Design Principles Required for Skills-Based Calibrated Item Generation

Hollis Lai

Mark J. Gierl

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Submitted to:

Dr. Krista Breithaupt

Director, Research and Development

Medical Council of Canada

March 31, 2012

INTRODUCTION

Test development is a costly endeavor. Rudner (2010) estimated that the cost for a single item on an operational licensure exam is from $1500 to $2000. Large numbers of items are needed to create the banks necessary for minimizing item exposure, enhancing test fairness, and supporting testing on-demand (Breithhaupt, Hare, & Ariel, 2010). As a result, testing programs are constantly developing new items. An alternative way to meet this demand for creating large numbers of new items is with automatic item generation (AIG). AIG can be described as the process of using models to generate statistically calibrated items with the aid of computer technology (Gierl & Haladyna, in press). AIG, an idea introduced more than 40 years ago (Bormuth, 1969), is gaining renewed interest and support. Significant developments in both AIG theory (i.e., Gierl & Lai, in press) and practice (e.g., Wendt, Kao, Gorham, & Woo, 2009; Masters, 2010) have recently been documented.

While AIG provides a solution for generating large numbers of new items, the quality and, hence, psychometric properties of these items must still be evaluated. Typically, item quality is determined through a field testing process where items are administered to examinees so their psychometric characteristics can be evaluated. But this typical solution—field test and item analysis—is not feasible when thousands of new items have been generated. An alternative method for estimating the psychometric properties of the generated items is with statistical models that permit item precalibration. With this approach, the psychometric properties of the items can be estimated during the item generation process. We refer to this amalgamated process as calibrated item generation (CIG).

TWO GENERAL METHODS FOR CALIBRATED ITEM GENERATION

CIG can be conceptualized as two different but complementary approaches. The first is family-based CIG where item clones (i.e., items that are very similar to one another) are calibrated with a parent item. For family-based CIG, Sinharay, Johnson, and Williamson (2003), and Sinharay and Johnson (2008) developed a statistical model where items with the same or similar features could be calibrated

as a family. Their approach, called the related siblings model, provides a method for estimating the statistical properties of siblings within an item family. Glas and van der Linden (2006) also demonstrated how a family-based CIG method could be operationalized by creating item clones and then estimating their psychometric properties using a hierarchical IRT model. In short, the family-based CIG approaches rely on the concept of a random item (de Boeck, 2008), where a generated assessment task is sampled from the parent item population, and the statistical properties of the generated items are estimated from the characteristics of the parent item.

The second general approach is skills-based CIG where generated items are calibrated using an analysis of the set of skills required to solve the item. In contrast to family-based CIG, generated items are calibrated based on their prerequisite skills. Embretson (1998) first demonstrated how items can be calibrated under a cognitive model of skills for abstract reasoning. More recently, Arendasy, Sommer, and Gittler (2010) generated items to assess spatial reasoning skills such as mental rotation and shape manipulation. Zeuch (2010) also demonstrated how a skills-based precalibration process could be used for latin-square tasks and statistical word problems. With skills-based CIG, the prerequisite skills required to solve a task are defined using an item-by-skill matrix called the Q-matrix. Items that measure these skills are either identified on an existing test or created specifically to measure a unique combination of skills and then calibrated to determine the psychometric properties of the items.

Family- and skills-based CIG differ in the criteria used to group items expected to elicit similar psychometric characteristics. With family-based CIG, generated items from the same model are assumed to elicit similar psychometric properties. This assumption limits the generative outcomes from an item model because the items are expected to probe the same set of content and skills. To maintain the alignment between the parent and siblings, the generated items typically need to be quite similar to each other. With skills-based CIG, generated items adhere to a prescribed set of skills as defined in the Q-matrix. But when items are generated, these skills can be organized in many different ways. The

generated items, therefore, need not be overtly similar to one another because different generated items can be developed to measure different skill sets. The focus of our paper is on skills-based CIG. Specifically, we focus on the use of the linear logistic test model with random effects (LLTM-RE) for estimating the difficulty level of generated items.

OVERVIEW OF LINEAR LOGISTIC TEST MODEL WITH RANDOM EFFECTS

Janssen (2010; see also Janssen, Schepers, & Peres, 2004) introduced a random item effects model for calibrating assessment tasks based on the prerequisite skills required to solve the item. The model serves as a modification of the linear logistic test model (LLTM), introduced by Fischer (1973), where the effect of each skill on the difficulty level of an item is first estimated, and then the effects are added together to provide an item difficulty estimate. That is, the probability that person j successfully answers item i is given by the LLTM as follows

$$P\left(X_{ij}=1\middle|\theta_j,\boldsymbol{q},\boldsymbol{\eta}\right)=\frac{\exp(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k)}{1+\exp(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k)}.$$

In this formula, the item difficulty parameter β_i for item i, typically found using the 1PL IRT model, is composed of a linear combination of skill predictors specified by $\beta_i=\sum_{k=1}^{K}q_{ik}\eta_k$, where the use of skill k for item i described by q_{ik} is multiplied by η_k, the difficulty weight corresponded to skill k, and then summed across all K skills.

One drawback of the LLTM is that it lacks a parameter to estimate the variability of the items within a set of skills (i.e., all items requiring the same skills are assumed to have the same difficulty estimate with the LLTM). To address this limitation, Janssen developed a model for estimating the variability within each set of skills by adding a random item effect term that could be used to account for item variation within a skill set. To estimate item difficulty of item i, an error term ε_i is added to every item such that

$$\beta_i=\sum_{k=1}^{K}q_{ik}\eta_k+\ \varepsilon_i=\beta_i^*+\ \varepsilon_i, \text{ where } \beta_i\sim N(\beta_i^*,\sigma_\epsilon^2).$$

De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx, and Partchev (2011) recently presented an estimation method using the LME4 package (Bates, Maechler, & Bolker, 2012) in R (R Core Development Team, 2012) for calibrating the LLTM-RE.

