by teachers to inform teaching and learning.

Keywords: Model-based diagnostic assessments; diagnostic scoring; complex sum scores

Introduction

Demands for diagnostic and formative feedback on student learning have led to significant changes in student assessment, including the ways tests are developed, administered, and scored. One such example is cognitive diagnostic assessment (CDA; Nichols, 1994). A cognitive model of task performance is used

to guide the development of a CDA, specifically when constructing test items that probe student mastery on a specific set of skills (Leighton & Gierl, 2007). In order to make inferences about student mastery on a set of assessed skills, various probabilistic scoring methods have been developed, with each method suited to measure different types of skills. The increasing complexity of probabilistic scoring methods, however, has raised the question of interpretability of the results obtained when such methods are used. To ensure diagnostic results are clearly understood by teachers and parents, subscore reporting has been used as an alternative to probabilistic methods (Wainer et al., 2001; Sinharay, Puhan, & Haberman, 2010). In this approach, scores on each cluster (i.e., subscale) are reported as diagnostic information about students' mastery/non-mastery on the skills assessed by a test. Although both approaches (probabilistic scoring and subscore reporting) are possible in scoring cognitive diagnostic assessments, the two approaches differ substantially in several ways (i.e., complexity, precision, etc.).

In this article, we address the complexity associated with diagnostic scoring and introduce an alternative approach that, in our opinion, can (a) ease computational intensity without unduly sacrificing precision, and (b) assist test item writers in developing cognitive diagnostic assessments and teachers in using CDA results to inform instruction. In our proposed alternative approach, skills to be assessed are organized using a cognitive diagnostic modeling method, Attribute Hierarchy Method (AHM; Leighton, Gierl, & Hunka, 2004), and students' responses are scored using a conditional score method, Complex Sum Scores (CSS; Henson, Templin, & Douglas, 2007). Before describing the alternative approach in detail, we start with a review of existing methods commonly used in the development and scoring of diagnostic assessments, highlighting the advantages and limitations of these methods. Using the real response data from an existing CDA program, we then demonstrate two applications of the proposed scoring approach, namely: a) illustrating and validating of the skills specified in the attribute hierarchy model, and b) partial mastery scoring using model-based distractors. We conclude with a discussion of why the proposed approach is a better alternative to more complex diagnostic scoring methods and, thus, may be appealing to testing programs that are interested in implementing cognitive diagnostic assessment but lack psychometric resources for this.

Review of Frameworks for CDA Development

Gorin (2007) conceptualized the development of a cognitive diagnostic assessment (CDA) consisting of two components: a) development of cognitive models to be used subsequently in item development, and b) statistical methods to be used in scoring students' responses. Various CDA development frameworks are used in defining cognitive models and creating items (see Mislevy, 1994; Embretson, 1994; Luecht, 2008). For the purposes of this paper, a CDA development framework can be generalized to include the following principles or assumptions: a) the assessed skill attributes are such that they can

be classified in the mastery/non-mastery manner, b) items are developed to probe a specified pattern of attributes, and c) the correct response on an item implies *evidence* for mastery of the probed attributes. To operationalize the CDA framework, cognitive psychologists and subject matter experts are involved in developing cognitive models, which item writers use to guide them in the development of items that probe the patterns of attributes as specified in the cognitive models (see Figure 1). After CDA models and items are developed and administered to students, statistical methods are required to determine student mastery of the assessed skills.

