Journal of Applied Testing Technology, *Vol 20(1) 16-33, 2019*

A Process for Automatically Generating Algebra Items

Audra E. Kosh

MetaMetrics, Inc. 1000 Park Forty Plaza # 120, Durham, NC 27713; audrakosh@gmail.com

Abstract

Developing high-quality test items is a costly and time-consuming endeavor, yet the need for large item banks increases as testing becomes more frequent and as computer-adaptive tests call for many items of varying difficulty. To address this need, some test developers have begun to implement Automatic Item Generation (AIG) to supplement item writing efforts. Research and practice on AIG continue to grow as technology and psychometric methods develop, but the field lacks content-specific processes for implementing AIG. The purpose of this paper is to present a process for creating AIG item models to assess middle school students' algebraic reasoning by building upon existing AIG methodology and literature regarding the cognitive complexity of algebra tasks.

Keywords: Assessment Design, Item Generation, Test Development

1. Introduction

Item development includes many stages such as writing item specifications, drafting items, designing relevant graphics, editing and revising items, and field-testing items (Haladyna & Rodriguez, 2013). Given the substantial effort to develop items, it is no surprise that Gierl and Haladyna (2013) estimate that developing an item bank for a high-stakes computer-adaptive test administered twice per year with forty items would cost three- to four-million dollars. As a result of item development demands, test developers have begun to turn to alternate methods of item development.

A hopeful solution to item development is AIG. AIG is an item development process that can be used to "supplement, rather than replace, traditional test development procedures by offering a systematic and strategic approach for manipulating content so items can be produced automatically" (Gierl, Zhou, & Alves, 2008, p. 32). AIG incorporates theoretical and empirical research about examinees' cognitive processes in order to carefully develop items with consideration to which features of the item impact difficulty.

AIG is typically classified as either weak or strong theory AIG. In weak theory AIG, test developers select existing items with desirable psychometric properties and then use those items as item shells (i.e., templates) to form new items by varying surface features of the item thought to not affect how students process the item (Drasgow, Luecht & Bennett, 2006; Gierl & Lai, 2013). The second approach to AIG, strong theory AIG, instead seeks to "generate *calibrated* items automatically from design *principles* by using a *theory* of difficulty based on a cognitive model" (Drasgow, Luecht, & Bennett, 2006, p. 474, emphases in original). The emphases on *principles* and *theory* to create *calibrated* items highlight that strong theory AIG begins with a theory of how examinees cognitively process within the domain of the test; items can then be designed according to that theory. The theory lends itself to the ability to vary features of the items in such a way that the psychometric properties of the items can be predicted, including the item's difficulty. There are two main requirements for strong theory AIG: 1. Items must be described in sufficient detail so that a computer program can produce the item, and 2. Variables impacting

item difficulty must be identified and systematically controlled (Gierl, Zhou & Alves, 2008).

Gierl and Lai (2016) outline a three-step process for strong theory AIG: 1. Developing cognitive models, 2. Developing item models, and 3. Using computer technology to generate items. The first step, developing cognitive models, includes characterizing "the knowledge and skills examinees require to solve items on assessments" (Leighton & Gierl, 2011, p. 49); this step is important to ensure that tests are "designed according to the science of human learning" (Leighton & Gierl, 2011). The cognitive model developed in step one becomes the basis for developing item models in the second step of AIG. An item model is "a template which highlights how the features in an assessment task can be manipulated to produce new items" (Gierl, Lai, Hogan, & Matovinovic, 2015, p. 2). In other words, the item model encompasses specifications for writing an item, and those specifications contain enough detail to warrant creation through algorithmic, computerized means (Bejar, 2010). After creation of item models, the third step is applying computer technology to generate items by combinatorically varying elements of the item model.

In the three-step process of AIG, a key step is developing item models. Despite item model creation serving as a critical methodological component of AIG, to my knowledge there are no published methods that describe the principles or standards used to create AIG item models. Instead, researchers merely present the cognitive model and item models they used without describing how those item models were created, raising a concern about the theory and research – or lack thereof – used to arrive at item models for AIG. Although other test development methods may use techniques that are similar to the item models used in AIG, these methods do not specifically focus on automating the item generation process. For example, in evidence-centered design, whereby test content is developed with the goal of collecting evidence to support a claim about an examiner, task models describe the components of a family of tasks such as the format of the tasks, the products the examinee creates, and variables allowed to differ within the tasks (Mislevy, Almond & Lucas, 2003). Unlike strong-theory AIG, evidence-centered design forms task models based on a student model and evidence model which describe a profile of a student and the necessary observations of that student required to make a claim, as opposed to AIG creating item models based on cognitive models of how examinees process within the domain. Also, in evidence-centered design, task shells (i.e., the templates for items created from task models) may be based on existing well-performing items as opposed to generating new items from item models as is the case in AIG (Zieky, 2014).

The lack of methodology for developing item models, particularly with respect to methods for translating the cognitive model into operational item models, occurs alongside other limitations of current work in AIG. Most strong-theory AIG studies follow the same general three-step methodology – developing cognitive models, developing item models, and generating items with technology – regardless of the intended population or purpose of the test. Although cognitive modeling certainly considers content, the three-step process fails to identify content-specific considerations that item developers should account for when writing item models. These limitations in current AIG work raise the need for domain-specific methodological processes to support item model development. The purpose of this paper is to propose a methodological process for developing AIG item models for generating items that assess middle school students' algebraic reasoning, first by reviewing current research and theory on the cognitive complexity of algebraic reasoning tasks in order to inform process development.

2. Features Affecting the Level of Challenge of Algebraic Reasoning Tasks

Because strong theory AIG requires systematically varying features of items thought to affect difficulty, a process of item development must account for previous research on content-specific features of tasks that impact difficulty. In this paper, the content under consideration is algebraic reasoning. Blanton and Kaput (2005) define algebraic reasoning as "a process in which students generalize mathematical ideas from a set of particular instances, establish those generalizations through the discourse of argumentation, and express them in increasingly formal and age-appropriate ways" (p. 413). The inclusion of algebraic reasoning content standards across many grade levels indicates that algebraic tasks represent a wide range of difficulty as student's progress in their mathematical development (Common Core State Standards Initiative, 2015).

As related to AIG, identifying features of tasks that impact difficulty is a key component of conducting AIG so that these features can be manipulated in item generation. I next present a review of literature regarding features that impact the difficulty of algebraic items. The first part of the review specifically covers literature on story problems (i.e., word problems), and the second part covers mathematical structures such as numbers, relations, and variables.

2.1 Story Problems and Semantics

One line of research about difficulty of algebraic tasks regards whether story problems or equation problems are more difficult. Teachers and researchers predominantly believe that story problems are more difficult for students because they pose an additional step of translating the story context to an equation prior to solving (Nathan & Koedinger, 2000). This belief makes the limiting assumption that students solve story problems by translating the context to an equation first. In actuality, students utilize an array of informal strategies, and research on whether story problems or equation problems are more difficult contains much more nuanced conclusions (Koedinger & Nathan, 2004).

One study on story problems and equation problems is that of Carraher and colleagues (1985). The researchers compared the results of three types of assessments for children who worked as street vendors in Brazil. Findings indicated that children completed 98.2% of calculations correctly when calculations occurred in the informal context of authentic money transactions, but performance declined to 73.7% and 36.8% correct on matched word and symbol problems, respectively. Results also showed that students used different strategies when completing computations in an authentic context as opposed to on a written test. Other researchers have likewise documented students' and adults' use of informal mathematical strategies when performing everyday mathematics in situated environments such as calculating change or making a budget (Goldman & Booker, 2009; Lave, 1988; Saxe, 1988a, 1988b; Taylor, 2009).