PURPOSE OF CURRENT STUDY

One important application of the LLTM-RE is in CIG. Janssen (2010) claimed that with a random item effect, it is now possible to use the LLTM-RE for skills-based CIG. She did not, however, demonstrate how her model could be used for CIG. Hence, the purpose of our study is to present two applications of skills-based CIG using data from operational testing programs to demonstrate how the LLTM-RE can be used. The first practical application focuses on generating items for a high-stake medical licensure examination. The second practical application focuses on generating items for a cognitive diagnostic assessment in mathematics. These two practical applications were selected to demonstrate, first, how real data could be used for CIG and, second, to highlight the importance of test design principles on the successful application of the LLTM-RE for CIG.

PRACTICAL APPLICATION #1: MEDICAL LICENSURE EXAMINATION

Test Design Principles

The first practical application featured in our study focuses on generating items for a high-stake medical licensure examination in the content area of surgery. The purpose of this examination is to assess the competencies of medical school graduates who wish to enter supervised clinical practice (i.e., medical residency). Using existing items and examinee response data from an operational medical licensure exam, the purpose of practical application #1 was to demonstrate how the LLTM-RE could be used with existing data to code for skills and then use these skill codes to estimate the difficulty for generated surgery items.

Skills-based CIG with the LLTM-RE required three steps for our first practical application. First, two content experts who were both practicing surgeons and experts in item development identified

skills that were expected to affect the difficulty level of surgery items on a medical licensure exam. The experts identified these skills by reviewing a set of existing medical licensure items and describing the skills they thought would affect medical student performance. The skills the medical experts identified were intended to be independent of one other and applicable to all surgery items. They also provided their justification for why they expected the skill to affect the item difficulty level. Second, the skill codes were applied to 43 existing surgery items used in the current study. A third content expert who is a practicing surgeon and expert in educational testing applied the skill codes to the 43 operational surgery items. Third, the 43 coded items were used to estimate the skill weights using the LLTM-RE. The calibration was based on a random sample of 1000 examinees who wrote the 43 surgery items.

Recall, the LLTM-RE estimates item difficulty using the skill codes specified in a Q-matrix. The LLTM-RE can therefore be evaluated by comparing the skills-based item difficulty estimates with a non-skill based estimation of difficulty from a more general statistical method like the 1PL IRT model. After the quality of the LLTM-RE item parameters are evaluated, the weights from the skills can be used to estimate item difficulty for newly generated items.

For the medical licensure exam application, existing test items were used. The skill codes were developed post-hoc meaning that the skills were extracted from a review of operational items and then applied to another set of items. This practice of applying skill codes to existing items is referred to as *retrofitting*. Retrofitting, in its most general sense, can be described as the addition of a new technology or feature to an older system. Similarly, we might consider skills-based retrofitting as the application of a new coding or classification scheme to an existing set of items. While retrofitting is common in practice, we contend that the use of retrofitting to skills-based CIG may not yield satisfactory results because the original items were never designed to meet the requirements of the new coding scheme. In other words, the skills-based CIG codes have specific requirements about the cognitive structure of items but this structure is unlikely to exist without some form of principled test design. Therefore, to

calibrate data that are retrofit requires strong assumptions—namely, that the sample of operational items represent the entire domain of surgery items on the medical licensure test; that the skills described by the experts adequately represent the skills that affect item difficulty in the domain of surgery; and that the generated items can be coded with the skills described by the experts. Given these assumptions, examinees' observed responses can be used to estimate item difficulty for each skill.

Results

The 10 skills expected to affect item difficulty on medical licensure exams in surgery are presented in Table 1. One example is case likelihood. If the content required to diagnose a problem was a common medical education problem, then it is more likely that the examinee would be familiar with those types of problems leading to a higher probability for a correct response. The skill code for this type of item would be 0. Conversely, if the content required to diagnose the problem presented in the test item was an uncommon medical education problem, then it is less likely the examinee would be familiar with those types of problems leading to a lower probability for a correct response. The skill code for this type of item would be 1.

The Q-matrix for the medical items coded in our study is presented in Table 2. Using these codes, the conditional p-value for each skill was calculated (see Table 3). The initial evidence from the p-value provides some support for the adequacy of our skills-based coding scheme. For example, items with outpatient cases were more difficult for examinees (i.e., had lower p-value) than items with inpatient cases (i.e., had higher p-value). In fact, the majority of items that were expected to be more difficult, according to our content experts, did in fact yield lower conditional p-values. Hence, the skills-based codes appear to predict the conditional item p-values.

Despite these initially promising results, the LLTM-RE was not successful at modeling item difficulty. Table 4 summarizes key statistics for the LLTM-RE and 1PL IRT model. The log likelihood values (LL) provide a comparative fit measure between the expected response data from the model and

observed response data from the examinees. This measure reveals that the LLTM-RE provided poor fit to the observed response data relative to the 1PL IRT model (i.e., LL of -25269.00 versus -25156.59, respectively). The LLTM-RE provides a person variance estimate that should match the observed performance distribution of the examinees, if the model fits the data. With the variance for the performance distribution of the examinees set to 1, the estimate for the 1PL IRT models was close to unity at 0.91. The person variance for the LLTM-RE, by comparison, was small at 0.25. The LLTM-RE also provides an item variance estimate that should be close to 0, if the model fits the data. The item variance estimate provides a measure of how well items measure a specific skill set, where more homogeneity (i.e., a lower item variance) indicates more uniformity among different items designed to measure the same skill. For the 1PL IRT model, the variance among the difficulty estimates across all items was 0.00. But for the LLTM-RE, the item variance was still comparatively large at 0.81 indicating, again, that this model does not fit the data. Finally, we can compare the correlation among the difficulty estimates by models. The correlation between the 1PL IRT model and the p-values was high at -0.97. Conversely, the correlation of the difficulty estimates between the 1PL IRT model and the LLTM-RE was low at 0.22. The LLTM-RE difficulty estimates were also poorly correlated with the p-values at -0.30. In short, the low correlation between LLTM-RE with other models suggests that the LLTM-RE did not accurately predict item difficulty for the medical licensure exam.