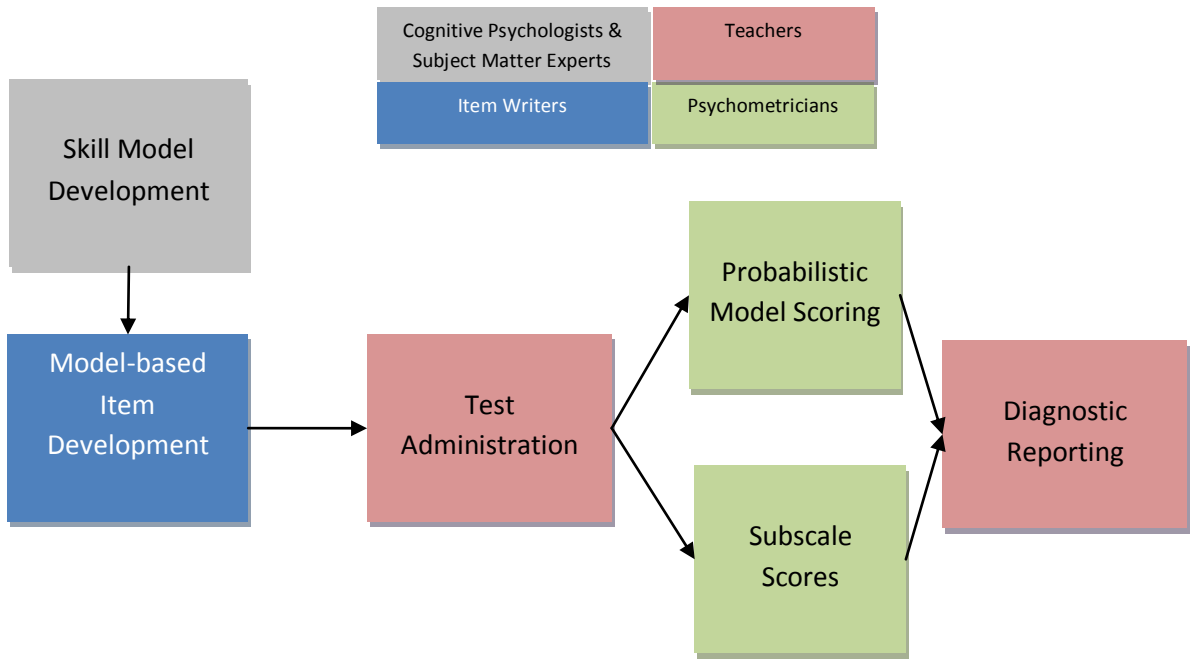


Figure 1. Illustration of the general CDA process

Two diagnostic scoring methods are used: probabilistic modeling and subscores. For probabilistic models, Fu and Li (2007) summarized 62 different methods that had been developed for diagnostic scoring. According to Sinharay, Puhan, and Haberman (2010), these methods have the following features in common: 1) test items require one or more skills to elicit a correct response, 2) students are assumed to have a latent ability associated with each skill, and 3) the likelihood of a correct response is a probabilistic function that can be determined based on the interaction of item characteristics and student's ability level. Each diagnostic scoring method differs in the assumptions made about the assessed skills. Specifically, models are created to describe skills that may be unidimensional or multi-dimensional (de la Torre & Patz, 2005), skill mastery may be classified dichotomously (DINA; Junker & Sijtsma, 2001) or polytomously (GDM; von Davier, 2008), and the structure of skills may be rigid (AHM; Leighton, Gierl, & Hunka, 2004) or flexible (RUM; Hartz, 2002). That is, characteristics of the assessed skills inform the choice of a diagnostic scoring method to be used to score student responses on a CDA.

In addition to the assumptions made about the assessed skills, there are several data requirements associated with the use of probabilistic scoring methods. First, these methods often require large data for estimating item parameters and student ability levels; however, such data are often not available at the initial development stage. Second, in the absence of real data, scoring methods are validated using simulated data; however, simulated data can provide invalid evidence of performance when students' actual responses do not fit the expected response pattern. To address these concerns, educational researchers have suggested the use of subscores for scoring student responses and reporting diagnostic results.

Subscale scores or subscores are parts of the total score that reflect student mastery on specific content areas that comprise the whole domain assessed by the test. Correspondingly, all subscores on a test can be summed to obtain the total score for each student, provide that each item is referenced to one and only one subscale. With each item being referenced to only one skill, subscores allow for a straightforward interpretation of the CDA results. However, in the CDA context, this is problematic because CDA items are designed to probe more than one skill. Earlier research also suggests that subscores provide little or no added value when subscale reliabilities are not high (Sinharay, 2010; Babenko & Rogers, 2014). Given the limitations associated with probabilistic scoring and subscore methods, we propose an alternative diagnostic scoring approach that will be of interest to assessment programs that may lack the expertise required for developing and scoring cognitive diagnostic assessments.

An Alternative Framework for CDA Development

Schematically, the alternative framework for CDA development is shown in Figure 2, and explained in subsequent sections. Development of cognitive diagnostic assessments requires a structure or model of cognitive skills. In this study, we applied the Attribute Hierarchy Method (AHM; Leighton, Gierl, & Hunka, 2004) to frame the skills to be assessed and provide guidance for item development. In AHM, skills are assumed to be mastered by students in a progression of an ordered hierarchy. This process requires involvement of experts familiar with both the cognitive processes of the target students and the content being assessed by the test. To validate the attribute structure, model data fit indices such as the Hierarchy Consistency Index (HCI; Cui & Leighton, 2009) are used to verify student response patterns against expected patterns from the model. Although the HCI and other indices provide an overall model data fit measure, they are of little help for item writers seeking to inform and refine their item development process. Further, little has been done to verify whether diagnostic scoring outcomes follow expected trend as specified by the attribute model. To address these limitations, we introduce an alternative scoring method, Complex Sum Scores (CSS), which also is used for the partial mastery scoring in the alternative CDA framework (see Figure 2). First, we review the sum score approaches, with a focus on the CSS method.