This aforementioned research of situated mathematics has the unique feature of occurring in informal, authentic contexts; other researchers have investigated whether story problems or symbol problems are more difficult for children in a school environment where mathematics tasks are typically fabricated to align to a learning objective. Nathan and Koedinger (2000) pointed out that teachers and researchers have a "symbol precedence model of development algebraic reasoning," meaning they believe that students first learn how to solve symbolic equations and then learn to solve story problems by using a strategy whereby the story context is translated to an equation and then solved (p. 168). The symbol-precedent view is corroborated in textbook design as well; in nine out of ten textbooks analyzed by Nathan, Long and Alibali (2002), equations were presented prior to story problems. Moreover, story problems are often presented toward the end of chapters as challenge problems.

The symbol-precedence view of algebraic development held by researchers and teachers has since been debunked as inaccurate because the issue is much more complex. In a study of high-school students, Koedinger and Nathan (2004) tested student performance on three different types of problems matched for mathematical structure and varied by presentation format: 1. Story problems (e.g., a question about a waiter making tips and an hourly rate), 2. Word equations (e.g., "Starting with \$81.90, I subtract \$66 and then divide by 6. What number do I get?"), and 3. Symbolic equations (e.g., Solve for x: $(81.90-66)/6 = x$). Students performed statistically significantly better on story problems and word equations as compared to symbolic equations, but there were no statistically significant differences on performance between story problems and word equations. The authors concluded that presenting problems verbally as opposed to symbolically is the key determinant of difficulty rather than the situational context.

A curiosity in Koedinger and Nathan's (2004) study was that student success in story problems and word equations was linked to the use of informal strategies such as guess-and-check or unwinding (i.e., working backwards from the answer), but such strategies may not be effective for more complex algebra problems. To test the hypothesis that there is a trade-off between problem presentation and complexity, Koedinger, Alibali and Nathan (2008) conducted a follow-up study with more complicated problems tested on college students. The 2008 study included double-reference problems where an unknown quantity is used twice in the matched symbolic equation. For example: the problem "Roseanne just paid \$38.24 for new jeans. She got them at a 15% discount. What was the original price?" translates to the equation $x - 0.15x = 38.24$ where the variable occurs twice (p. 370). As expected, since these double-reference problems

devalue informal strategies, results showed that students performed better on symbolic equations than story problems.

Enright, Morley and Sheehan (2002) conducted a study similar to those of Nathan and Koedinger that examined the impact of particular story and equation problem features on difficulty. Their goal was to identify item features that predicted psychometric characteristics in order to reduce sample sizes for item pretesting. The authors systematically varied characteristics of two sets of algebraic story problems related to rates and probability in a sample of GRE examinees. For the rate problems, they varied whether the item included variables or numbers, the context of the problem (i.e., cost or distance), and the level of complexity of the constraints in the problem. The factor that impacted difficulty the most was the presence of variables as opposed to numbers. Interestingly, the authors found that the effect of context depended on whether or not variables were required: for rate problems without variables, a cost context (e.g., prices with dollar signs to calculate a unit rate) made the item statistically significantly easier than a distance context (e.g., miles per hour). But, for items with variables, there was no statistically difference between cost and distance rate problems. Similar to previous findings, these results indicate that context matters less for more mathematically complex problems such as problems using variables as opposed to numbers.

For the probability problems in Enright and colleagues' (2002) study, the authors varied whether the item was phrased as a problem about probability (i.e., "What is the probability of…") or percentage (i.e., "Which percentage of…") and whether the context was a real-life scenario or a pure number context (e.g., "An integer is chosen at random from…"). The authors also varied the complexity of counting when describing the sample space in a probability item (e.g., "integer between 200 and 399" versus "integer beginning with the digits 2 or 3 and ending with the digits 8 or 9"). Results showed that complexity of counting had the greatest impact on item difficulty and that items phrased as probability questions were more difficult than items phrased as percentage questions. Real-life versus pure mathematical context had no statistically significant differences on item difficulty.

Even within the domain of story problems, subtle changes to the phrasing of an item can dramatically affect difficulty. As a simple example, Hudson (1983) showed first-graders pictures of worms and birds and found that only 64% of first-graders correctly answered the question "How many more birds than worms are there?", but 100% of the same children answered correctly when rephrasing the question as "How many birds won't get a worm?" Although these questions are identical in terms of their mathematical structure, the latter question provides the student with additional context clues that invites a strategy of directly modeling a one-to-one correspondence between birds and worms. Moreover, the phrasing of the second question does not require an understanding of the word "more" as the first question does.

Additional studies of story problems further demonstrate how minor semantic changes impact difficulty, particularly through the use of keywords that signal students to use particular operations or strategies. Martin and Bassok (2005) defined translation cues as "standardized phrases and keywords that are highly correlated with correct solutions" which allow students to go directly from words to solution strategies with little need to interpret the context of the story problem (p. 471). For example, students identify *altogether* to mean addition, *difference* to mean subtraction, and *times* to mean multiplication.

Translation cue strategies can backfire when a mismatch exists between the translation cue and solution strategy. For example, in the statement "There are six times as many students (S) as professors (P)," 37% of undergraduate engineering students incorrectly translated this sentence to an expression, usually by answering with 6S=P (Clement, 1982). This type of error, known as a reversal error, commonly occurs when the student tries to directly translate the keywords in the statement to an expression without making sense of the relationship between quantities. Another example of how semantics can complicate mathematics is the bat and ball problem ("A bat and a ball cost $1.10. The bat costs one dollar more than the ball. How much does the ball cost?"), for which over 50% of students at elite universities responded with the incorrect answer of ten cents (Kahneman, 2011).

To further investigate how students use word cues, Martin and Bassok (2005) conducted a study to test their hypothesis that story problems also have semantic cues in addition to translation cues that affect which strategies students use and the likelihood of a correct response. They hypothesized that certain objects represent symmetrical relationships usually modeled by addition or subtraction (e.g., blue marbles and red marbles) whereas other words represent asymmetrical relationships

associated with multiplication and division (e.g., apples and baskets; chairs and tables). Problems have semantic alignment when the symmetrical or asymmetrical relationship between the words in the problem matches the correct solution strategy (Bassok & Martin, 1998). Martin and Bassok (2005) presented seventh-graders, ninth-graders, eleventh-graders, and college students with different types of problems that varied in their semantic alignment and whether students were asked to provide a numerical answer or write an expression or equation. As hypothesized, semantic alignment affected whether students answered story problems correctly; more students answered correctly to semantically aligned problems, as expected. However, semantic alignment had no effect on expression or equation writing tasks. Also, although students performed better on problems with semantic alignment, this effect was stronger for younger students and diminished as age increased. These results imply that word cues matter less for higher-ability students answering more complex questions, probably because the mathematical complexity of the problem trumps context cues.

Koedinger, Alibali and Nathan (2008) conducted similar research on story problem phrasing with high-school and college algebra students. The authors tested for difficulty differences between story-implicit problems expressed with everyday language of mathematical operations (e.g., gave away, kept) and story-explicit problems that utilized mathematics terminology (e.g., subtracted, added). They found no statistically significant differences between performances on story-implicit versus story-explicit problems; however, this result is likely due to the age of the participants. It is possible that younger students are still developing understanding of mathematical terminology, as was the case in the aforementioned lower-elementary and middle school examples, whereas high-school and college students already have this foundational knowledge.

Combing all of the above research on mathematics story problems illustrates that the role of context as a predictor of item difficulty is not a simple question about whether story problems are more difficult than word problems. Instead, the context of a problem interacts with other features of the problem and the student, with evidence supporting the idea of a tradeoff between story problem context, mathematical complexity of the problem, and the age and ability of the student. Research findings indicate that context affects lower-level students more so than students completing higher-level mathematics.