Application to Calibrated Item Generation

Different sources of empirical evidence were used to evaluate the quality of the LLTM-RE for CIG in the medical licensure examination example. To begin, we reviewed the conditional p-values for each skill category. The ordering of the p-values suggested that our skill codes were predictive of item difficulty. Then, item difficulty was estimated for the LLTM-RE and compared with other statistical models of difficulty such as the 1PL IRT model and the item p-values. Our results suggested that the LLTM-RE did not provide an adequate description of the observed response data. The 1PL IRT provided

a better fit to the observed response data as it yielded a higher person and lower item variance estimate than the LLTM-RE. The correlation between difficulty estimated from LLTM-RE with p-values and 1 PL IRT model were also low. These results suggest the LLTM-RE did not adequately predict item difficulty.

Because the LLTM-RE did not adequately fit the data, the LLTM-RE skills weights (see Table 5) were not expected to provide a consistent estimate of difficulty for the generated items. When we apply the LLTM-RE skill weights to the 43 surgery items to estimate item difficulty, the results differ dramatically across the three estimates (i.e., LLTM-RE, p-value, 1PL IRT). As shown in Figure 1, the scatterplot yields varied difficulty estimates where the p-values and 1PL IRT model yield results that are strongly correlated with one another by weakly correlated with the LLTM-RE. From these results, we conclude that the LLTM-RE is not appropriate for CIG in this application.

PRACTICAL APPLICATION #2: COGNITIVE DIAGNOSTIC ASSESSMENT

Test Design Principles

The second practical application featured in our study focuses on generating items for a cognitive diagnostic assessment in mathematics. The purpose of the diagnostic assessment is to evaluate students' knowledge and problem-solving skills (herein called skills) on key curricular outcomes in mathematics and provide students' with diagnostic feedback on those skills that are deemed to be weak (see Gierl, Alves, & Taylor-Majeau, 2010). However, unlike the test design in the medical licensure testing example where existing items were used, the diagnostic math project was based on a principled test design approach where the items were developed specifically to measure specific problem-solving skills. As a result, practical application #2 provides a demonstration of how to use the LLTM-RE for CIG when the items are designed and generated intentionally to measure specific skills.

Test design principles and practices have been developed to guide the creation of items. Embretson (1998), for instance, proposed a cognitive design system where a cognitive model of task performance guides the development of test items. Luecht (2007) introduced an assessment

engineering approach to item development where skills are mapped onto different dimensions of ability and items are systematically developed to meet the specific cognitive requirements within these dimensions. Cognitive diagnostic assessment (CDA) is a form of testing that relies on test design principles to evaluate student mastery on a specific set of knowledge and skills. The cornerstone of CDA is the use of a cognitive model to define the necessary skills required to solve problems in a particular content area. The cognitive model also guides the test development process by providing a blueprint of the knowledge and skills that each item must measure so that all the skills in the cognitive model can be properly evaluated. Gierl, Leighton, and Hunka (2007) summarized three key design principles required for CDA. First, a cognitive model is needed so the knowledge and skills required to solve the test items can be identified. Often, the knowledge and skills are ordered in a hierarchical manner implying that certain prerequisite and predictable knowledge structure exists to characterize cognitive problem solving. Second, the cognitive skills must be measurable with test items. Third, the knowledge and skills measured with the test items need to be instructionally relevant and meaningful to students, teachers, and parents. Using these design principles, a cognitive diagnostic assessment developed for Grade 3 mathematics is featured in our practical application #2.

A cognitive model was first created to define and organize the knowledge and skills required to solve math problems in the Grade 3 curriculum. Eight cognitive skills were defined by a team of four content experts who were also experienced test developers. The skills were organized in a hierarchical sequence ranging from simple to complex, where mastery of a more difficult skill implied mastery of an easier skill in the hierarchy. The skills are presented in Table 6. After the cognitive model and skills were defined, items were created based on the definitions of the skills in Table 6.

In total, 20 operational math items were developed to measure the eight skills in Table 6 (different numbers of items were created for each skill because in the original design of the Diagnostic Mathematics Assessment the skill requirements varied by test form). The Q-matrix is presented in Table

7. Notice, the skills in this table are organized from least to most complex. Results from 102 students who responded to each item within two weeks of completing a given instructional topic were used.

Results

A summary of the conditional p-values from the student results are presented in Table 8. For the most part, as the p-values decrease, the complexity of skill increases. This suggests that the items are, in fact, ordered correctly and measure cognitive skills of increasing complexity. Table 9 summarizes key statistics for the LLTM-RE and 1PL IRT model. The log likelihood values (LL) provide a comparative fit measure between the expected response data from the model and observed response data from the examinees. For our second practical example, the LLTM-RE and the 1PL IRT model provided a comparable fit to the observed response data (-1092 versus -1084). The person variance estimate should match the observed performance distribution of the examinees at unity, if the model fits the data. The LLTM-RE and the 1PL IRT model both yield high person variance estimates (0.99 and 0.98, respectively). The item variance estimate should be small if the model fits the data thereby indicating that different items measure the same skill uniformly on the test. For the LLTM-RE in this example, the item variances was small (0.03), indicating the skills can be accurately modeled. Finally, the correlation of the difficulty estimates across models should be strong, as this result would suggest that the difficulty estimates are comparable across models. The correlation between the 1PL IRT model and the p-values was high at -0.99. But the correlation of the difficulty estimates between the LLTM-RE and the 1 PL IRT model was also high at 0.96 and with the p-values at -0.97. In short, the high correlations between LLTM-RE with the other models confirms that the LLTM-RE accurately predicted item difficulty.

Application to Calibrated Item Generation

As with the medical licensure exam, different sources of empirical evidence were used to evaluate the quality of the LLTM-RE for CIG in the cognitive diagnostic assessment example. We reviewed the conditional p-values for each skill category. The ordering of the p-values suggested that

the skill codes were predictive of item difficulty. The LLTM-RE provided excellent fit to the observed response data relative to the 1PL IRT model using the log likelihood value. The LLTM-RE produced a higher person and lower item variance estimate. The correlation between difficulty estimated from LLTM-RE with p-values and 1 PL IRT model were also high. Taken together, these results suggest that the LLTM-RE can be used to predict item difficulty for the generated items.