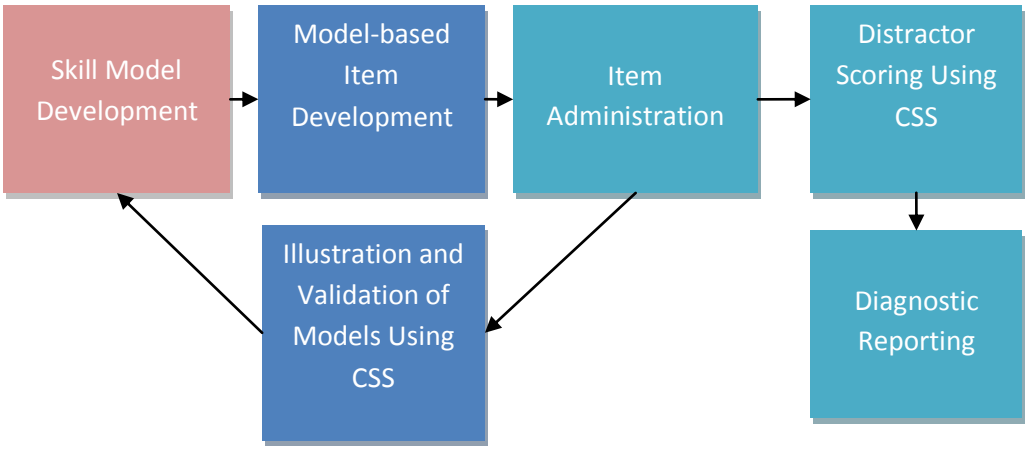


Figure 2. An illustration of the alternative framework for CDA development

Sum Score Approaches

Sum score approaches can be conceptualized as a compromise between subscore and probabilistic scoring methods. Recognizing the need for simplified model-based diagnostic scores, Henson, Templin, and Douglas (2007) proposed diagnostic scoring methods using sum scores. In diagnostic assessments, the relationship between test items and specific sets of skills they probe is defined by a Q-matrix. Henson et al. (2007) suggested that skills could be scored using the conditional sums of the correct responses to corresponding items as defined under the Q-matrix. For example, under a dichotomously classified Q-matrix, where each row represents an item and each column represents a skill or attribute, the concept of sum scores is represented as:

$$X_k = \sum^{j} (x_j \times q_{jk})$$

where x_j is the dichotomously scored student response for item j, with the responses being summed if item j requires the use of attribute k in the Q-matrix. Based on this concept, Henson et al. (2007) introduced three types of sum scores. The first and the simplest, called simple sum score (SSS), is statistically identical to the subscore method used in diagnostic scoring, with each item representing only one attribute. Recognizing that each item may probe more than one attribute as specified in the Q-matrix, the complex sum score (CSS) method was introduced, with items contributing to more than one sum score. The third and most complex type, called weighted-complex sum score (WCSS), was introduced to provide weighted representation of skills on a given item, because a dichotomous representation of skill mastery used in CSS and SSS may not reflect the process of skill acquisition as it occurs in reality. In the present study, we extend the use of the sum score methods, in particular the CSS method, to scoring diagnostic assessments and demonstrate that the results obtained using sum score methods may be as accurate as the results obtained using probabilistic scoring methods.

CSS is a conditional sum approach to diagnostic scoring. What it means is that for an item that probes, for example, attributes 1 and 3 out of four attributes assessed by a test, the correct response on the item contributes to the conditional sum of attributes 1 and 3, whereas an incorrect response does not change the conditional sum of the attributes. The higher the value of the conditional sum, the higher the level of attribute mastery is inferred (Henson et al., 2007). In order to provide a better measure of skill mastery, we suggest the use of conditional *means* in place of conditional *sums*. The conditional mean of the CSS is given as:

$$X_k = \frac{\Sigma_{k=1}^{j} x_j}{n_k},$$

where CSS of attribute k is a factor of the number of items probing attribute k, n_k. This modification makes a CSS a *proportional measure* of mastery rather than being a raw value that depends on the number of items probing each attribute. Next, diagnosticity of a scoring method is highly dependent on both the specificity of the attribute and the alignment of the item with the attribute it is supposed to probe (Gierl, Cui, & Zhou, 2009). Therefore, CSS is a model-dependent scoring method, requiring attributes to be defined in a structure (such as a hierarchy) prior to item development. The next section describes how CSS can be applied to provide validation and illustration for the structure of attributes, and in assessing partial mastery using distractor responses.

Illustrating and Validating the Structure of Skill Attributes

The CSS scoring method provides a raw-score measure for multidimensional diagnostic data and can be used to describe individual attributes and their interactions. Consider the model of attributes in Figure 3 to be probed or assessed by a diagnostic assessment. This model includes six attributes that are organized linearly.

A model of linearly ordered attributes (i.e., skills) suggests that: (a) each attribute is acquired in a sequential manner, and (b) each attribute is a pre-requisite of subsequent attributes. This implies that the mean CSS for one attribute should not exceed the ratio of its parent or previous attributes:

$$CSS_{A2} = CSS_{A1} + CSS_{A2 \mid A1}$$

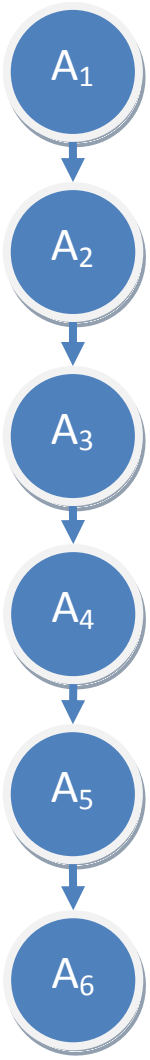


Figure 3. A model of linearly ordered attributes

The mean CSS on attribute k represents the average proportion of mastery for this attribute from the given sample of students. This information is important because it provides a p-value equivalent or difficulty at the attribute level. To confirm or validate the hypothesized structure of this model, we expect that the size of differences between mean CSSs should be in the order specified in the model. That is, the expectation for the structure of attributes to be valid is that the differences increase between non-adjacent attributes (e.g., the difference between A1 and A2 is smaller than the difference between A1 and A3, which in turn is smaller than the difference between A1 and A4, etc.). To describe distances between attributes, a mean deviation statistic is applied to the obtained CSS's. In the present study, the Mean Absolute Difference (MAD) is used to describe the relationship between pairs of attributes, and to quantify the distances among attributes in order to make meaningful inferences about student mastery. For example, to determine MAD between attributes A2 and A1, with A1 being a prerequisite skill of A2 (see Figure 3), the average of absolute differences across all examinees is computed using the formula:

$$\mathrm{MAD}_{\mathrm{A2-A1}} = \frac{\sum_n |CSS_{A2} - CSS_{A1}|}{N},$$

where CSS is the complex sum score of the respective attribute, and *N* is the total number of examinees. This outcome can be used to describe differences between two attributes.

