Comparison of Attribute Coding Procedures for Retrofitting Cognitive Diagnostic Models

Amy Clark

Neal Kingston

University of Kansas

Corresponding Author: Amy Clark, 1122 West Campus Rd., Rm. 447, Lawrence, KS 66045, email: akclark@ku.edu

Abstract

This paper explores the process of coding items when retrofitting a cognitive diagnostic model to determine if the coding process impacts model-data fit. Three approaches for coding items for the cognitive attributes required to provide a correct response were compared. The coding approaches were implemented with three groups of coders: a group without additional training, a group receiving a training set, and a group of content experts reaching consensus. Inter-rater reliability was calculated for the first two groups, and the resulting Q matrices for the groups were compared to determine which approach achieved the best fit to the data using the Ox Metric software.

Keywords: attribute, coding, raters, diagnostic, assessment

Comparison of Attribute Coding Procedures for Retrofitting Cognitive Diagnostic Models

Arguably the most important aspect of retrofitting a cognitive diagnostic model to an assessment already in use is the correct coding of attributes to items on the test form. Without correct alignment between the attributes and the items, model misfit is likely to occur, leading to challenges in interpreting scores at the attribute level. Despite the importance of this step in the diagnostic classification procedure, few studies provide an in depth description of the processes raters go through when coding items for cognitive attributes. The current study seeks to compare three unique approaches to coding items for cognitive attributes, which are used to define the Q matrix for retrofitting a cognitive diagnostic model to data from a previously administered test form. Fit is compared across the three resulting Q matrices to determine how procedure used for coding attributes to items ultimately impacts model-data fit.

Literature Review

Cognitive Diagnostic Modeling

Cognitive diagnostic models have gained popularity in recent years as an assessment approach that can be used to diagnose skill mastery in a domain. By accounting for whether or not examinees have mastered various attributes, examinees can be diagnostically classified into knowledge states that exemplify the masetery/non-mastery of skills. Diagnostic score reports can be constructed using these knowledge states to inform teachers, students, and parents of the examinee's strengths and weaknesses.

Retrofitting models. When an assessment is already in use, rather than writing new items to assess the cognitive attributes included in the diagnostic model, the attributes

can be retrofit to the test form. During the process of retrofitting the model to the test form, coders examine items for evidence that an item does or does not require the attributes specified in the cognitive model. Since the items on the form were not originally written to assess these specific attributes, it is imperative that various raters code the attributes in a reliable fashion in order to ensure that each item is associated with the correct cognitive attributes.

Impact of coding on Q matrix specification. Many cognitive diagnostic models require the specification of a Q matrix, which provides an association between the items and the attributes they measure. The Q matrix is an item by attribute matrix, where a value of 1 indicates an item requires a particular attribute for a correct response, and a 0 indicates an item does not require an attribute to obtain a correct response. Correct specification of the Q matrix is essential for model-data fit. When a Q matrix is incorrectly specified, model-data fit is impacted, which can result in low classification rates (Svetina, Gorin, & Tatsuoka, 2011), poor discrimination between masters and non-masters, (DiBello, Roussos, & Stout, 2007), spuriously high or low expected scores (Liu, Douglas, & Henson, 2009), or inflated slipping and guessing parameters (Junker & Sijtsma, 2001). Thus, in order to ensure that cognitive diagnostic models correctly classify examinees into knowledge states, it is imperative that the Q matrix be correctly coded.

Attribute Coding for Cognitive Diagnostic Modeling

Despite the recent popularity of specifying cognitive diagnostic models for assessments already in use, the literature pertaining to retrofitting models contains relatively little description of the process actually used by raters to code items for cognitive attributes. As a result, a consensus on the procedure for coding items has not been

developed, and the majority of retrofitted cognitive diagnostic modeling studies each follow a unique coding approach.

One way in which the coding approach differs by study is in the selection of coders. In many applications of retrofitting diagnostic models to existing data, the authors of the study are included as coders (e.g. Buck, Tatsuoka, & Kostin, 1997; Gierl et al., 2009). While outside raters have been recommended to avoid bias in the coding, one author in particular specified that due to funding issues outside coders could not be included (Buck & Tatsuoka, 1998). In contrast, other applications of retrofitting diagnostic modeling have employed the use of content experts for coding the items for attributes (e.g. von Davier, 2008; Wang, Gierl, & Leighton, 2006). Still others recruited graduate students to assist with the coding process (e.g. Jang, 2005; Wang & Gierl, 2011). Regardless of the approach taken for selecting coders, it is imperative that coders accurately assign attributes to items in order to ensure accuracy of the Q matrix.