Because the LLTM-RE adequately fit the data, the LLTM-RE skills weights (see Table 10) were expected to provide a consistent estimate of the difficulty level for the generated items. Item difficulty for the generated items can be calculated by taking the fixed effects for the LLTM-RE presented in Table 10, multiplying each skill by the item, and then adding a random effect based on the item variance. Figure 2 is a plot of the estimated parameters for the 20 items used in the calibration process along with their corresponding p-values. As highlighted by the high correlation between LLTM-RE with other models (see Table 9), the difficulty estimates were similar across models. In sum, the LLTM-RE was able to accurately predict item difficulty based on the skills that were designed to measure in each item.

With an adequate model for estimating difficulty, the generated items can now be calibrated. A fixed set of five items was generated for each skill (i.e., 40 items in total) and then the difficulty of these newly generated items was estimated using the fixed-effect and random-effects weights from the LLTM-RE. Because of the diagnostic assessment design where multiple forms were required, we also had a parallel set of five operational items that were designed to measure each skill. In other words, we had 40 generated items (i.e., five items each measuring eight skills) and 40 operational items (i.e., five items measuring eight skills) developed using two different methods: The first item set contained generated items and the second item set contained assessment tasks developed by content specialists. Figure 3 contains a scatterplot of the p-values and the LLTM-RE difficulty estimates for the items. That is, a 1, for example, in Figure 3 is 1 of 5 items measuring skill 1. The x-axis is the LLTM-RE difficulty estimate for the generate item and the y-axis is the p-value for this operational item. The correlation between the p-

values and the generated item difficulty estimate is -0.70, indicating that there is a strong relationship between the student responses and the skills-based estimates on the two sets of items in this comparison. To provide some perspective on the adequacy of this correlation, Embretson (1998) claimed that skills-based diagnostic items created using a principled design approach typically correlate in 0.70's to 0.80's when these items are implemented in an operational assessment. If this rule of thumb is extended to CIG, then the correlation of -0.70 between our skills-based CIG estimates with generated items and the p-values for the operational items demonstrates that our generated items are functioning in a comparable manner, psychometrically speaking, to real items.

DISCUSSION AND CONCLUSIONS

The purpose of our study was to demonstrate how skills-based calibrated item generation (CIG) can be conducted using the LLTM-RE. Two diverse and practical applications using real data were presented. But the most important characteristic that differentiated these two examples was the test design. For the medical licensure examination, skills were retrofitted to existing items on the test. The skills were identified by content experts who initially coded existing items for the features thought to affect difficulty, and then the codes were applied to a new set of operational items. Using the LLTM-RE to model item difficulty was not successful and, as a result, generated items could not be precalibrated. For the cognitive diagnostic assessment, skills organized from least to most complex were first identified in a cognitive model and then items were developed specifically to measure these skills. The skills and the items were created by content specialists for the diagnostic test. Using the LLTM-RE to model item difficulty was successful in this application, thereby allowing us to precalibrate the generated items. The correlation between item difficulty estimated using p-values and the LLTM-RE for generated items were high, indicating that the LLTM-RE could, in fact, be used to precalibrate the generated math items.

We contend that the different LLTM-RE outcomes across the two applications were not a result of the content areas, but rather the design principles used to align items with skills. The medical items,

without a readily available set of skills codes, required retrospective coding that likely contributed to the miss-identification and/or misalignment of items and skills. The math items, by comparison, were designed specifically to require the use of the skills. As a result, the item responses were better aligned to the skills thereby producing more accurate item difficulty estimates. It appears, then, that one important factor in the successful application of skills-based CIG is the use of a design principle that can help align the items and their associated skills.

In addition to proper alignment, the skills used for the cognitive diagnostic assessment were ordered by complexity from least to most difficult. The skills on the medical exam were neither ordered nor organized. Rather, they were thought to be independent. We contend that even though the skills were identified by content experts, the lack of a cognitive structure allowed many different patterns of skills to be used in the estimation process. The availability of a cognitive model in the math example limited the number of skill patterns and may have contributed to a better estimate of difficulty.

In the development of a CDA, Gierl and Leighton (2007) warned against retrofitting or repurposing existing items for diagnostic use because there is no guarantee that the items initially developed for one purpose can be recoded and, hence, repurposed for a different use. The results from our results suggest a similar caution is warranted when using the LLTM-RE for skills-based CIG. Without a set of ordered skills and a purposefully developed item designed specifically to measure a set of skills, the alignment between the item and the skills will be poor. Without a proper item-by-skill alignment, item difficulty may be challenging, or even impossible, to estimate.

Directions for Future Research

Demands for larger numbers of test items are increasing. The LLTM-RE modeling approach demonstrated in this study serves as one example of how generated item calibration can be conducted. By following design principles from CDA, generated items can be calibrated for item difficulty with high precision. Calibrating items using retrofitting, on the other hand, yield less desirable results. AIG

researchers have often described item calibration as a separate process from item generation when, in fact, outcome of both processes depend on one another. With a principled item development process, a known order of skill complexity, and a robust estimation method, items can be generated and calibrated in both an efficient and effective manner.

Just a few short years ago, the idea of calibrating generated items seemed as far fetched as the idea of automatic item generation itself. Yet skills-based CIG, as demonstrated in this study, can be a feasible and viable method for estimating the psychometric characteristics of generated items. Of course, some important hurdles still remain. For example, the organization, definition, and alignment of skills are rarely discussed in the literature on automatic item generation. While different statistical models are available for calibrating items, it is not clear how the organization and definition of skills may impact the calibrated outcomes. Also, as technology permits the generation of more diverse item types, (Gierl & Lai, 2012), research is needed on how CIG can be used to estimate the psychometric properties of these diverse outcomes.

Implications of this Study for the Medical Council of Canada

The Medical Council of Canada has been evaluating automated procedures for operational item generation (see Gierl, Lai, & Turner, in press). In this paper, we present a statistical method for calibrating the generated items. However, as our study makes clear, a principled approach to test design is also required for skills-based CIG. Principled test design requires two general stages. To begin, a cognitive model representation is required, initially, to identify the knowledge and skills required to make a medical diagnosis and, then, these skills must be ordered in a hierarchical manner. Cognitive model development could be conducted during the AIG process by adding a step to the development phase where the cognitive skills required to diagnose a medical problem are articulated by content specialists, organized in a model, and ordered by complexity. Because cognitive model development plays an important role in CIG, these models would also need to be validated, preferably by a second

group of medical content experts. Steps for cognitive model development are described by Gierl, Alves, Roberts, and Gotzmann (2009) and by Gierl, Alves, and Taylor-Majeau (2010).