Distractor Scoring

In multiple-choice tests, examinees are required to select the correct response from a set of options. Options that do not contain the correct response (i.e., distractors) are created based on common misconceptions or errors that examinees are likely to encounter when solving the item. Until recently, it was considered that inferences about skill mastery could be made based only on the correct response on a test item, whereas an examinee's choice of a distractor was scored as non-mastery. For example, on a CDA of a skill with six attributes, the following is an example of how a multiple choice item would be scored:

Response Option	Associated Attribute Pattern
A	0,0,0,0,0,0
B*(correct)	1,1,1,1,0,0
C	0,0,0,0,0,0
D	0,0,0,0,0,0

* indicates the option selected by a student

Given B is the correct response for this item, two inferences can be made under this approach. First, if the examinee selected the correct response, then he/she has demonstrated mastery of the skill (Associated Attribute Pattern). Second, if other responses (i.e., any of the three distractors) were selected, then the student has not demonstrated any evidence of mastery. This approach is inefficient in the sense that information from distractors is not used in the scoring process, and a large number of test items are required to probe a small set of attributes because each attribute pattern needs to be probed by a set of items.

Although various scoring methods are available, distractors are rarely used in scoring because of the difficulty of incorporating them into scoring models (Luecht, 2007). A general approach to distractor scoring is through the use of item response theory (IRT), in which polytomous or graded latent response models can incorporate distractor information in the scoring process (Thissen et al, 1999). Luecht (2007) suggested the use of multiple scoring strategies to produce multiple scoring matrices in order to incorporate information from distractors. To implement this concept, Luecht (2007) suggested a set of Augmented Data matrices to be added in addition to a matrix of correct response used for scoring. For example, an augmented data matrix may include student responses to an often selected but incorrect option. In the present study, Luecht's (2007) approach is used with the CSS scoring method.

For distractors to be used in the scoring of diagnostic assessments, distractors must be developed diagnostically. Specifically, distractors have to represent mastery of a subset of attributes in the correct response. That is, it is

assumed that an examinee's response to a distractor indicates that the student has demonstrated mastery of a limited set of attributes (i.e., partial mastery). To incorporate this additional information in the scoring process, each distractor contributes to the CSS using its own associated attribute pattern. As a result, one additional inference can be made based on the examinee's response. Consider an examinee who answers incorrectly on the same item, for example, by selecting response option A.

Response Option	Associated Attribute Pattern
A*	1,1,0,0,0,0
B(correct)	1,1,1,1,0,0
C	1,0,0,0,0,0
D	0,0,0,0,0,0

* indicates the option selected by a student

Two inferences can be made. First, the examinee did not master the entire attribute pattern associated with the correct response (i.e., option B). Second, the examinee demonstrated mastery of attributes associated with one of the distractors, namely response option A. Conceptually, the overlapping attributes (the first two attributes in the Attribute Pattern) provide partial mastery evidence, and attributes probed by the correct response (i.e., option B) but not the distractor (i.e., attributes 3 and 4 in the Attribute Pattern) are considered as not mastered. From this approach, the length of patterns to be considered is no longer the number of items presented, but with a minimum of the item length and a maximum of twice the item length. Consequently, a conditional average (i.e., CSS) is needed to score a diagnostic assessment with distractors because both attribute patterns (i.e., options A and B) are used in scoring.

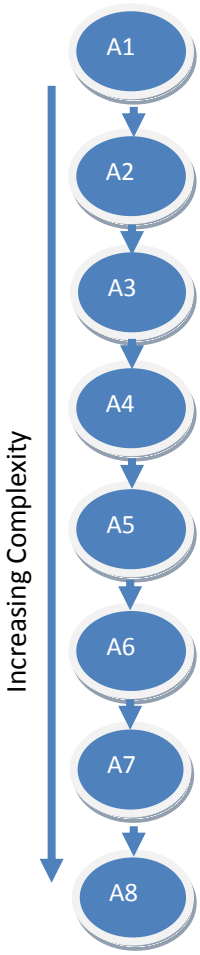
Method

Data

To demonstrate our CDA development framework, field test results from a provincial diagnostic assessment program were used. In total, 680 Grade 3 students participated in model-based diagnostic assessments for Mathematics. Within this program, a total of 48 items were administered to probe student mastery in two skills that are taught in classrooms as part of the Grade 3 Mathematics curriculum. To provide diagnostic information on student mastery, each of the two skills is further broken into hierarchies of attributes, with each skill described by a hierarchy of 8 attributes, organized in a linear pattern. Hierarchy A probed student mastery on place value representations (Figure 4), and hierarchy B probed student mastery on the ordering of numbers (Figure 5)**.** Each unique attribute combination is probed by three items, with a total of 24 items for each hierarchy. The hierarchies were developed by cognitive and subject matter experts and based on cognitive models of task performance (Gierl et al., 2007).

Hierarchy A

- Represent and describe numbers to 1,000, concretely, pictorially and symbolically.
- Illustrate, concretely and pictorially, the meaning of place value for numerals to 1,000.