An additional area that differs across studies is the number of coders that assign attribute to items. The most common number of raters is two (e.g. Birenbaum & Tatsuoka, 1993; Buck et al., 1998). One study in particular made use of five coders (e.g. Jang, 2005). In some instances, only a single rater is used to assign codes (e.g. Buck & Tatsuoka, 1998; Leighton, Cui, & Cor, 2009). However, using a single coder could be problematic; because there is no evidence of inter-rater reliability or consensus among coders, attribute codes are confounded with the sole coder's level of consistency in providing ratings. For this reason, it is generally recommended to have more than one coder assign attribute codes to items.

When multiple coders are used to assign attribute codes, the number of items coded by each coder sometimes differs. In some examples, each coder assigns attribute codes to all items (e.g. VanderVeen et al., 2007; Wang & Gierl, 2011). Although time consuming, this approach allows for the greatest potential inter-rater reliability because no items are excluded. In contrast, other studies make use of a single coder to code all items, while a second coder reviewed the codes or independently coded a subset of the items (e.g. Svetina et al., 2011). While there are benefits to having a second rater code only a subset of items or simply review the codes, including requiring less time and monetary resources, this approach may again impact the reliability of the codes assigned to items. Much like the situation in which only a single coder was implemented, the use of a second rater only coding a subset of the items could drastically impact the codes assigned to each item. In such a situation it becomes essential that the first coder assign codes to the items in a consistent manner throughout the coding process.

When more than one rater is used to code items for cognitive attributes, indices of rater agreement are often calculated between the coders. Typically the raters code the attributes independently without discussing assignments, and the level of agreement is calculated after all items have been coded with their requisite attributes. The most commonly used metric when codes are dichotomously assigned is percent agreement (e.g. Birenbaum & Tatsuoka, 1993; Buck et al., 1998). When attributes are coded continuously, such as for count variables, Pearson correlation values may be used to provide an estimate of rater agreement (Buck & Tatsuoka, 1998). Other measures of rater agreement include Cohen's kappa for two raters, and Fleiss's kappa and intraclass correlation for groups of

raters. Each of these indices provides an estimate of the level of agreement between raters and can in some cases inform the final selection of items or attributes.

In contrast, rather than estimate levels of agreement, the goal of many studies is to attain a consensus regarding the coding of attributes to items. This consensus then forms the final Q matrix used to fit the diagnostic model to the data. In such cases, raters may independently code the items for attributes, but meet to discuss codes and reach an agreement for any items with discrepancies (e.g. Gierl et al., 2009; VanderVeen et al., 2007). While levels of agreement or interrater reliability are not reported, this approach allows for a single Q matrix to be used that retains all items and attributes, as disagreement is resolved prior to its construction.

A final way coding procedures often differs for retrofitted models is with regard to the extent of the instructions provided to the coder(s). Coders may first meet to code a subset of items together in order to establish a common agreement regarding the application of attributes to the items (e.g. Buck et al., 1997; Wang & Gierl, 2011). In contrast, coders may simply be provided with a set of instructions without discussing the attributes or reaching an initial consensus regarding their application (e.g. Jang, 2005). The depth of instructions may also differ from including a simple instruction to code all present attributes to including a detailed set of coding instructions that contains examples and explanations of the attributes (e.g. Buck et al., 1997; Svetina et al., 2011). Despite these differences in procedure, the effects of the level of instruction or training provided has not been analyzed to determine what differences, if any, exist in rater agreement between these approaches and how differences subsequently impact model-data fit.

Coding in Other Domains

Coding procedures are not limited to cognitive diagnostic modeling for assessments. Coding is also common in various educational research domains as well as in research that makes use of qualitative analyses. Since there has been relatively little research into the impact of coding procedures on classification using cognitive diagnostic models, the research in these areas was consulted as a means of further guiding the current study.