Once a cognitive model has been developed, the AIG process could be guided by the structure of this model. In other words, the cognitive model could be used to design item models that systematically measure diagnostic problem-solving skills from least to most complex. Items could be generated to measure these ordered skills. A template used for item generation with a diagnostic math assessment is presented in Figure 4. The template contains three sections: the cognitive model, attribute structure, and the item models. Content specialist could use this type of template to develop medical item models thereby promoting a principled approach to test design that, in turn, would permit skills-based CIG.

REFERENCES

Arendasy, M., Sommer, M., & Gittler, G. (2010). Combining automatic item generation and experimental designs to investigate the contribution of cognitive components to the gender difference in mental rotation. *Intelligence,* 38, 506-512.

Bates, D., Maechler, M., & Bolker, B. (2012). LME*4: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999375-42, URL http://CRAN.R-project.org/package=lme4.

Bormuth, J. (1969). *On a theory of achievement test items*. Chicago: University of Chicago Press.

Breithaupt, K., Ariel, A., & Hare, D. (2010). Assembling an inventory of multistatge adaptive testing systems. In W. van der Linden & C. Glas (Eds.), *Elements of Adaptive Testing* (p. 247-266), New York, NY: Springer.

De Boeck, P. (2008) Random item IRT models. *Psychometrika,*73, 533-559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software,* 39, 1-28.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 300-396.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Gierl, M.J., Alves, C., & Taylor-Majeau, R. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing,* 10, 318-341.

Gierl, M.J., & Haladyna, T. (in press). *Automatic item generation: Theories and applications.* New York, NY:Routledge.

Gierl, M. J., Alves, C., & Taylor-Majeau, R. (2010). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing, 10,* 318-341.

Gierl, M. J. Alves, C., Roberts, M., & Gotzmann, A. (2009, April). *Using judgments from content specialists to develop cognitive models for diagnostic assessments.* In J. Gorin (Chair), How to Build a Cognitive Model for Educational Assessments. Paper presented in symposium conducted at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Gierl, M.J., & Lai, H. (2011, April). *The role of item models in automatic item generation.* Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Gierl, M.J., & Lai, H. (2012, April). *Methods for creating and evaluating item model structures for automatic item generation.* Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, CA.

Gierl, M. J., Lai, H., & Turner, S. (in press). Using automatic item generation to create multiple-choice items for assessments in medical education*. Medical Education.*

Gierl, M. J., & Leighton, J.(2007). Summary and conclusion. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 341–354). Cambridge, UK: Cambridge University Press.

Gierl, M.J., Leighton, J.P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242–274). Cambridge, UK: Cambridge University Press.

Glas, C., & van der Linden, W. (2006). Computerized adaptive testing with item cloning. *Law School Admission Council Computerized Testing Report,* 1, 1-13.

Janssen, R. (2010). Modeling the effect of item designs within the Rasch model. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 227-245). Washington DC: American Psychological Association.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and group predictors. In P. DeBoeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and non-linear approach* (pp. 189-212). New York: Springer.

Luecht, R. M. (2007, April). *Assessment engineering in language testing: From data models and templates to psychometrics*. Invited paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Masters, J. (2010). *A comparison of traditional test blueprinting and item development to assessment engineering in a licensure context.* Doctoral Dissertation. University of North Carolina Greensboro. Greensboro, NC.

R Development Core Team (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing*,* Vienna, Austria. URL http://www.R-project.org/.

Rudner, L. (2010). Implementing the Graduate Management Admission Test computerized adaptive test. In W. van der Linden & C. Glas (Eds.), *Elements of Adaptive Testing* (p. 151-165), New York, NY: Springer.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28,* 295-313.

Sinharay, S., & Johnson, M. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing, 8*, 209-236.

Wendt, A., Kao, S., Gorham, J., & Woo, A. (2009). Developing item variants: An empirical study. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved April 1, 2011 from www.psych.umn.edu/psylabs/CATCentral/

Zeuch, N. (2010). *Rule-based item construction: Analysis with and comparison of linear logistic test models and cognitive diagnostic models with two item types.* Doctoral Dissertation. Westfalische Wilhelms-Universitat. Münster, Germany.

Table 1.

Summary of the Skill Codes for the Medical Licensure Exam in the Content Area of Surgery

Skill	Description	Justification	Code
Clinical Case Likelihood (CL)	Clinical case likelihood describes if a case is likely to be taught during students' education (i.e., common) or not likely to be taught (i.e., uncommon). This feature is not to be confused with likelihood of case occurrence, which describes how likely a case occurs.	An examinee is more likely to provide a correct response if the presenting case was taught during their education. Hence, items that measure common cases are expected to be easier than items that measure uncommon cases.	0 - Common 1 - Uncommon
Amount of relevant case information (RI)	Whether the case provides just enough information about the diagnosis (i.e., sufficient), or more than enough information (i.e., ample) for a diagnosis.	An item that provides just enough information about the case for the correct response (i.e., sufficient) is expected to be more difficult than an item that provides ample information. Hence, items with sufficient information are expected to be more difficult than items with ample information.	0 - Sufficient 1 - Ample
Cue Typicality (CT)	A case contains cues about the patient that can lead to a diagnosis. Such cues can be either typical of the diagnosis (e.g., classic symptoms and demographic) or atypical (e.g., rare symptoms or unlikely demographic).	The presentation of an item with cues that are typical of the diagnosis would be easier for students than items with cues that are atypical.	0 - No 1 - Yes
Extremes of patient age (AG)	Whether the patient in the case is within the normal age range (i.e., between 18-70 years old) or outside of this range.	Items that measure cases with patients in extremes of age (i.e., less than 18 and greater than 70) are not common to most students. Students therefore are exposed to these cases less frequently. Hence, items with cases in the normal age range are deemed to be easier that items with extremes of age.	0 - Normal age range 1 - Extremes of age
Setting of Case Scenario (SE)	Cases with patients already admitted into a hospital (i.e., inpatients) compared to cases with patients not yet admitted to the hospital (i.e., outpatients).	Items that probe cases with outpatients are less common to students because they receive limited exposure to these situations. Hence, items with outpatients are expected to be more difficult than items with inpatients.	0 - Inpatient 1 - Outpatient
Specificity of Case Information (CI)	Each case provides different information (e.g., demographic, test results, physical exam findings) about a patient or a situation. Such information can be either specific to one diagnosis (i.e., high specificity), or generic information that can fit different diagnoses (i.e., low specificity).	When specificity of the case information on a given item is high, the item is considered to be easy. In contrast, when an item is presented with low specificity in case information, then the item is considered more difficult.	0 - High 1 - Low
Distracter Difficulty (DD)	Whether the solution of the task is provided with one, or more than one, plausible distracters for student selection.	The availability of multiple plausible distracters increases item difficulty, as students are required to only choose one.	0 - One plausible distracter (excluding key) 1 - Two or more plausible distracters
Knowledge Recall (KR)	Whether the correct solution to the task requires recall of a specific knowledge (i.e., rote memorization).	The item is easier when it depends only on knowledge recall compared to non-memory intensive item that requires multiple-steps as well as knowledge and skills.	0 - No 1 - Yes
Common Mistake (CM)	Whether the correct solution to the task is colluded with a mistake commonly committed by a novice examinees.	Items that prompt for common mistakes in novices are more difficult for students as they are novices in the field.	0 - No 1 - Yes