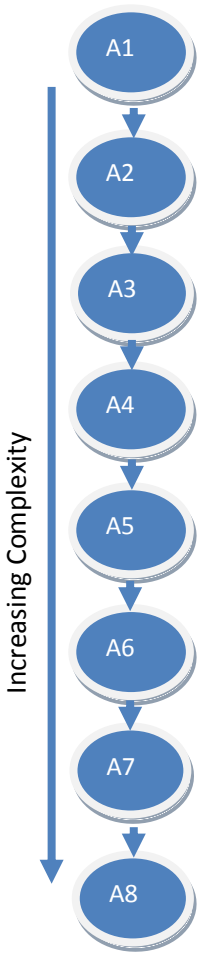


Attribute	Description
A1	Express number words in symbolic form using numbers 100 to 1,000
A2	Express the symbolic form of a number in words using numbers 100 to 1,000
A3	Represent a number concretely or identify a pictorial representation using numbers 100 to 1,000
A4	Identify the place-value meaning of a digit in a number using numbers 100 to 1,000
A5	Identify a number on a number line using numbers 100 to 1,000
A6	Identify a number that is a multiple of 10 or 100 greater than or less than a given number using numbers 100 to 1,000
A7	Identify a number when given place- value data using numbers 100 to 1,000
A8	Represent a number in more than one way using numbers 100 to 1,000

Figure 4. Hierarchy A – Place Value Representations

Hierarchy B

- Compare and order numbers to 1,000.



Attribute	Description
A1	Identify three missing numbers in a hundred chart using numbers 100 to 1,000
A2	Identify numbers on a number line using numbers 100 to 1,000
A3	Order numbers in ascending order using numbers 100 to 1,000
A4	Order numbers in descending order using numbers 1,000 to 100
A5	Correct an error in an ordered sequence using numbers 100 to 1,000
A6	Create 3-digit numbers from three numerals and order them in ascending order using numbers 100 to 1,000
A7	Create 3-digit numbers from three numerals and order them in descending order using numbers 1,000 to 100
A8	Verify the larger or smaller number of two numbers using place value concepts with numbers 100 to 1,000

Figure 5. Hierarchy B – Ordering of Numbers

Results

Results of the analyses carried out in the present study are organized in three parts. First, the descriptive results for the two diagnostic assessments are described. The results are summarized at the examinee, item, attribute, and hierarchy (i.e., test) levels. Second, we demonstrate how CSS results can be used in the model illustration and validation, using the mean absolute difference (MAD) of CSS. Third, the CSS results when distractor scoring is used are presented and compared with the results from the CSS without distractor scoring.

Descriptive Statistics of CSS

The results of diagnostic assessments are examined at four levels: examinee, item, attribute, and hierarchy or test. In total, 295 students participated in the diagnostic assessment for hierarchy A, and 385 students

participated in the diagnostic assessment for hierarchy B. As shown in Table 1, at the examinee level, student responses for the two hierarchies (A and B) follow a normal distribution, with the mean correct responses on hierarchy A and hierarchy B being 12.06 and 14.05, respectively.

Table 1. Diagnostic assessment results at the examinee level

	Hierarchy A	Hierarchy B
Mean	12.06	14.05
SD	5.62	4.72
Min	0	0
Max	24	24
N	295	385

The results at the item level are presented in Table 2. The percent correct for each item (i.e., p-values) indicated that, as expected, the test items that probed the attributes of higher complexity tended to have lower p-values than the items that probed the attributes of lower complexity.

Table 2. Diagnostic assessment results at the item level

		p-value	
Attribute	Item	Hierarchy A	Hierarchy B
A1	Item 1	0.824	0.820
	Item 2	0.753	0.870
	Item 3	0.610	0.747
A2	Item 4	0.631	0.698
	Item 5	0.363	0.589
	Item 6	0.668	0.620
A3	Item 7	0.590	0.758
	Item 8	0.722	0.646
	Item 9	0.597	0.716
A4	Item 10	0.481	0.802
	Item 11	0.566	0.760
	Item 12	0.512	0.820
A5	Item 13	0.488	0.599
	Item 14	0.393	0.372
	Item 15	0.495	0.635
A6	Item 16	0.559	0.354
	Item 17	0.393	0.456
	Item 18	0.319	0.417
A7	Item 19	0.444	0.378
	Item 20	0.380	0.438
	Item 21	0.458	0.500
A8	Item 22	0.237	0.435
	Item 23	0.231	0.375
	Item 24	0.353	0.286

At the attribute level, the CSS's were computed for each attribute of hierarchies A and B. The results are shown in Table 3, with the number of items probing each attribute shown in the last column on the right side. As described earlier, the CSS is the mean proportion of correct responses out of the total number of examinees' responses on the items used to probe each attribute. As seen in Table 3, the CSS values decrease with the increase in the attribute level. To corroborate the CSS results, the AHM results are also shown in Table 3. As mentioned earlier, the AHM is a probabilistic scoring method used, and indicates a probability of mastery for a student on a given attribute (i.e., attribute probability). Similar to the CSS values, the AHM values decrease with the increase in the attribute level, although in a non-linear way as compared to the linear nature of the CSS (see Figure 6).