One approach commonly used in domains outside of educational assessment is the use of a codebook. The codebook is created through an iterative process to encompass all the codes included in the study. After an initial set of items is coded, the coders meet to discuss the codes and make revisions to the codebook. This approach may or may not be useful when assigning attribute codes to items for a cognitive diagnostic modeling study. In the case where a well-researched set of attributes is being applied to the items, the researchers may not want to revise the attributes simply based on the results of the first set of codes assigned to the item. Perhaps instead of a full revision to the attributes, examples or explanations could be included in the codebook to provide a more precise definition of the attribute without modifying its content.

In addition to implementing an iterative codebook, the coding process in domains outside of diagnostic classification made use of many similar procedures. Multiple coders are included and estimates of inter-rater reliability are calculated using Cohen's Kappa or Fleiss's kappa, depending on the number of coders included. Benchmarks for Kappa values, as identified by Landis & Koch (1977b) are included in Table 1. Similarly, coding in outside areas also makes use of training sets and examples (e.g. Larsson, 1993; Novak, Hoffman, & Duhachek, 2003). These findings provide evidence that the coding procedures used in

areas beyond diagnostic classification models may prove helpful in establishing a procedure to maximize inter-rater reliability when coding attributes to items.

Table 1

Benchmark Kappa values

Kappa Value	Strength of Agreement
< 0.0	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Method

Materials

Prior to beginning the coding process, the researchers prepared a set of materials. Among the materials were three forms of a reading comprehension assessment. The forms each included 35 passage-based reading comprehension items from a large-scale assessment administered annually to high school students. As only the passage-based items were included, item numbers spanned from 9-24, and 30-48. In addition, a list of cognitive attributes was prepared. The attributes included 11 skills that were found to underlie passage-based reading comprehension items by Wang and Gierl (2011). Table 1 includes a summary of the cognitive attributes. The researchers also prepared an item by attribute coding sheet using Excel to record the attributes required for a correct response to each item.

Table 1

Summary of Cognitive Attributes

A1	Basic language knowledge, such as word recognition and basic grammar
A2	Determining word meaning by referring to context
A3a	Literal understanding of sentences with minimal amount of inference
A3b	Understanding sentences by making inferences based on the reader's experience and background knowledge
A4a	Literal understanding of larger sections of text with minimal amount of inference
A4b	Understanding larger sections of text by making inferences based on the reader's experience and world knowledge; building coherence across, summarizing, and evaluating larger sections of text
A5	Analyzing author's purposes, goals, and strategies
A6	Understanding text with difficult vocabulary
A7	Understanding text with complex syntactic structure
A8	Using rhetorical knowledge
A9	Evaluating response options

Procedure

Based on the previous literature, three distinct coding approaches were incorporated in the current study. Each coding approach was selected to follow a unique procedure in order to determine if differences in inter-rater reliability and model-data fit would be observed. A total of nine coders were recruited for the current study. All coding took place in two distinct stages.

Stage one. During the first stage, two coding groups were created using six of the nine coders. The six coders included in the first stage were graduate students working at a university-based educational testing company in the Midwest. Each of the six coders was randomly assigned to one of two coding groups.

Of the two coding groups, one was selected to code items for the requisite attributes based on the attributes alone. Prior to beginning the coding process, these raters were provided with a brief set of instructions, the list of the 11 cognitive attributes, copies of the three test forms, and an electronic copy of the Excel coding sheet. The brief set of instructions highlighted that the coder was to work independently to code each item for the attributes an examinee would need to have mastered in order to obtain a correct response to the item. The list of attributes they were provided with only included the attribute code and the description, with no additional information on any of the attributes. The raters were to use the coding sheet to enter their codes, using a 0 to indicate if an item did not require an attribute for mastery, and a 1 to indicate if an item did require an attribute for mastery.

The second coding group began the coding process by holding a meeting to discuss procedures. A codebook was provided to each of the coders that included the 11 cognitive attributes, the description, along with an expanded explanation of each. During the meeting, the group first reviewed the cognitive attributes and the corresponding explanations, and the group discussed the meaning of any areas that were unclear. Next, the group practiced coding a training set of five items using the attributes. Each coder independently coded the items, and then the group discussed the codes. All areas of disagreement were discussed, and group members explained divergent thinking. Explanations in the codebook were revisited to clarify the meaning of the attributes. Following the meeting, each coder independently coded the items using the codebook, and entered values of 0 and 1 in their electronic copy of the Excel coding sheet.