Severity/ Urgency of Diagnosis (SD)	The correct solution to the task requires either a severe diagnosis (i.e., life and death situation) or a non-severe diagnosis (i.e., non-life threatening).	Items that require a severe diagnosis are more likely to be recognized by students as they are an important entity in their education compared to diagnoses that are non-severe. As a result, items with a severe diagnosis will be easier than items that require a non-severe diagnosis.	0 – Severe/Urgent 1 – Non-severe/Non-Urgent

Table 2.

Q-Matrix Highlighting the Item-by-Skill Coding for 43 Surgery Items

	Skill									
Item	CL	RI	CT	AG	SE	CI	DD	KR	CM	SD
1	1	0	0	0	0	1	0	0	0	1
2	0	1	1	0	0	0	1	0	1	0
3	0	1	1	0	0	0	1	0	1	0
4	1	1	1	0	0	0	1	1	0	0
5	0	1	0	0	0	1	1	0	0	1
6	0	0	0	0	1	1	1	1	0	0
7	1	1	1	0	1	0	1	0	0	1
8	1	1	1	1	0	0	0	0	0	0
9	0	1	1	0	0	0	1	0	0	0
10	0	1	0	0	1	0	1	0	0	0
11	0	0	1	0	1	1	1	0	0	1
12	0	0	0	1	1	1	1	0	0	1
13	1	0	1	0	0	0	1	1	0	0
14	1	1	1	1	1	0	1	1	0	1
15	0	1	1	1	0	0	0	1	0	0
16	0	1	1	0	0	0	1	0	0	0
17	1	1	1	0	1	0	1	1	0	1
18	0	1	0	0	1	0	1	1	0	1
19	1	0	0	0	0	1	0	0	0	1
20	0	1	1	1	0	0	0	0	1	0
21	1	0	0	0	0	1	1	0	0	0
22	0	0	0	0	1	0	1	1	1	0
23	0	1	0	0	1	0	1	1	0	1
24	0	0	0	0	0	0	1	1	1	1
25	0	0	0	0	1	1	1	0	0	1
26	1	1	0	0	0	0	1	0	1	1
27	0	0	0	0	0	0	1	0	1	1
28	1	1	1	0	1	0	1	0	0	1
29	0	1	1	0	1	0	1	0	0	1
30	0	0	0	1	0	1	0	0	0	1
31	0	0	1	1	1	0	1	1	0	1
32	1	0	0	1	0	1	1	0	1	0
33	0	0	0	0	0	0	1	1	0	0
34	1	1	0	1	0	0	0	0	0	0
35	0	0	1	1	0	0	1	1	0	0
36	0	0	0	0	1	1	1	1	0	1

37	0	0	1	0	0	1	0	0	0	0
38	1	0	1	0	1	1	0	0	0	0
39	0	1	1	0	1	0	0	1	0	1
40	1	1	0	0	1	0	1	1	0	1
41	0	1	1	1	1	0	1	0	0	0
42	0	1	1	0	1	0	1	0	0	0
43	0	0	0	1	0	1	1	0	1	0

Table 3.

Summary of the Conditional P-value By Skill

	Code	
Feature	0	1
CL	0.67	0.55
RI	0.65	0.60
CT	0.64	0.61
AG	0.64	0.61
SE	0.60	0.65
CI	0.60	0.69
DD	0.62	0.63
KR	0.61	0.65
CM	0.63	0.62
SD	0.64	0.61

Table 4.

Summary of the LLTM-RE Model Using Surgery Items

	LL	Person Variance	Item Variance	Correlation with 1PL IRT	Correlation with p-value
IPL IRT	-25156.50	0.91	0.00	1.00	-0.97
LLTM-RE	-25269.00	0.25	0.81	0.22	-0.30

Table 5.

Fixed Effects of LLTM-RE for the Surgery Items

Fixed Effects	B	SE
LK	-0.02	0.30
IN	0.08	0.34
TY	0.40	0.30
AG	0.02	0.31
SE	-0.32	0.35
SP	0.37	0.34
DI	0.14	0.33
RE	0.69	0.31
CO	0.06	0.38
SV	-0.07	0.32

Table 6.

Eight Skills Measured by the Cognitive Diagnostic Assessment

Skill	Skill
1 (Simple)	Apply a skip counting pattern of 10 forward using any starting point from 100 to 1 000
2	Apply a skip counting pattern of 5 forward using any starting point from 100 to 1 000
3	Identify the place value meaning of a digit in a number using numbers 100 to 1 000
4	Represent a number concretely or identify a pictorial representation using numbers 100 to 1 000
5	Apply a mental math strategy to add two 2-digit numbers where one addend is a multiple of 10 greater than 50 using regrouping with sums greater than 99
6	Apply a mental math strategy to add two 2-digit numbers using regrouping with sums to 99
7	Solve a two-step word problem where one step is either division or multiplication using numbers from a 5 x 5 grid
8 (Complex)	Identify a word problem for a given division expression

Table 7.