Table 3. Diagnostic assessment results at the attribute level

	Hierarchy A		Hierarchy B		
	AHM	CSS	AHM	CSS	Items
A1	0.926	0.496	0.960	0.607	24
A2	0.924	0.466	0.959	0.577	21
A3	0.862	0.454	0.932	0.562	18
A4	0.804	0.418	0.882	0.529	15
A5	0.712	0.395	0.766	0.455	12
A6	0.559	0.378	0.587	0.421	9
A7	0.425	0.357	0.403	0.419	6
A8	0.154	0.274	0.304	0.375	3

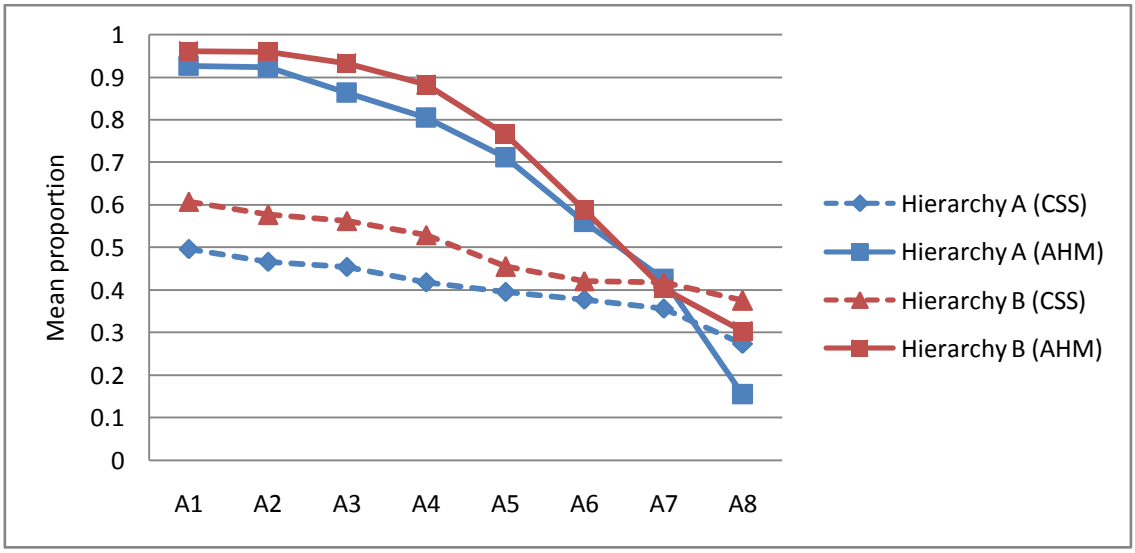


Figure 6. Diagnostic results at the attribute level

At the hierarchy level, different statistics or indices can be used to determine the overall fit of examinee responses with the design of the hierarchy.

A summary of results using two statistics at the hierarchy level are shown in Table 4. First, Chronbach's alpha, a coefficient often used to determine the internal consistency of items on a test, indicates that both assessments had high internal consistency (Cronbach's α >0.80). Second, the Hierarchy Consistency Index (HCI) indicates a fit between the observed response pattern and the expected response pattern (i.e., model-data fit), with larger HCI values indicating a better fit. In this study, the distributions of the HCI were determined to be bi-modal; therefore, medians were used to describe the central tendency of the HCI. The medians of the HCI indicated that, overall, hierarchy B tended to have a better model-data fit than hierarchy A. Next, percentages of examinees with an HCI value greater than 0 were computed to determine the percentage of examinees with the same patterns of observed responses as those expected for each hierarchy. With 59.5% and 78.4% examinees for hierarchy A and hierarchy B, respectively, this suggested that each hierarchy or the arrangement of the attributes used in the two assessments fit moderately well with the observed response patterns.

Table 4. Diagnostic assessment results at the hierarchy (test) level

	Hierarchy A	Hierarchy B
Median HCI	0.254	0.464
Examinee with HCI > 0	59.50%	78.40%
Cronbach's Alpha	0.863	0.801

Overall, the results for diagnostic scoring using the CSS method indicated an adequate model-data fit for the items representing the attributes and confirmed the order of the attributes in each hierarchy. Based on these results, the novel applications of the CSS method are demonstrated next.

Illustrating and Validating the Structure of Attributes

To demonstrate how the CSS method can be used to refine and validate the structure of attributes specified by test developers and content specialists, the mean absolute difference (MAD) is computed to determine the mean differences between any two attributes in the hierarchy. These values are then used to describe the distance or relatedness of attributes in terms of their complexity levels. The mean differences of the CSS's between any two attributes in hierarchy A and hierarchy B are shown in Tables 5 and 6, respectively. Distractor scoring was not used at this stage.

Table 5. ***Mean absolute differences (MAD) between two attributes in Hierarchy A***

	A2	A3	A4	A5	A6	A7	A8
A1	0.04	0.06	0.09	0.13	0.15	0.19	0.30
A2		0.03	0.06	0.10	0.13	0.17	0.28
A3			0.05	0.08	0.12	0.15	0.27
A4				0.05	0.09	0.13	0.24
A5					0.06	0.10	0.22
A6						0.09	0.20
A7							0.16

Table 6. ***Mean absolute differences (MAD) between two attributes in Hierarchy B***

	A2	A3	A4	A5	A6	A7	A8
A1	0.04	0.06	0.09	0.16	0.20	0.22	0.29
A2		0.04	0.07	0.13	0.17	0.19	0.27
A3			0.05	0.12	0.16	0.18	0.26
A4				0.08	0.13	0.15	0.23
A5					0.07	0.12	0.21
A6						0.09	0.18
A7							0.15

As shown in Tables 5 and 6, the mean absolute differences between any two attributes follow the expected linear pattern, namely the absolute CSS differences become larger as the level of attribute complexity increases. This is consistent with the structure of the hierarchy, the attributes in which are organized linearly. Next, the MAD values on the diagonal in Tables 5 and 6 provide a measure of differences that can be used to illustrate the distance between any two adjacent attributes. As the differences in the complexity among attributes become larger, the MAD values increase respectively, providing validation evidence for the attribute structure of both hierarchies.