After each of the coders in groups one and two completed their independent coding using the Excel coding sheet, they submitted their codes to the researchers. Two measures of inter-rater reliability were calculated within each group of raters. These measures included Fleiss's kappa and intraclass correlations, both of which are designed to assess the level of agreement within a group when ratings are provided by groups rather than pairs of raters. Following the calculation of interrater reliability indices, a Q matrix was constructed for each training group. As both training groups included an odd number of raters, the Q matrix included codes agreed on by at least two of the three raters. In addition, the item by attribute codes were synthesized across groups to create a summary sheet that included counts of the number of raters that coded a 1 for each attribute. For example, if all 6 raters coded a 1 for an item requiring an attribute, then a 6 was placed in that location of the matrix.

Stage two. The final coding group consisted of three content experts. All three expert raters were full time staff specializing in K-12 assessment at a university-based educational testing company in the Midwest. Rather than code items individually and assess inter-rater reliability, the expert coding group convened for a series of meetings to reach consensus on the attributes required by each item. The expert coders were provided with the codebook that included attribute descriptions and explanations, as well as the summary sheet that included the ratings from the two groups from stage one. During the meetings, each item was independently reviewed and the group of content experts reached a consensus as to the attributes required to provide a correct response. One of the researchers facilitated the meetings and recorded the expert consensus in the coding sheet, but the researcher did not participate in the process of determining required attributes for

each item in order to avoid the introduction of any potential bias on the part of the researcher.

Following the completion of coding in stage two, a Q matrix was constructed to reflect the codes agreed upon by the expert raters. The entries in this Q matrix were compared with the entries of the two Q matrices created for groups one and two in the first stage of the research using Cohen's kappa.

To determine which of the three Q matrices provided the best model-data fit, each of the Q matrices was used to create a diagnostic model of reading comprehension using the generalized DINA model. The Q matrices were each fit to a random sample of 2000 examinees' responses to the 35 items using the Ox Metric software (de la Torre, Chiu, & Chen, 2012). Output included fit statistics as well as recommendations for improvements to the Q matrix based on the student response patterns evident in the data.

Results

Within each group in stage one, pairwise comparisons were made between each of the raters to determine the level of consistency with which they coded the items for attributes. For the group of raters that did not receive the training set, percent agreement across the three forms was around 64% for each pair of raters. Across the three forms Cohen's kappa values between the pairs of raters were all approximately 0.2. In contrast, for the group of raters that received the training set pairwise agreement was slightly higher. Across the three forms, the pairwise percent agreement between each of the raters was around 70%. Cohen's kappa values between the pairs were all approximately 0.4. The group that did not receive the training set coded attributes A1 and A8 with the highest level of agreement, while the group that received the training set coded attributes A1 and A9

with the highest level of agreement. Form B had the highest level of consistency between the pairs of raters in both groups.

When examining ratings for the entire group, rather than pairwise comparisons, the non-training set group had overall lower inter-rater reliability. The average Fleiss's kappa value for forms A, B, and C were 0.1, 0.2, and 0.2, respectively. Ideal values for Fleiss's kappa are those values ≥ 0.6 (Landis & Koch, 1977a). A total of two, three, and one items were identified as having interrater agreement values that were at least 0.6 for forms A, B, and C, respectively. Similarly, average intraclass correlation values for forms A, B, and C were 0.2, 0.4, and 0.4, respectively. Ideal values for intra-class correlations are those values ≥ 0.7. A total of 7, 11, and 8 items were identified as having interrater agreement values that were at least 0.7 for forms A, B, and C, respectively. These values suggest overall, the non-training set group experienced low levels of coding agreement when codes were assigned independently without the opportunity to practice coding with a training set.

In contrast, the group of three raters that received practice with the training set, or the training set group, had moderate overall inter-rater reliability. The average Fleiss's kappa value for forms A, B, and C were 0.2, 0.5, and 0.4, respectively. A total of 0, 12, and 5 items were identified as having interrater agreement values that were at least 0.6 for forms A, B, and C respectively, indicative of meeting the threshold for ideal interrater reliability. Similarly, average intraclass correlation values for forms A, B, and C were 0.4, 0.7, and 0.6, respectively. A total of 12, 24, and 21 items were identified as having interrater agreement values that were at least 0.7. for forms A, B, and C, respectively, indicative of meeting the threshold for ideal interrater agreement. These values suggest overall, the training set group experienced slightly higher levels of coding agreement when codes were assigned

independently, but values were still generally below the ideal values for both Fleiss's kappa and intraclass correlation.