2.2 Numerical and Relational Mathematical Structures

In addition to research on story problems, research and theory also examine the numerical and relational components of mathematics. Regarding theories, Mayer (1982) introduced the idea of classifying equations by the number of *moves* and *computes* needed to isolate a variable. A move refers to applying inverse operations to each side of the equation, and a compute refers to simplifying one side of the equation (e.g., combining like terms or applying the distributive property). Later, Mayer, Larkin and Kadane (1984) expanded this idea to develop a theory of problem-solving which accounts for how students make sense of the problem given to them and then implement a plan for completing the problem, called problem representation and problem execution.

In more recent work, Bickel and colleagues (2015) proposed a semantic and syntactic theory of mathematics task difficulty which mirrors methods for measuring text complexity in reading with semantic components (i.e., word difficulty) and syntactic components (i.e., sentence length). In mathematics, semantic refers to the meaning of elements in a task and syntactic refers to how those elements fit together, for example through identifying the conceptual demand of a task (semantic) and the number of steps required to complete the task (syntactic). A limitation of this theory is that it can be difficult to distinguish between syntactic and semantic components of a mathematical task. Instead, the complex nature of mathematics inherently intertwines the two.

Regardless of which theories most thoroughly describe mathematics task difficulty, certain replicated research findings allow identification of a set of factors that consistently impact a student's likelihood of answering an item correctly. First, the number system matters. Problems with whole numbers are generally easier than problems with fractions or decimals (Graf & Fife, 2013; Koedinger & Nathan, 2004). Moreover, phrasing numbers as prices (e.g., with dollar signs) or as decimals in multiples of .25 to mirror a quarter makes items easier as compared to numbers without a money context (Baranes, Perry & Stigler, 1989; Sebrechts, et, al., 1996). Also, the presence of variables makes items more difficult as compared to

items with numbers only, and items become increasingly difficult when they have more than one variable (Graf & Fife, 2013; Sebrechts et. al., 1996). In addition to the mathematical structures included in the item, the type of task students are asked to complete also impacts difficulty. For example, items that require students to express answers in the form of an expression are more difficult than those only requiring a numerical answer (Graf & Fife, 2013).

Procedural mathematics tasks can also be ordered by classifying which problems are contained within a more complex problem. For example, a student must know how to compute 1+4 in order to compute 31+24 because adding the ones digits in the latter requires adding 1+4 (O'Rourke, Andersen, Gulwani & Popovic, 2015). In other words, the problem 1+4 is contained inside the problem 31+24, thereby making the former easier. This type of classification system – stemming from the computer science field – determines difficulty order by specifying which algorithms must be programmed first before programming a more sophisticated algorithm.

2.3 Modeling Cognitive Complexity of Algebraic Reasoning Tasks

Much of the aforementioned research focuses on one task feature at a time when predicting task difficulty. In addition to studying task complexity by experimentally testing a single task feature, researchers have also holistically modeled the cognitive complexity of algebraic reasoning tasks by considering multiple features simultaneously as applied to a variety of problem types. Ideally, models predicting item difficulty explain a large percent of variance in item difficulty because this implies that the construct-relevant features used in item generation account for the majority of cognitive processes impacting difficulty. In some highly successful cases, models explain up to 90% of variance in item difficulty; more commonly, results fall within the 30 to 60% range (Gorin & Embretson, 2013).

As an example of item difficulty modelling, Embretson and Daniel (2008) investigated which cognitive processes were predictive of item difficulty on a wide range of quantitative GRE items administered to undergraduate students. The authors coded GRE items according to cognitive processes required to solve the problem such as recalling an equation, generating an equation to represent the problem context, the maximum grade level of knowledge required to solve the problem, the number of computations required to solve the problem, and other similar variables. Results showed that almost all of the variables used in the cognitive model statistically significantly predicted item difficulty and explained about half of the variance in item difficulty. In a subsequent study, Daniel and Embretson (2010) generated items according to the cognitive model used in their 2008 study. Again, using undergraduate participants with variants of GRE items, the researchers found that items that required examinees to generate or recall equations were more difficult than items that contained equations without words.

Sometimes, as few as one or two powerful item features explain a large percent of variance in item difficulty. For example, Simpson and colleagues (2015) found that a key predictor of third-grade multi-digit addition and subtraction item difficulty was whether or not the problem required regrouping (i.e., carrying or borrowing). Similarly, Sebrechts and colleagues (1996) found that as few as two item features (e.g., number of variables in an item and the number of equations required to solve an item) predicted 47% of variance in algebraic items from the GRE.

Yet another study using GRE items is that of Enright, Morley and Sheehan (2002) described earlier with probability and rate problems. Their study perhaps is one of the most successful – if percent of variance of difficulty explained by the model is used as a measure of successful item generation – with a model for rate problems that accounted for 90% of variance in item difficulty. In other words, the correlation between item difficulty as predicted by the model and empirical item difficulty was around .95, an extremely high correlation for modeling item difficulty. The model for probability problems was less successful but still satisfactory, explaining 61% of variance in item difficulty.

A trend in the literature on model-based item generation in mathematics is that the majority of studies use GRE items administered to adults. Research with K-12 learners is severely lacking, and conclusions about algebraic complexity from studies with adults cannot be generalized to younger students who are still rapidly growing in their mathematical development. Yet, K-12 learners are extensively studied in the learning sciences field. It is clear that psychometricians working on item generation and learning scientists studying K-12 algebraic thinking could both benefit from more collaboration.

Psychometricians would be able to produce better items that most successfully capture the cognitive processes students complete when performing a mathematical task, and learning scientists would have additional methods for experimentally testing theories about the science of learning. In the next section I present a process for generating middle school algebraic reasoning items that builds upon both AIG and learning sciences research.

3. The Item Model Creation Process - Algebra

In this section, I propose a process for using AIG to develop items that assess middle school students' algebraic reasoning, subsequently referred to as the Item Model Creation Process - Algebra (IMCP-A), that demonstrates a process for translating cognitive models into item models for AIG. The IMCP-A includes five stages of item model development: 1. Identify a broad-grained schema, 2. Identify features impacting cognitive complexity through analyzing factors of structural complexity, contextual support, and extent of generalization, 3. Determine item type, 4. Specify item meta-model, and 5. Write item models. I next describe each section of the IMCP-A in detail along with a concurrent example, and a visual representation of the IMCP-A is presented in Figure 1.

3.1 Step 1: Identify Broad-Grained Schema

The IMCP-A begins with identifying an algebraic broad-grained schema characterizing the desired content of the item. Schemas represent classes of problems that students' group together in their cognitive processing by categorizing problems that share the same underlying mathematical structures (Singley & Bennet, 2002). The grain-size of a students' schema can vary and will expand as students develop. For example, a student might initially develop a schema for solving one-step linear equations, then two-step linear equations, then equations with variables on both sides, and so on until eventually the student develops an overarching, broad-grained schema for solving equations. Another example is a schema for identifying relationships in bivariate tables (i.e., input/output tables) which entails finding missing numbers in tables and writing equations to model a relationship in a table across a variety of relationships (e.g., linear, quadratic, logarithmic). This skill appears in content standards across multiple grade levels with varying difficulty, such as in the Common Core State Standards for Mathematics whereby grade five students generate numbers from a pattern with two different rules and grade eight students write functions to model the relationship between two quantities (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). I use the term broad-grained