Q-Matrix Highlighting the Item-by-Skill Coding for 20 Mathematics Items

	Skill							
Item	1	2	3	4	5	6	7	8
1 (Simple)	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0
4	1	1	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0
7	1	1	1	0	0	0	0	0
8	1	1	1	1	0	0	0	0
9	1	1	1	1	0	0	0	0
10	1	1	1	1	0	0	0	0
11	1	1	1	1	1	0	0	0
12	1	1	1	1	1	0	0	0
13	1	1	1	1	1	0	0	0
14	1	1	1	1	1	1	0	0
15	1	1	1	1	1	1	0	0
16	1	1	1	1	1	1	1	0
17	1	1	1	1	1	1	1	0
18	1	1	1	1	1	1	1	1
19	1	1	1	1	1	1	1	1
20 (Complex)	1	1	1	1	1	1	1	1

Table 8.

Summary of the Conditional P-value By Skill

	Coding	
Skill	1	0
1	0.89	0.59
2	0.82	0.58
3	0.66	0.61
4	0.75	0.59
5	0.60	0.62
6	0.52	0.63
7	0.29	0.65
8	0.39	0.66

Table 9.

Summary of the LLTM-RE Model Using the Mathematics Items

Model	LL	Person Variance	Item Variance	Correlation with P-Value	Correlation with 1PL IRT
1PL IRT	-1084.00	0.99	0.00	-0.99	1.00
LLTM-RE	-1092.00	0.98	0.03	-0.97	0.96

Table 10.

Fixed Effects of LLTM-RE for the Mathematics Skills

Fixed Effects	**B**	**SE**
1	-2.41*	0.27
2	0.59	0.31
3	1.00*	0.26
4	-0.52*	0.25
5	0.81*	0.22
6	0.65*	0.23
7	0.98*	0.26
8	-0.57*	0.24

*p<.05

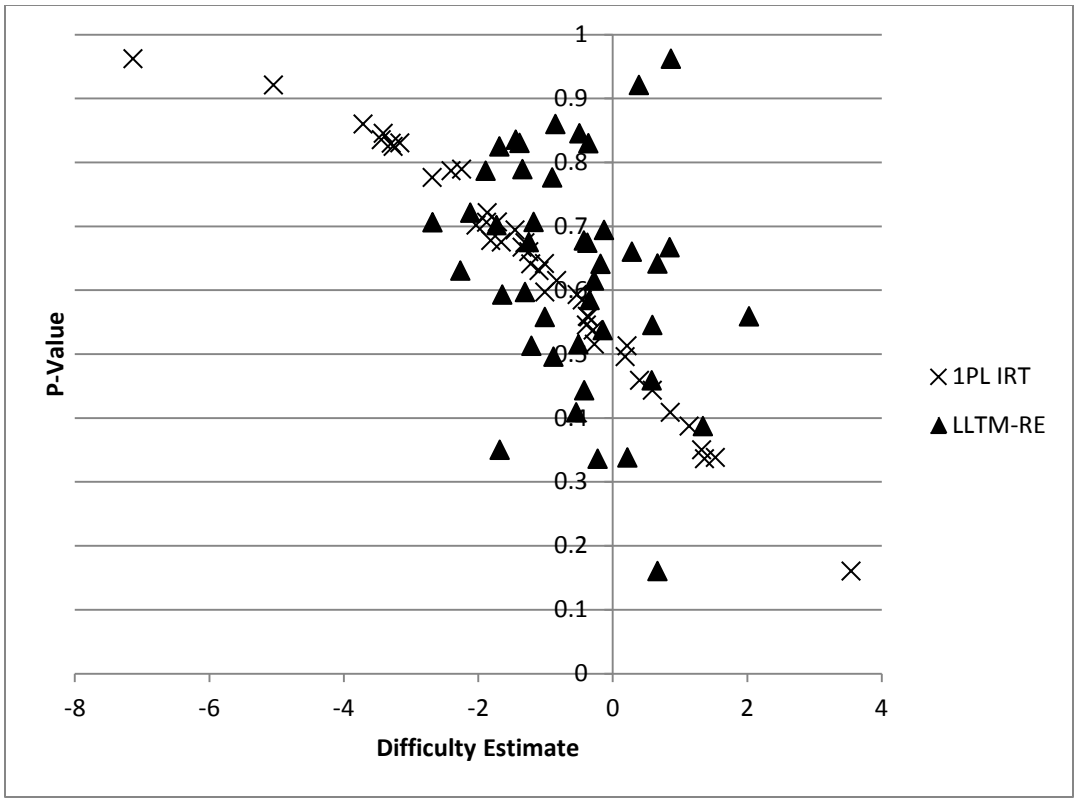

Figure 1. ***Plot of p-values, 1PL IRT and LLTM-RE difficulty estimates for medical licensure items.***