Tables 2 and 3 include Fleiss's kappa values for both of the training groups in stage one. Tables 4 and 5 include the intraclass correlation values for both of the training groups in stage one. These tables show that overall the raters that received the training set had higher agreement across items than the raters that did not receive the training set.

When comparing the codes from the stage one raters to the codes provided by the expert raters, the group that received the training set of items had closer alignment. Cohen's kappa values across the rating groups for each of the three forms are presented in Table 6.

Overall, the Q matrix based on the codes obtained from the expert raters had the best fit to the data for each of the three forms. Table 7 includes the fit values for each of the three rating groups. A smaller value indicates better model-data fit.

	Form A			Form B			Form C		
	Expert	Training Set	No Training Set	Expert	Training Set	No Training Set	Expert	Training Set	No Training Set
-LL2	56738	55369	55563	57383	55167	56243	56620	54135	55120
AIC	62192	67031	61593	63533	70685	61801	61866	63205	60694
BIC	76681	98013	77613	79871	111910	76567	75803	87301	75502

Table 2

No training set Fleiss's Kappa

Items:	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	30	31
Form A	0.0	0.2	0.1	-0.1	0.3	0.0	0.0	0.1	0.4	0.3	0.1	0.0	0.9	0.3	0.4	0.0	0.2	0.0
Form B	0.0	-0.2	0.0	0.2	0.0	0.5	0.3	0.1	0.0	0.4	0.0	0.3	0.0	0.6	0.0	0.0	0.3	0.3
Form C	0.0	0.2	-0.1	-0.1	0.3	0.2	0.3	0.7	0.4	0.5	0.1	0.1	0.1	0.0	0.3	-0.3	0.2	0.4

Items:	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Form A	0.2	-0.1	0.3	-0.1	0.2	-0.2	0.2	0.6	-0.3	-0.1	0.5	0.1	0.5	0.0	-0.1	0.0	-0.2
Form B	0.3	0.3	0.2	0.3	0.5	0.4	0.0	0.1	0.1	0.3	0.4	0.3	0.6	0.7	0.2	-0.2	0.1
Form C	0.5	0.3	-0.1	0.0	0.1	-0.1	0.2	0.0	0.1	0.5	0.0	-0.2	0.0	0.4	0.2	0.2	0.0

Table 3

Training set Fleiss's Kappa

Items:	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	30	31
Form A	0.4	0.0	0.1	-0.2	0.0	0.1	0.0	0.0	0.3	0.1	0.1	0.8	-0.1	0.1	0.2	-0.1	-0.1	0.3
Form B	-0.1	0.3	-0.1	0.5	0.8	0.4	0.3	0.7	0.3	0.1	0.5	0.3	0.6	0.4	0.4	0.3	0.2	0.6
Form C	0.8	0.0	0.2	-0.1	0.1	0.2	0.4	0.4	0.2	0.1	0.4	0.8	0.5	0.2	0.4	0.0	0.0	0.9

Items:	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Form A	0.5	0.5	0.5	0.4	0.4	0.4	0.4	0.3	0.5	0.3	0.4	0.2	0.2	0.2	0.4	0.3	0.0
Form B	0.6	0.5	0.8	0.8	0.6	0.6	0.4	0.6	0.9	0.5	0.5	0.4	0.5	0.3	0.3	0.3	0.6
Form C	0.6	0.3	0.4	0.5	0.3	0.5	0.5	0.3	0.5	0.6	0.5	0.4	0.5	0.2	0.5	0.5	0.3

Table 4

No training set intraclass correlation

Items:	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	30	31
Form A	0.1	0.5	0.4	0.1	0.6	0.2	0.3	0.5	0.7	0.0	0.4	0.1	1.0	0.6	0.7	0.2	0.5	0.2
Form B	0.0	-0.2	0.0	0.6	0.3	0.8	0.6	0.4	-0.1	0.7	0.3	0.6	0.3	0.8	0.2	0.3	0.7	0.6
Form C	0.1	0.4	0.1	-0.1	0.6	0.5	0.6	0.9	0.7	0.8	0.4	0.3	0.4	0.3	0.5	-1.2	0.5	0.7
Items:	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
Form A	0.4	-0.3	0.7	-0.2	0.4	-0.7	0.6	0.8	-2.2	-0.1	0.8	0.4	0.7	0.3	-0.1	0.2	-0.4	
Form B	0.6	0.7	0.5	0.7	0.8	0.7	0.2	0.3	0.4	0.6	0.7	0.6	0.8	0.9	0.5	-1.3	0.3	
Form C	0.8	0.7	0.2	0.3	0.5	0.2	0.5	0.3	0.5	0.8	0.4	-0.3	0.3	0.7	0.6	0.6	0.3	