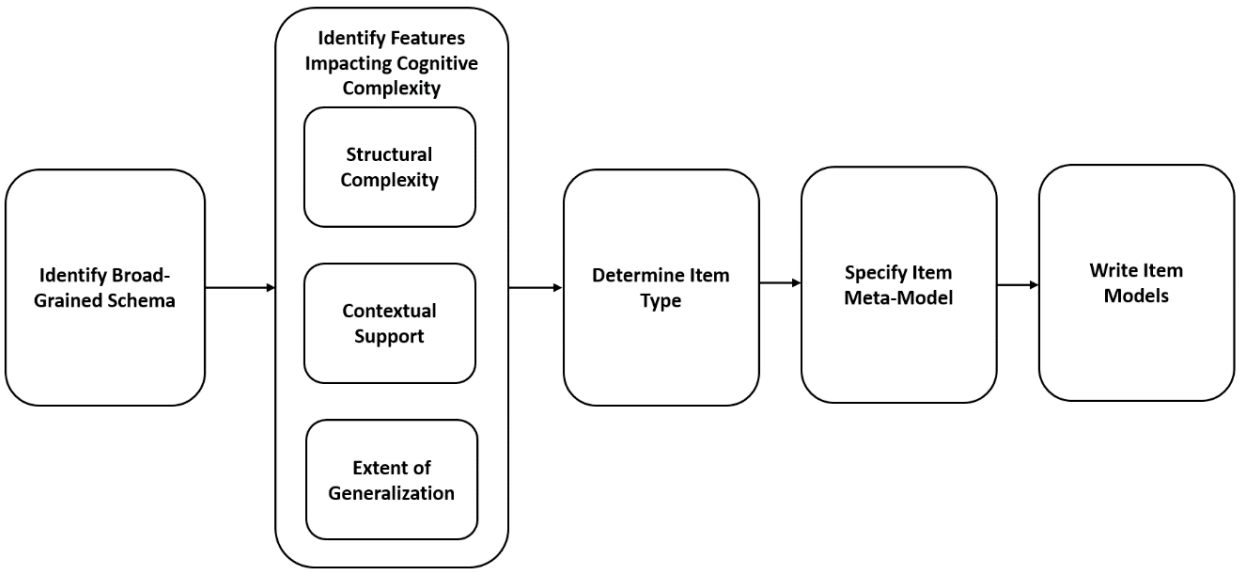

Figure 1. The Item Model Creation Process – Algebra.

schema to clarify that the schema represents a group of fine-grained schemas with a common theme. Often, the broad-grained schema is similar to curricular unit plans or textbook chapters except it can span multiple grade levels with increasing complexity. The chosen schema will depend on the content of the items the test developer wants to generate, which are likely determined by test specifications.

In practice, test developers can accomplish step one by examining test specifications and determining which skills share common characteristics. For example, solving one-step, two-step, and multi-step equations could all fall into the same schema because solving equations relies on similar cognitive processes regardless of the number of steps involved. Similarly, as another example, graphing linear equations and graphing quadratic equations may also group well together as one schema, again because the cognitive processes involved in graphing are similar regardless of the type of function. Depending on the granularity of test content, a test may include a single schema or several schemas.

3.2 Step 2: Identify Features Impacting Cognitive Complexity of Broad-Grained Schema

The next step draws on existing research and theory to identify features impacting the cognitive complexity of tasks within the broad-grained schema. The IMCP-A includes three categories to consider when identifying key features – structural complexity, contextual support, and extent of generalization – which stem from the previously summarized research on algebraic task complexity. At this stage, cognitive models, learning trajectories, empirical research, and theory are useful sources for identifying features impacting cognitive complexity.

3.2.1 Structural Complexity

The first category of consideration is structural complexity. The cognitive complexity of a mathematics task varies based on the complexity of the mathematical structures used in the task. Star and colleagues (2015) define structure as:

> The underlying mathematical features and relationships of an expression, representation, or equation. Structure includes quantities, variables, operations, and relationships (including equalities and inequalities). More complex structures are built out of simple ones, and recognizing structure involves the ability to move between different levels of complexity complexity. (p. 6).

This definition notes that structure includes many different components of the task. The exact elements of structure will vary depending on the specific mathematical content. Structural complexity for the purpose of AIG may include, but is not limited to:

- The number system (e.g., all real numbers, natural numbers, unit fractions, decimals rounded to the nearest tenth, variables only, variables with numbers);
- Relational complexity, defined as the "processing load imposed by interacting components of a task" and quantified based on how many relationships the student must simultaneously track to complete a task (Halford, Wilson, & Phillips, 1998, p. 805); and
- Complexity of operations (e.g., exponents, division).

Table 1 shows an example of structural complexity features related to the broad-grained schema of identifying relationships in bivariate tables.

3.2.2 Contextual Support

The second category of features impacting cognitive complexity is contextual support. I define contextual support as the degree to which the presence of context aids or hinders a student in solving a problem. As I have illustrated by reviewing the work of Koedinger, Nathan, Enright, and others, context matters greatly as a predictor of item difficulty. In practice, item writers often adhere to specifications that merely indicate whether an item should be a word problem or not, with little attention given to the spectrum of possibilities for incorporating context. Attention to context is particularly important for middle school tests because of the research summarized above indicating that context plays a greater role with younger, mathematically developing students than it does for adults that are already mathematically proficient.

Providing contextual support does not mean phrasing an item as a story problem, and the presence of context does not necessarily mean that the item is easier or harder. Still, the type of context impacts item difficulty, either by making items easier or more difficult depending on the

Table 1: Structural Complexity Features Impacting Cognitive Complexity

Number System	Relational Complexity	Complexity of Operations
Values in table can include: • Natural numbers • Integers • Positive fractions and decimals • Positive and negative fractions and decimals	The number of operations completed on x in order to obtain $f(x)$. For example: • $f(x) = 3x - 1$ requires 2 operations (multiply x by 3 and subtract 1) • $f(x) = 3x^2 - 1$ requires 3 operations (square x, multiply by 3, and subtract 1)	• Proportional, $f(x) = 3x$ • Linear, $f(x) = 3x\text{-}1$ • Quadratic, $f(x) = 3x^2$ • Exponential, $f(x) = 3x$ • Logarithmic, $f(x) = log_3(x)$

student and the other features of the item. Contextual support can be thought of as a continuum ranging from meaningful context that aids in students' interpretation of the problem to purely abstract situations with no real-world connections. Figure 2 shows an example of a continuum of contextual support for bivariate tables, ranging from providing the student with a relatable, familiar context (i.e., time and distance) to using abstract function notation. In this case, the level of contextual support varies by the type of header in the table. Again, I do not claim that this continuum is equivalent to a continuum of item difficulty – in fact, it is probably not the case that items progress from easiest to hardest along the continuum and rather the case that context interacts with other features of the item and student age and ability to predict difficulty. The purpose of the continuum is to illustrate how an item can vary in the amount of context.

3.2.3 Extent of Generalization

The third category of features impacting cognitive complexity is extent of generalization, defined as the amount of generalizing and formalizing across relationships, properties, and patterns required by the student in order to correctly answer the item. Because generalizing and formalizing relationships represent the core of algebraic reasoning, extent of generalization is included in the IMCP-A in order to produce items that will solicit evidence to locate a student on a trajectory of algebraic reasoning development. Explicitly varying items in terms of the amount of generalization results in items that will represent a range of abilities pertaining specifically to algebraic reasoning.

Figure 3 presents an example of a continuum of extent of generalization, again for the broad-grained schema of identifying relationships in bivariate tables. The lowest

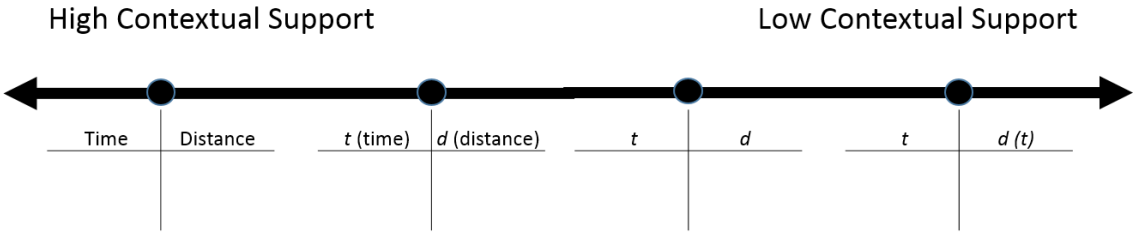

Figure 2. An example of a continuum of contextual support for algebra items.