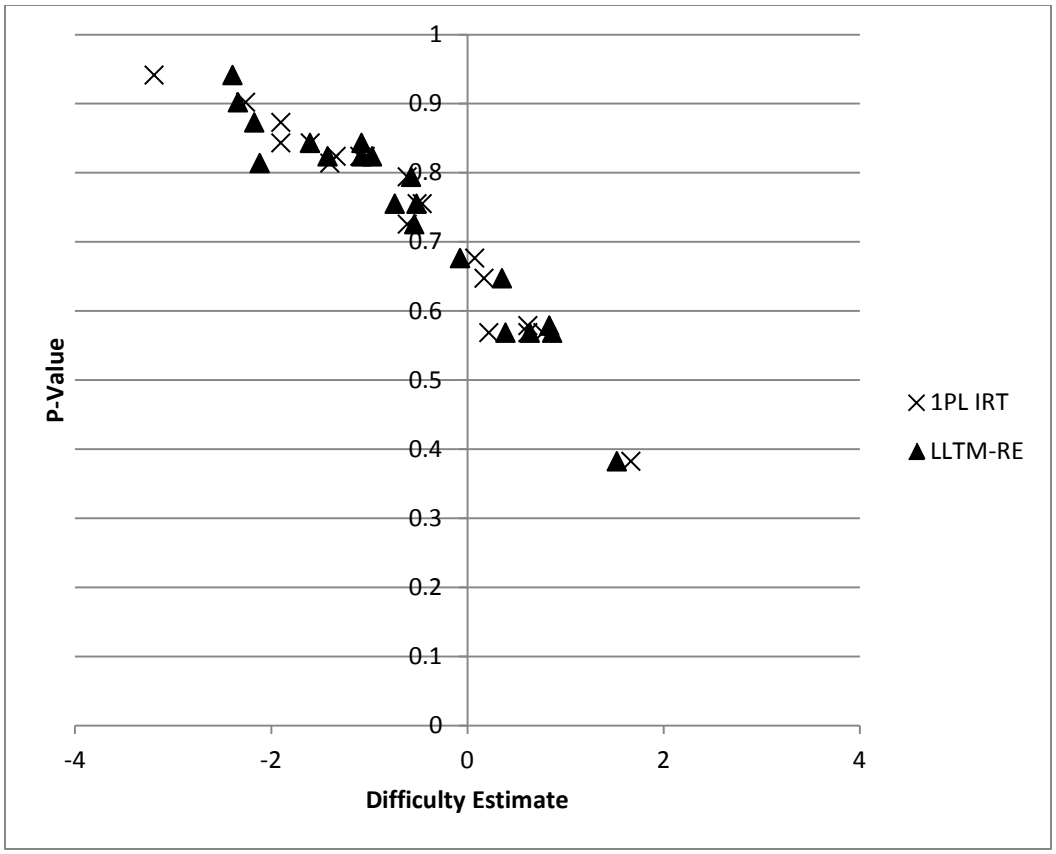

Figure 2. Plot of 1PL IRT and LLTM-RE difficulty estimates for the math items.

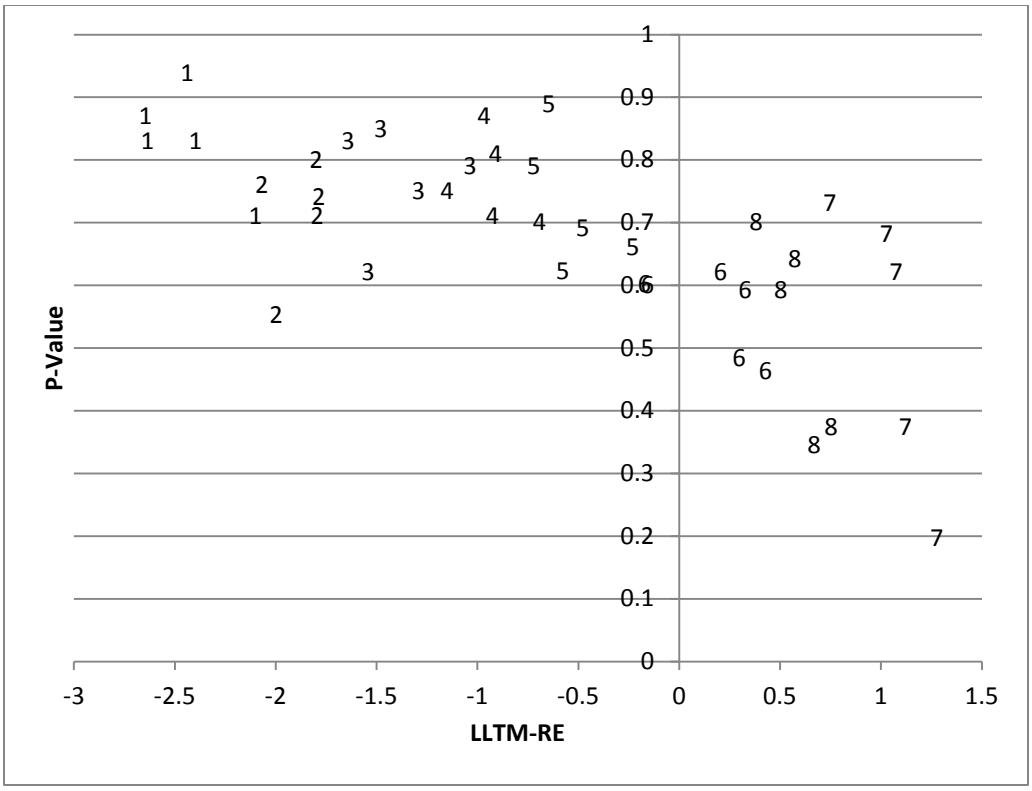

Figure 3. ***Scatterplot of estimated difficulty for generated items and p-values for operational items as a function of the skill required to solve the assessment task.***

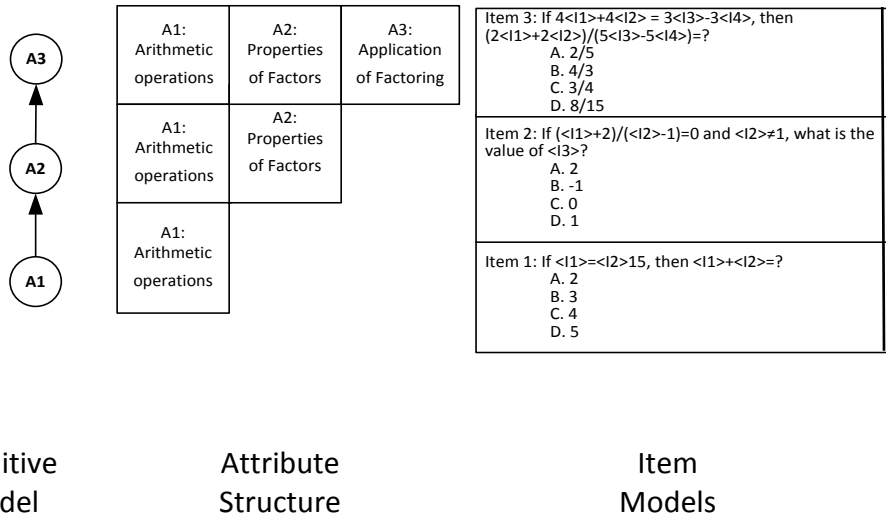

Figure 4. Principled design approach for item model development in mathematics.