Table 5

Training set intraclass correlation

Items:	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	30	31
Form A	0.7	0.1	0.4	-0.3	0.2	0.3	0.4	0.3	0.6	0.4	0.5	0.9	-0.6	0.3	0.5	-0.1	-0.5	0.6
Form B	-0.3	0.6	-0.4	0.8	0.9	0.7	0.6	0.9	0.6	0.5	0.8	0.6	0.8	0.7	0.7	0.6	0.5	0.8
Form C	0.9	0.1	0.5	0.0	0.3	0.5	0.7	0.7	0.6	0.3	0.7	0.9	0.8	0.5	0.7	0.1	0.3	1.0
Items:	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
Form A	0.8	0.8	0.8	0.7	0.7	0.7	0.7	0.7	0.8	0.6	0.7	0.5	0.4	0.5	0.7	0.5	0.1	
Form B	0.9	0.8	0.9	0.9	0.9	0.8	0.7	0.9	1.0	0.8	0.8	0.7	0.8	0.6	0.6	0.7	0.9	
Form C	0.8	0.6	0.7	0.8	0.6	0.8	0.8	0.7	0.8	0.8	0.8	0.7	0.8	0.5	0.8	0.8	0.6	

Discussion

Upon comparison of the two coding approaches from stage one, the group of three raters that received the practice training set appeared to code the items with greater consistency. Across forms, the group that received the training set achieved Kappa values that averaged .10 larger and intraclass correlation values that averaged .30 larger. These findings suggest that providing raters with an opportunity to practice the coding process as a group may increase the likelihood that the raters agree overall.

Despite having generally high percent accuracy, particularly for the raters that received the training set, the majority of Kappa values fell below .60, which was the suggested threshold for inter-rater agreement. Since there were only two codes possible, 0 and 1, the likelihood that the pairs of raters would agree by chance alone was quite high. Similarly, for attributes that most items assessed, coders were more likely to code 1 over 0, inflating the likelihood of agreement by chance. These aspects of the coding process impacted the magnitude of the Kappa values.

One aspect that likely allowed the coding group that received the training set to code more similarly was the ability to discuss their perceptions regarding the attributes. During the training, the raters discussed the meaning of A1 and reached the conclusion that all items would require basic language knowledge. This led to complete accuracy in the coding of A1. This is in stark contrast to the group that did not receive the training set, where two sets of raters had a percent agreement of only 25%.

Similarly, the group that received the training set initially had some confusion over when to code A9. After practicing with the training set, the group concluded that in instances where the answer could be arrived at without consulting the options, the item would be coded 0. In instances where the test taker had to read all options prior to selecting an answer, the item would be coded 1. This discussion likely led to the much higher percent agreement for A9 for the group receiving the training set than the group that did not.

The results of this study indicate that when coding items for attributes, conducting an instructional meeting to providing the coders with additional explanations of the attributes and to practice coding with a training set helps ensure greater percent agreement and inter-rater reliability among the coders. Attributes that are the most discussed within the group and a consensus is reached regarding when to code an item for that attribute may increase the level of agreement between raters as well.

Several suggestions were developed for the hierarchy as a result of this study. Consistent with the intent of Wang and Gierl, 2011, the expert raters coded A1 as a required skill for all items. In addition, the expert raters also coded A3a as a required skill for all items. During the coding process, it was determined that all items required a literal understanding of the content, form, and function of sentences with minimal amount of inference. This finding may serve to inform future iterations of the hierarchy. The content experts also discussed at length what constituted “larger sections of text” during the coding process. The experts concluded that in this study, larger sections referred to those instances where

students had to move within the text or read the entire passage, as opposed to those instances where the correct answer could be determined from a single small selection of sentences or even a paragraph. Additionally, the content experts determined that A9 would be coded in those instances where the correct answer could be provided without consulting the options. When the student was required to discern between the answer choices in order to provide a correct response, for example, in instances where the stem elicited students to select "the best" response option. Another distinction made by the content experts in coding the items pertained to A5. The experts encountered several items that specifically asked the students to gauge the author's purposes. However, upon closer inspection, some of these items were actually just asking for the main idea. Thus, the experts determined that not all items that explicitly ask for the identification of "author's purpose" are actually tapping into that particular skill. In future iterations of the hierarchy, raters will want to be careful to determine what is actually being measured by the item as opposed to what the item intended to be measure.