Distractor Scoring

In order to incorporate distractors into the CSS scoring process, distractors need to be coded using partial mastery attribute patterns. In the present study, such coding was performed by two subject matter experts. Attribute patterns for each response option are shown in Table 7 for all the items for hierarchy B. As shown in Table 7, some distractors were not coded for any attribute mastery because these distractors did not elicit any skill related to the hierarchy.

Table 7. ***Attribute patterns for all the items with distractor scoring for Hierarchy B***

Item	Key	Options	Attribute Mastery							
			A1	A2	A3	A4	A5	A6	A7	A8
1	2	1	0	0	0	0	0	0	0	0
		2	1	0	0	0	0	0	0	0
		3	0	0	0	0	0	0	0	0
		4	0	0	0	0	0	0	0	0
2	3	1	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0
		3	1	0	0	0	0	0	0	0
		4	0	0	0	0	0	0	0	0
3	4	1	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0
		3	0	0	0	0	0	0	0	0
		4	1	0	0	0	0	0	0	0
13	2	1	1	1	1	0	0	0	0	0
		2	1	1	1	1	1	0	0	0
		3	1	1	0	0	0	0	0	0
		4	1	1	0	0	0	0	0	0
14	2	1	1	1	1	1	0	0	0	0
		2	1	1	1	1	1	0	0	0
		3	1	1	1	0	0	0	0	0
		4	1	1	1	0	0	0	0	0
15	3	1	1	1	1	0	0	0	0	0
		2	1	1	1	0	0	0	0	0
		3	1	1	1	1	1	0	0	0
		4	1	1	1	0	0	0	0	0

											16									
4	4	1	0	0	0	0	0	0	0	0		6	4	1	1	1	1	1	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	1	0	0	0	0
		3	0	0	0	0	0	0	0	0				3	1	1	1	0	0	0	0	0
		4	1	1	0	0	0	0	0	0				4	1	1	1	1	1	1	0	0
											17									
5	2	1	0	0	0	0	0	0	0	0		7	1	1	1	1	1	1	1	1	0	0
		2	1	1	0	0	0	0	0	0				2	1	1	1	1	0	0	0	0
		3	0	0	0	0	0	0	0	0				3	0	0	0	0	0	0	0	0
		4	0	0	0	0	0	0	0	0				4	0	0	0	0	0	0	0	0
											18									
6	3	1	0	0	0	0	0	0	0	0		8	4	1	1	1	1	0	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	0	0	0	0	0
		3	1	1	0	0	0	0	0	0				3	1	1	1	0	0	0	0	0
		4	0	0	0	0	0	0	0	0				4	1	1	1	1	1	1	0	0
											19									
7	1	1	1	1	1	0	0	0	0	0		9	4	1	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	0	0	0	0	0
		3	0	0	0	0	0	0	0	0				3	0	0	0	0	0	0	0	0
		4	0	0	0	0	0	0	0	0				4	1	1	1	1	1	1	1	0
											20									
8	3	1	0	0	0	0	0	0	0	0		0	3	1	1	1	1	0	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	0	0	0	0	0
		3	1	1	1	0	0	0	0	0				3	1	1	1	1	1	1	1	0
		4	0	0	0	0	0	0	0	0				4	0	0	0	0	0	0	0	0
											21									
9	4	1	0	0	0	0	0	0	0	0		1	2	1	1	1	1	1	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	1	1	1	1	0
		3	0	0	0	0	0	0	0	0				3	1	1	1	1	0	0	0	0
		4	1	1	1	0	0	0	0	0				4	0	0	0	0	0	0	0	0
1											22									
0	1	1	1	1	1	1	0	0	0	0		2	1	1	1	1	1	1	1	1	1	1
		2	0	0	0	0	0	0	0	0				2	1	1	1	1	0	0	0	0
		3	0	0	0	0	0	0	0	0				3	1	1	1	1	0	0	0	0
		4	0	0	0	0	0	0	0	0				4	1	1	1	1	0	0	0	0
1											23									
1	3	1	0	0	0	0	0	0	0	0		3	4	1	1	1	1	0	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	0	0	0	0	0
		3	1	1	1	1	0	0	0	0				3	1	1	1	0	0	0	0	0
		4	0	0	0	0	0	0	0	0				4	1	1	1	1	1	1	1	1
1											24									
2	1	1	1	1	1	1	0	0	0	0		4	2	1	1	1	1	0	0	0	0	0
		2	0	0	0	0	0	0	0	0				2	1	1	1	1	1	1	1	1
		3	0	0	0	0	0	0	0	0				3	1	1	1	0	0	0	0	0
		4	0	0	0	0	0	0	0	0				4	1	1	1	0	0	0	0	0

With response patterns added to the scoring process in the form of distractors to probe for partial mastery, there are more opportunities for examinees to demonstrate skill mastery. Given that two attribute patterns can be used per each item with distractor scoring, the total number of opportunities to demonstrate mastery of a given attribute across the entire test increases, and thus, contribute to the precision of the estimation of attribute mastery. Table 8

summarizes the number of opportunities for demonstrating attribute mastery for the two hierarchies, both when the CSS is used with and without distractor scoring. As shown in the table, using distractor patterns with the CSS method increases the number of opportunities for examinees to demonstrate mastery as compared to the CSS method when used without distractor scoring. However, attribute 8 in hierarchy A and attributes 5 through 8 in hierarchy B were not affected by distractor scoring because no partial mastery patterns were found to be associated with these attributes.