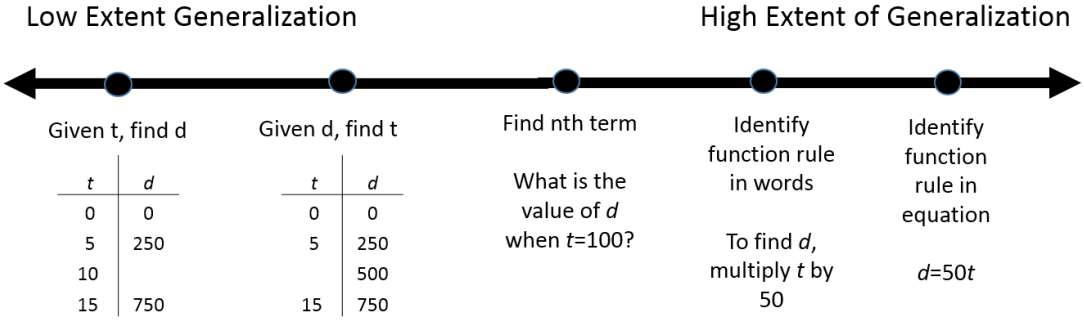

Figure 3. An example of a continuum of extent of generalization for algebra items.

level of generalization (e.g., the left-most item) is a computational example. The student must only identify a pattern for one particular case: identifying the value of *d* when *t* equals 10. The student could accomplish this in several ways, some of which do not require identifying the relationship between *t* and *d*, such as recognizing that *d* increases by 250 in each row of the table. The item in this case, then, is nearly equivalent to finding the missing number in the pattern 0, 250, ___, 750 and requires no generalizing. Although it is true that a student might solve this by identifying the general relationship between *t* and *d*, the key idea here is that the item does not require the student to do so. In other words, the student could use a variety of non-generalizing strategies to successfully answer the item. The second item on the continuum is similar, although slightly more complex because it is phrased as the inverse of typical table problems. Still, a student could similarly solve this item by only identifying the missing term in 0, 5, __, 15.

In the third item, the student is given a table and must identify a term outside the bounds of the presented table. The student could solve this item in several ways, for example by extending the pattern one row at a time or making a number-specific generalization about the relationships (e.g., as *t* increases by 15, *d* increases by 750) to more efficiently arrive at the solution. The magnitude of *n* when finding the *n*th term will also impact the students' need to generalize in this case, which can move this item's position along the continuum in either direction: a larger *n* will require more generalization and vice versa.

In the last two right-most examples of the continuum, the student has to generalize the pattern for the entire table by specifying a rule applicable to any value of *t*. Writing this rule in an equation or function notation represents a greater degree of formalization as compared to writing the rule in words and therefore is farther along the extent of generalization continuum. Determining where tasks fall on the continuum of extent of generalization should be guided with cognitive models and empirical evidence of students' thinking as they solve problems.

Test developers may specify structural complexity, contextual support, and extent of generalization by examining existing content (e.g., reviewing curricular resources or retired test items). Qualitative coding may be performed on existing content for each of the three categories – structural complexity, contextual support, and extent of generalization – in order to determine how content might vary within each of those categories. Additionally, literature reviews of features affecting task difficulty, as was presented earlier for a range of algebraic topics, may be conducted specifically for the targeted schema to gain insight about how task features may vary.

3.3 Step 3: Determine Item Type

The next step, after identifying factors impacting cognitive complexity, is to choose the item type. Common item types include multiple-choice, grid-in, constructed response, and technology-enhanced items. The item type should align with the intended purpose of the item while also considering resource limitations such as access to technology or the ability to score constructed-response items. Another consideration when choosing the item type is that certain item formats can affect the cognitive complexity of features of an item; for example, a multiple-choice item requiring a student to solve an equation invites the strategy of substituting the answer choices into the equation whereas a grid-in response does not. A final point about the item type is that, although tests often include multiple item-types on a single form, the IMCP-A should only be applied to one item type at a time because item generation methods will vary based on the type of item. Test developers may choose item type based on operational constraints (e.g., no computers available to administer technology-enhanced item) or may consider other aspects such as providing constructed-response items to elicit more information about student reasoning. Test developers wishing to use multiple item-types can repeatedly apply the IMCP-A as needed.

3.4 Step 4: Specify Item Meta-Model

The fourth step in the IMCP-A is specifying the item meta-model, which specifically aids in writing item models. The item meta-model is a model for the item model – it describes the structure of the item models and how the structures will vary based on the identified features of cognitive complexity. To specify the item meta-model, test developers should holistically consider the overall intent of the content standard aligned to the desired generated items and the main idea of that content standard. The item meta-model is then a generic representation of how to ask a student about the main idea captured in the content standard. Figure 4 shows an example of an item meta-model for the schema of identifying relationships in bivariate tables, with gray-highlighted portions of the item meta-model representing components that will

[Introductory Content] The table shows the relationship between [input] and [output]

[input]	[output]
$[x_1]$	$[y_1]$
$[x_2]$	$[y_2]$
$[x_3]$	$[y_3]$
$[x_4]$	$[y_4]$

[Question]

[Distractors and Key]

Figure 4. Item meta-model for identifying relationships in bivariate tables.

vary across item models. In this example, the main idea is finding relationships in tables, thus the item meta-model includes generic language regarding a relationship between an input and output along with a visual of a table since tables are critical to the content standard. The remainder of the item meta-model leaves flexibility for varying aspects of the items, such as the specific question asked, and the answer choices used.

As the figure illustrates, the item meta-model in this example conveys many aspects of the item models: 1. Items must include a table with four inputs and outputs, 2. The structure is multiple choice, and 3. The phrase "The table shows the relationship between ___ and ___" is present in all item models, etc. Item meta-models may also include a list of common student misconceptions and errors as related to the broad-grained schema that can then become distractor rationales. Explicitly identifying the structure of the item models upfront allows for a methodological process of writing item models while still allowing enough freedom in item models to produce heterogeneous items.

The item meta-model approach has two additional benefits. First, beginning with an item meta-model allows for generation of more diverse items as compared to approaches that start with item models because the item meta-model creates a broader range of item models from which items are then generated. In other words, the item meta-model supplies an additional layer of item variation by implementing a process of *item meta-model* → *item model* → *item* instead of just *item model* → *item*. Second, since creation of the item meta-model is based on critical features of algebra cognitive complexity (i.e., structural complexity, contextual support, and extent of generalization), generation of items using the item meta-model approach will likely result in items with high correlations between predicted and actual difficulties.

3.5 Step 5: Write Item Models

Item meta-models are not in a programmable format, yet AIG relies on creating items from computer programs. The next step is to use the item meta-model to create item models in sufficient detail for computer interpretation. This step entails listing out combinations of features from step two and then determining which combinations may work well together to form an item model. For example, Table 2 illustrates three combinations derived from combing the features from the continuums in Figure 2 & 3 along with structural complexity features. In these combinations, Combination A and B would produce an item model that is appropriate to the content standard, yet Combination C, which requires a proportional relationship with three unique operations, cannot mathematically happen due to the definition of a proportional relationship. Thus, Combination C can be ruled out as a feasible item model.

The item model writer uses the item meta-model to vary elements of the item models according to contextual support, structural complexity, and extent of generalization specified for the particular desired combination of features.

Table 2. Combinations of Features Used to Create an Item Model

Combination	Number System	Relational Complexity	Complexity of Operations	Contextual Support	Extent of Generalization
A	Natural numbers	Two operations	Linear	Low	High
B	Natural numbers	One operation	Exponential	High	Low
C	Integers	Three operations	Proportional	Low	Low

Note: Combination C is shown only as an example that would not work as an item model because it is mathematically impossible to create a proportional relationship with three distinct operations.