Another interesting discussion that came from the expert coding sessions was the idea of multiple pathways or variant hierarchies. The content experts pointed out multiple instances where a student could have correctly responded using one of two possible routes, each requiring a different combination of attributes. In these instances, the experts decided to specify the collection of attributes required for the lowest ability student to correctly respond to the item. However, this approach could potentially cause the model to fit less well for some students and thereby increase the number of "slips" and "guesses" that occur when

the model does not adequately fit to the student's response pattern. Future research on retrofitted cognitive diagnostic models will want to further explore this idea of multiple pathways or Q matrices that provide better fit for examinees of different ability levels or skill sets.

The coding findings presented here also have implications for cognitive diagnostic assessments, where items are written to assess specific attributes. As shown here, the nine individuals examining items did not all have complete agreement on the skills being assessed. When items are written to assess specific attributes, it is imperative that item writers thoroughly understand the construct. Furthermore, items should all be reviewed externally prior to being included on an assessment and additional analyses should be conducted to evaluate whether the items appear to truly assess the attribute.

Despite the time and monetary constraints associated with using multiple raters during the coding process and finding time to collaborate in the coding process, the present findings suggest that these strategies may be associated with better model-data fit and inter-rater agreement. Additional studies should be conducted to confirm this finding and evaluate additional conditions by which items might be coded for a retrofitted model.

References

Birenbaum, M., & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students' knowledge states in multiplication and division with exponents. *Applied Measurement in Education, 6*(4), 255-268. doi: 10.1207/s15324818ame0604_1

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*, 423-466. doi: 10.1111/0023-8333.00016

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Buck, G., VanEssen, T., Tatsuoka, K. K., Kostin, I., Lutz, D., & Phelps, M. (1998). Development, selection, and validation of a set of cognitive and linguistic attributes for the SAT I verbal, sentence completion section. Princeton, NJ: Educational Testing Service.

de la Torre, J., Chiu, C-Y., Chen, J. (2012, April). *Cognitive diagnosis modeling: A general framework approach.* Training session provided at the annual meeting of the National Council on Measurement in Education, Vancouver.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 979-1030). Amsterdam: Elsevier.

Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2009). Validating cognitive models of task performance in algebra on the SAT®. New York: The College Board.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* (Ph.D. 3182288), University of Illinois at Urbana-Champaign, United States -- Illinois. ProQuest Dissertations & Theses (PQDT) database.

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.

Landis, J. R., & Koch, G. G. (1977a). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Landis, J. R., & Koch, G. G. (1977b). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Larsson, R. (1993). Case survey methodology: Quantitative analysis of patterns across case studies. *Academy of Management Journal*, 1515-1546.

Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchy consistency index. *Applied Measurement in Education, 22*(3), 229-254. doi: 10.1080/08957340902984018

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*(8), 579-598.

Novak, T. P., Hoffman, D. L., & Duhachek, A. (2003). The influence of goal-directed and experiential activities on online flow experiences. *Journal of Consumer Psychology, 13*(1), 3-16.

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing, 11*(1), 1-23. doi: 10.1080/15305058.2010.518261

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT Reasoning Test™. In D. S. McNamara (Ed.), *Reading comprehension strategies* (pp. 137-171). Mahwah, NJ: Erlbaum.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307. doi: 10.1348/000711007x193957

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*(2), 165-187. doi: 10.1111/j.1745-3984.2011.00142.x

Wang, C., Gierl, M. J., & Leighton, J. P. (2006). *Investigating the Cognitive Attributes Underlying Student Performance on a Foreign Language Reading Test: An Application of the Attribute Hierarchy Method.* Paper presented at the annual

meeting of the National Council on Measurement in Education, San Francisco, California.