Table 8. ***Summary of the number of items representing each attribute***

	A1	A2	A3	A4	A5	A6	A7	A8
CSS								
Hierarchy A	24	21	18	15	12	9	6	3
Hierarchy B	24	21	18	15	12	9	6	3
CSS with Distractor Scoring								
Hierarchy A	54	51	45	42	17	14	11	3
Hierarchy B	54	51	46	24	12	9	6	3

The means and differences between the CSS method with and without distractor scoring (DS) are shown in Table 9. The mean proportions of CSS values with distractor scoring increased as a result of the increased number of opportunities for examinees to demonstrate attribute mastery when partial mastery patterns were used in scoring. As expected, there were no changes for the attributes for which partial mastery patterns were not used in scoring (i.e., A8 in hierarchy A and A5-A8 in hierarchy B).

Table 9. ***The CSS results (means and mean differences) with and without distractor scoring (DS)***

	Hierarchy A			Hierarchy B		
	CSS	CSS + DS	Difference	CSS	CSS + DS	Difference
A1	0.50	0.67	0.17	0.61	0.69	0.08
A2	0.47	0.61	0.14	0.58	0.67	0.09
A3	0.45	0.63	0.18	0.56	0.67	0.11
A4	0.42	0.58	0.17	0.53	0.59	0.06
A5	0.40	0.45	0.05	0.46	0.45	0.00
A6	0.38	0.49	0.11	0.42	0.42	0.00
A7	0.36	0.51	0.15	0.42	0.42	0.00
A8	0.27	0.27	0.00	0.38	0.38	0.00

Conclusion

In educational assessment, demands for diagnostic information from test results have prompted the development of model-based diagnostic assessment to inform teaching and learning. To determine student mastery of specific skills,

a number of scoring methods have been developed to score diagnostic assessments. However, current diagnostic scoring methods are at two extremes. On the one hand, probabilistic scoring methods are complex to be implemented in educational assessment programs, with results being difficult for teachers to understand and use in class. On the other hand, the method of subscore reporting provides little information about the level of skill mastery. Depending on the context and purposes of diagnostic assessments, the scoring approach presented in this paper – Complex Sum Scores (CSS) – can be a useful scoring solution, in particular when there is a shortage of psychometric resources required for implementing diagnostic assessment. If the purpose of a diagnostic assessment is to determine the level of an examinee's skill mastery based only on the evidence available from the test, or if there is a small number of students whose skill mastery is assessed by the test, then the CSS method can be a viable alternative to estimate skill mastery in such assessment programs.

However, several limitations associated with the CSS scoring method need to be acknowledged. First, given that the CSS method is a raw score approach to diagnostic scoring, the CSS scale can be problematic. Thus, some transformation of the raw score scale would be needed. Second, no comparisons of classification rates for the CSS and other diagnostic scoring methods have been provided in the present study. Although classification rates can be obtained using cut-score methods (Henson, Templin, & Douglas, 2007), a comparison of classification rates for the CSS and other methods was not a purpose of the present study. Rather, the purpose was to demonstrate the CSS method as a less complex alternative to current diagnostic scoring methods. Further, in the context of this study, a comparison of classification rates would involve the use of simulated data, and thus, make the accuracy of the results dependent on the simulation environment. However, since the CSS method is a non-probabilistic alternative to Gierl et al.'s (2007) neural network approach for diagnostic scoring and classification, recurrent neural networks can still be easily adapted with the CSS method to perform classification tasks.

References

Babenko, O., & Rogers, W. T. (2014). Comparison and properties of correlational and agreement methods for determining whether or not to report subtest scores. *International Journal of Learning, Teaching and Educational Research, 4*(1), 61-74.

Cui, Y., & Leighton, J. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46,* 429-449.

De la Torre, J., & Patz, R. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30,* 295-311.

Embretson, S. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective (pp. 107-136). New York: Plenum.

Fu, J., & Li, Y. (2007). An integrated review of cognitively diagnostic psychometric models. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement, 44,* 325–340.

Gierl, M., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46(3),* 293-313.

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl, (Eds.) Cognitive Diagnostic Assessment in Education: Theory and Practice. Cambridge University Press.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33(2),* 204–229.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation, Department of Statistics, University of Illinois, Urbana-Champaign.

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44,* 361-376.

Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3),* 258-272.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26,* 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule space approach. *Journal of Educational Measurement. 41(3),* 205-237.

Luecht, R. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting in J. Leighton and M. Gierl (Eds.), Cognitive diagnostic assessment for education: Theory and applications. New York, NY: Cambridge. pp. 319-340.

Luecht, R. (November, 2008). Assessment engineering in test design, development, assembly, and scoring. Keynote address at the East Coast Organization of Language Testers (ECOLT), Washington, DC.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64(4),* 575-603.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150-174.

Sinharay, S., Puhan, G., & Haberman, S. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45,* 553-573.

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61(2),* 287-307.

Wainer, H., Vevea, J., Camacho, F., Reeve, B., Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented scores - "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.) Test Scoring. Mahwah, NJ: LEA. pp. 343-387.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46(6),* 716-730.