In the item meta-model example, the first element of the item meta-model is the introductory content. An item model could omit introductory content altogether or include phrases such as "Sarah is taking a road trip to the beach and tracks her driving time in a table."

The item meta-model specifies that the next sentence starts with "The table shows the relationship between…," but the item model varies how the input and output are specified. For the example item meta-model, an input/output might be *time driving/distance traveled* or *t/ d*, which varies contextual support based on inclusion or exclusion of variables. The next part of the item model includes the values of the table, which may include specifying if one value is missing or not depending on the extent of generalization. Next, different questions can be asked, again by varying the extent of generalization, such as "How far has Sarah traveled after 8 hours of driving?" or "Which function represents the relationship in the table?". Lastly, the item model indicates the correct answer and distractors based on common student errors and misunderstandings (for multiple-choice item types).

The table shows the relationship between *x* and *y*.

x	*y*
$[x_1]$	$[a][x_1] + b$
$[x_2]$	$[a][x_2] + b$
$[x_3]$	$[a][x_3] + b$
$[x_4]$	$[a][x_4] + b$

Which equation represents the relationship in the table?

A) y = [a]*x*
B) y = *x*+[b]
C) y = [a]*x*+[b]
D) y = [b]*x*+[a]

Distractor Rationales:

A) Student omitted constant term in equation.
B) Student omitted slope in equation.
C) Key
D) Student interchanged constant and rate.

Specification of elements:

$a, b \in \{N : 2 \le a, b \le 12\}$

$x_1 \in \{N : 2 \le x_1 \le 10\}, x_{i+1} = x_i + k \; where \; k \in N, k \le 10$

Figure 5a. Item model from item meta-model.

Figure 5a and 5b portray two item models created from the same item meta-model and aligned to Combination A and B from Table 2, respectively, although many other possibilities exist. The item model in Figure 5a represents minimal structural complexity and contextual support because it uses whole numbers with a linear relationship and no real-life context to aid interpretation or solving. The item requires a fair amount of generalization, however, because the student must identify an equation as opposed to finding a missing value in the table. The item model in Figure 5b is high in contextual support

The [content word 1] in [content word 2] [grows OR decays] over time. The table shows the relationship between time and [content word 1].

Time	[content word 1]
$[x_1]$	$[a]^{[x_1]}$
$[x_2]$	$[a]^{[x_2]}$
$[x_3]$	$[a]^{[x_3]}$
$[x_4]$	?

Which is the missing value in the table ?*

A) $[x_4] + ([a]^{[x_2]} - [a]^{[x_1]})$
B) $[a]^{[x_4]}$
C) $[a]^{[x_3+2]}$
D) $[x_4] * [a]^{[x_1]} / [x_1]$

*After calculating distractors and key, reorder from least to greatest.

Distractor Rationales:

A) Student identified change in first two y values and added that change to the third y value.
B) Key
C) Student calculated one power more than needed.
D) Student identified the multiplicative relationship in the first row of the table and applied that relationship to $[x_4]$.

Specification of elements:

(*content word 1, content word 2, grows or decays*) ∈ {(*bacteria, food, grows*), (*money,bank accounts, grows*), (*toxins, chemicals, decay*), (*population, cities, grows*)}

$x_1 \in \{N : 1 \le x_1 \le 10\}, x_{i+1} = x_i + 1$

$a \in \{2, 3, 4\}$

Figure 5b. Item model from item meta-model.

because it includes a real-life context. It has a moderate level of structural complexity (i.e., whole numbers with an exponential relationship) and low level of extent of generalization because the student must identify a missing value as opposed to writing an equation.

These are merely two examples of the possibilities for item models developed from the item meta-model. As another example, the [Question] part of the item meta-model could be inserted with "Which type of relationship is shown in the table?" with fixed answer choices: A) linear, B) quadratic, C) exponential, or D) logarithmic. Alternatively, the [Introductory Content] could provide the student with the equation represented in the table as an additional layer of scaffolding and ask the student to find a specific value of the table. The item meta-model is a critical component of the IMCP-A because it promotes diversity of item models, which results in diversity of generated items.

The last step of AIG is to programmatically generate items from item models, methods for doing so are already well-established and beyond the purpose of the IMCP-A. During item generation, variables of the item meta-model and item model should be tracked in order to potentially pre-align items to content standards, as it is most likely the case that items generated from a single item meta-model align to different content standards in more than one grade level. Examples of generated items from the item models in Figure 5a and 5b are presented in Figure 6.

Although the steps of the IMCP-A were presented linearly here, users should recognize that – as with any item development process – new information in later stages may require revisions to earlier stages. This may particularly occur with respect to identifying features in step two that yield the most content-appropriate items. After attempting to write item meta-models and item models, it may be necessary to revise the initial features specified in step two. As another example, a particular item type may seem reasonable in step three, but attempts to write item models of that particular type may prove difficult. The need to revisit an earlier step could occur at any stage of the process, and test developers should plan for such fluidity in the item development process.

4. Methods for Evaluating the IMCP-A

The IMCP-A currently has not undergone a validation or evaluation study to determine the efficacy of the

The table shows the relationship between x and y.

x	y
2	8
4	14
6	20
8	26

Which equation represents the relationship in the table?

A) $y = 3x$
B) $y = x+2$
C) $y = 3x+2$
D) $y = 2x+3$

The bacteria in food grow over time. The table shows the relationship between time in hours and bacteria in millions.

Time	Bacteria
1	2
2	4
3	8
4	?

Which is the missing value in the table?

A) 8
B) 10
C) 16
D) 32

Figure 6. Items generated from item models.

IMCP-A as a test-development process for creating high-quality automatically-generated items. As such, future evaluation and validation is necessary. Since the purpose of the IMCP-A is to create large quantities of operational items on tests that produce valid scores of middle school students' algebraic reasoning, evaluating the IMCP-A necessitates evaluating the items produced with the IMCP-A both at the item level (i.e., evaluating each item independently from other items) and holistically across sets of generated items (i.e., evaluating multiple items together as if they were on a single test form). I describe

three main areas of consideration for evaluating the IMCP-A and the items produced from it: 1. Validity of scores obtained from automatically-generated items, 2. Heterogeneity of items, and 3. Instructional relevance of tests that contain generated items. Methods for evaluating the IMCP-A within each category follow.

4.1 Validity

In order to support a validity argument, the features used in item generation as obtained from the IMCP-A should account for a large percentage of variation in item difficulty. Highly-successful AIG models of item features have shown results with as much as 90% of variance in item difficulty explained by item features used in generation (Enright et al., 2002). It is important that an AIG model explains a large percent of variance in item difficulty in order to provide evidence that the model captures important construct-relevant contributors to item difficulty.

Many methods exist for modeling item difficulty with item features. A few popular methods include the linear logistic test model (Fischer, 1973), constrained two-parameter logistic model (Embretson, 1999), identical siblings model (Hombo & Dresher, 2001), related siblings model (Glas & van der Linden, 2003), linear item cloning model (Geerlings, Glas & van der Linden, 2011), IRT family modeling (Daniel & Embretson, 2010), and hierarchical IRT model (Embretson & Daniel, 2008). A limiting factor of most of these models is that they require raw item response data for each student's response to each item as opposed to aggregated item difficulty statistics. Sinharay and Johnson (2013) provide a review of several of these methods.

Standard regression techniques can also be applied to item data and are particularly useful when only aggregated item statistics are available. In the simplest case, multiple regression can be applied with item features as independent variables and item difficulty as the dependent variable. Another technique is tree-regression (also known as random forest regression), which makes it possible to identify interaction effects between item features without necessarily specifying interactions prior to running the model (Enright et al., 2002). A limitation of regression methods is that they produce greater standard errors as compared to methods designed specifically for applications to item data (Embretson & Daniel, 2008).

An additional dependent variable to consider is response time (i.e., the number of seconds an examinee spends responding to an item). Although response time data are typically less readily-available, response time is an indicator (with some limitations) of the cognitive complexity of the item because it serves as a proxy measurement for the amount of processing the examinee had to do in order to answer the item. Longer response times are typically, but not always, associated with more difficult items and vice versa. Thus, researchers can model how features used in item generation predict response time (Embretson & Yang, 2006). When response time data are used, a common approach is to model the log of response time in order to make a more normal distribution (Daniel & Embretson, 2010). Usually, when modeling response time, a hierarchical linear model must be used to account for the data structure whereby a single student responds to multiple items (i.e., items nested in students).

All models of item features should be evaluated alongside theory and expected results. As a caution, Bejar (2010) reminds test developers that statistically non-significant item features should not necessarily be removed from item generation parameters. In these cases where models show counterintuitive effects of item features, meaningful reasons for the discrepancy may exist. For example, Kosh, Simpson & Bickel (2015) found that upper-elementary students did *not* respond statistically significantly faster to addition fluency items as compared to subtraction items, despite cognitive theory that subtraction problems are more difficult for students than addition problems (Fuson, 1984). A plausible explanation for the finding is that the sample included high-ability students who likely already memorized addition and subtraction facts, thereby returning equivalent response times for any number fact regardless of the operation. This point illustrates that, before making decisions to remove variables from item generation, researchers should investigate alternate explanations and conduct additional research when theoretical predictions about item difficulty and empirical data conflict. It is especially important that researchers consider the amount of instruction students have received on the topic and students' ability levels before making conclusions.

4.2 Heterogeneity

A second area of evaluation for the IMCP-A is its ability to produce diverse items. A concern for automatically-generated items is a lack of item diversity both within and

across skills and concepts. Often, items generated from the same item model look similar. Test users criticize these items for lack of heterogeneity and call them "clones," "ghost" items, or "Franken-items" (Gierl & Lai, 2016). Items that appear too similar to each other are susceptible to coaching for test preparation and memorization by examinees (Gierl & Lai 2013; Mayer 1981). On the other hand, when items are sufficiently different, coaching is equivalent to learning because the items represent a range of formats that assess knowledge of a concept or skill (Embretson, 2002). Additionally, as Enright and Sheehan (2002) point out to again illustrate the importance of item heterogeneity, "an undesirable consequence [of AIG] would be to restrict the kinds of skills assessed because it's easier to automatically generate some types of problems than others" (p. 156-157).

These two unintentional results of AIG would indeed be unfortunate and raise the need to evaluate the diversity of generated items. One such method for measuring the heterogeneity of items is the cosine similarity index (CSI; Gierl & Lai, 2016). The CSI uses techniques similar to latent semantic analysis by creating vectors of word and number frequencies in each item stem and then computing the cosine of the angle between those vectors as a measure of vector proximity within a vector space (Landauer, McNamara, Dennis & Kintsch, 2013). The resulting CSI is a quantitative measure ranging from zero to one of the similarity of words and numbers in an item's stem; stems with low CSIs (i.e., fewer similar words and numbers) represent greater heterogeneity of items. In Gierl and Lai's (2016) generated items, CSIs ranged from 0.20 to 0.66. The field of AIG has no current recommendations for acceptable or "good" CSI cosines, but the CSI can be used to compare different two AIG methods in terms of item heterogeneity. Thus, CSI cosines from items generated with the IMCP-A could be compared to CSI cosines from other item generation methods to determine which methodology produces more diverse items.

A second method of assessing item heterogeneity is to experimentally test if subject matter experts can distinguish between generated items and hand-written items (Gierl & Lai, 2016). In this method, researchers provide subject matter experts with a batch of items, half of which are hand-written items and the other half of which are generated items. Then, content experts are asked to identify whether an item was automatically generated or not. When subject matter experts accurately classify less than fifty percent of items as generated or hand-written, results imply that items cannot be detected as clones.

4.3 Instructional Relevance

The final evaluation category is the instructional relevance of scores obtained from tests with generated items. As Leighton and Gierl (2011) highlighted, given all the money that is spent on testing, "it seems reasonable to ask whether [tests] will provide additional information about student learning aside from their performance at a single point in time, including partial knowledge and skills, misconceptions, and areas of genuine strength" (p. 6-7). Additionally, the U.S. Department of Education's (2015) *Testing Action Plan* called for seven principles leading to fewer and smarter assessments, one of which is that tests should be "tied to improve learning." As the report describes, "In a well-designed testing strategy, assessment outcomes are not only used to identify what students know, but also inform and guide additional teaching, supports, or interventions that will help students master challenging material" (U.S. Department of Education, 2015). Another term for this is instructional validity, which is the applicability of tests for informing instructional decisions (Penuel, Confrey, Maloney & Rupp, 2014).

Since one goal of assessment is to gain useful information about students that can be used to inform instructional decisions, the IMCP-A should also be evaluated in terms of its ability to produce test scores that provide instructionally-relevant information about what students can and cannot do instead of merely classifying them as basic, proficient, or advanced. Instructional relevance could apply to a variety of stakeholders beyond teachers, such as district leaders and policy makers. In fact, since the IMCP-A mostly applies to large-scale testing, it is unlikely that teachers will have information from tests that they can use to inform day-to-day lesson and unit planning. However, large-scale tests can provide actionable instructionally-relevant information on a broader level, for example by identifying themes across students and skills that highlight gaps in curriculum, grade levels of concern, or teaching weaknesses and strengths. One such application could be identifying that students using a particular curriculum do well with complicated numbers (e.g., decimals, fractions, and negative numbers) but perform poorly on expressing generalized relationships. This information might provide

insight about a curriculum that emphasizes drill-like practice rather than meaningful, conceptually-rich tasks.

The instructional relevance of a test depends not only on the design and content of the test but also largely on how scores are reported. A full discussion of score reporting is beyond the scope of this paper, but AIG does lend itself to more detailed score reports because items include systematic variation of key features, which allows data to illustrate themes of student performance on multiple levels. Assuming quality score reports exist and assuming information from tests will be analyzed at the large-scale level, researchers could evaluate the instructional relevance of scores by observing district and state planning meetings related to designing curriculum, training teachers, funding interventions, and the like. A researcher observing a meeting where score reports serve as a source of data would be able to answer the research question: when provided with score reports of tests, how do stakeholders use that information to guide their decisions? Researchers could see if policymakers use score reports to demonstrate deficiencies, highlight strengths, or target particular areas of need.

5. Conclusion

As AIG becomes more heavily used, test developers will write item models as opposed to writing items. I have presented a process for creating item models to generate items that measure middle school students' algebraic reasoning, named the Item Model Creation Process - Algebra. The IMCP-A facilitates a science of item model development so that item models – and subsequently items – are based on the science of learning, which will ultimately result in test scores with more meaningful and interpretable information about what students know and can do.

The IMCP-A guides test developers through a five stage process for developing item models for AIG: 1. Identifying a broad-grained schema, 2. Outlining features impacting the cognitive complexity of tasks within the schema in terms of structural complexity, contextual support, and extent of generalization, 3. Determining the item type, 4. writing an item meta-model to outline the structure of item models, and 5. Writing item models from the item meta-model. To my knowledge, this process is the first to explicitly address a process for creating item models and is also the first to present a domain-specific approach to AIG by focusing on items that assess middle school students' ability to reason algebraically. Next steps for this work include applying the IMCP-A in practice and evaluating the items produced from it with respect to validity and instructional relevance of score inferences obtained from tests using generated items and heterogeneity of generated items. Finally, I encourage others to develop additional item model creation processes in other content areas, as this paper only described the case of middle school algebraic reasoning.

6. References

Baranes, R., Perry, M. & Stigler, J. (1989). Activation of real-world knowledge in the solution of word problems. *Cognition and Instruction, 6(4)*, 287–318. https://doi.org/10.1207/s1532690xci0604_1.

Bassok, M. C. & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology, 35,* 99–134. https://doi.org/10.1006/cogp.1998.0675. PMid: 9570897.

Bejar, I. I. (2010). Recent development and prospects in item generation. In: S. Embretson (Ed.), *Measuring psychological constructs (pp. 201-226). Washington, DC: American Psychological Association*. https://doi.org/10.1037/12074-009.

Bickel, L., Simpson, M. A., Sanford-Moore, E., Price, R., Durakovic, L., Totten, S. & Kosh, A. E. (2015). Family group generation theory: Large scale implementation. *Paper presented at the 2015 National Council on Measurement in Education Conference*, Chicago, IL.

Blanton, M. L. & Kaput, J. J. (2005). Characterizing a classroom practice that promotes algebraic reasoning. *Journal for Research in Mathematics Education, 36(5),* 412–446.

Carraher, T. N., Carraher, D. W. & Schliemann, A. D. (1985). Mathematics in the streets and schools. *British Journal of Developmental Psychology, 3,* 21–29. https://doi.org/10.1111/j.2044-835X.1985.tb00951.x.

Clement, J. (1982). Algebra word problem solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education, 13(1),* 16–30. https://doi.org/10.2307/748434.

Common Core State Standards Initative (2015). *Standards for Mathematical Practice*. Retrieved from http://www.corestandards.org/Math/Practice/.

Daniel, R. C. & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement, 34(5),* 348–364. https://doi.org/10.1177/0146621609349801.

Drasgow, F., Luecht, R. M. & Bennett, R. E. (2006). Technology and testing. In: Brennan, R. L. (Ed.), *Educational*

Measurement, Westport, CT: Praeger Publishers. p. 471–516.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433. https://doi.org/10.1007/BF02294564.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In: Irvine, S. H. & Kyllonen, P. C. (Eds.), *Item generation for test development*. Mahwah, NJ: Erlbaum. p. 219-250.

Embretson, S. E. & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50(3),* 328.

Embretson, S. & Yang, X. (2006). Automatic item generation and cognitive psychology. *Handbook of Statistics, 26*, 747–768. https://doi.org/10.1016/S0169-7161(06)26023-1.

Enright, M. K. & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In: S. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. p. 129–157.

Enright, M. K., Morley, M. & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15(1),* 49–74. https://doi.org/10.1207/S15324818AME1501_04.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37(6),* 359–374. https://doi.org/10.1016/0001-6918(73)90003-6.

Fuson, K. (1984). More complexities in subtraction. *Journal for Research in Mathematics Education, 15,* 214–225. https://doi.org/10.2307/748350.

Geerlings, H., Glas, C. A. & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika, 76(2),* 337–359. https://doi.org/10.1007/s11336-011-9204-x.

Gierl, M. J. & Haladyna, T. M. (2013). Automatic item generation: An introduction. In: Gierl, M. J. & Haladyna (Eds.), *Automatic Item Generation: Theory and Practice*. New York: Routledge. p. 3-12.

Gierl, M. J. & Lai, H. (2013). Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples. In: Gierl, M. J. & Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York: Routledge. p. 26–39.

Gierl, M. J. & Lai, H. (2016). Automatic item generation. In: Lane, S., Raymond, M., & Haladyna, T. M. (Eds.), *Second Handbook of Test Development*. New York: Routledge. p. 410–429.

Gierl, M. J., Lai, H., Hogan, J. B. & Matovinovic, D. (2015). A method for generating educational test items that are aligned to the common core state standards. *Journal for Applied Testing Technology, 16(1),* 1-18.

Gierl, M. J., Zhou, J. & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning and Assessment, 7(2)*.

Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 247–261. https://doi.org/10.1177/0146621603027004001.

Goldman, S. & Booker, A. (2009). Making math a definition of the situation: Families as sites for mathematical practices. *Anthropology and Education Quarterly, 40(4),* 369–387. https://doi.org/10.1111/j.1548-1492.2009.01057.x.

Gorin, J. S. & Embretson, S. E. (2013). Using cognitive psychology to generate items and predict item characteristics. In: Gierl, M. J. & Haladyna (Eds.), *Automatic Item Generation: Theory and Practice*. New York: Routledge. p. 136-156.

Graf, E. A. & Fife, J. H. (2013). Difficulty modeling and automatic generation of quantitative items: Recent advances and possible next steps. In: Gierl, M. J. & Haladyna (Eds.), *Automatic Item Generation: Theory and Practice*. New York: Routledge. p. 157–179.

Graf, E. A. (2009). Defining mathematics competency in the service of cognitively based assessment for grades 6 through 8 (ETS Research Report. No. RR-09-42). Princeton, NJ: Educational Testing Service.

Haladyna, T. & Rodriguez, M. (2013). Developing the test item. In: T. Haladyna & M. Rodriguez (Eds.), *Developing and Validating Test Items*. New York: Routledge. p. 17–27. https://doi.org/10.4324/9780203850381.

Halford, G. S., Wilson, W. H. & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21,* 803–865. https://doi.org/10.1017/S0140525X98001769. PMid: 10191879.

Hombo, C. & Dresher, A. (2001). A simulation study of the impact of automatic item generation under NAEP-like data conditions. *Paper presented at the annual meeting of the National Council of Measurement of Education*, Seattle, WA.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development, 54(1)*, 84–90. https://doi.org/10.2307/1129864.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux. PMid: 21714386.

Koedinger, K. R. & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13(2),* 129–164. https://doi.org/10.1207/s15327809jls1302_1.

Koedinger, K. R., Alibali, M. W. & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science, 32(2),*

366–397. https://doi.org/10.1080/03640210701863933. PMid: 21635340.

Kosh, A. E., Simpson, M. A., Bickel, L. (2015). *Validity of task models in automatically generated mathematics items*. Poster presented at the Research Symposium Honoring the Legacy of Carol E. Malloy, Chapel Hill, NC.

Landauer, T. K., McNamara, D. S., Dennis, S. & Kintsch, W. (Eds.). (2013). *Handbook of Latent Semantic Analysis*. Psychology Press. https://doi.org/10.4324/9780203936399.

Lave, J. (1988). *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609268.

Leighton, J. P. & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511996276.

Martin, S. A. & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory & Cognition, 33(3),* 471–478. https://doi.org/10.3758/BF03193064.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10(2),* 135–175. https://doi.org/10.1007/BF00132515.

Mayer, R. E. (1982). Different problem-solving strategies for algebra word and equation problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8(5),* 448–462. https://doi.org/10.1037/0278-7393.8.5.448.

Mayer, R. E., Larkin, J., & Kadane, J. B. (1984). A cognitive analysis of mathematical problem solving ability. In: R. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence,* Hillsdale, NJ: Lawrence Erlbaum. Vol. 2, p. 231–273.

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A brief introduction to evidence-centered design (ETS Research Report. No. RR-03-16).* Princeton, NJ: Educational Testing Service.

Nathan, M. J. & Koedinger, K. R. (2000). Teachers' and researchers' beliefs about the development of algebraic reasoning. *Journal for Research in Mathematics Education, 168-90*. https://doi.org/10.2307/749750.

Nathan, M. J., Long, S. D. & Alibali, M. W. (2002). Symbol precedence in mathematics textbooks: A corpus analysis. *Discourse Processes, 33,* 1–21. https://doi.org/10.1207/S15326950DP3301_01.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors.

O'Rourke, E., Andersen, E., Gulwani, S. & Popovic, Z. (2015, April). A process for automatically generating interactive instructional scaffolding. In: Begole, B. & Kim, Jinwoo (Eds.), *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery. p. 1545–1554.

Penuel, W. R., Confrey, J., Maloney, A. & Rupp, A. A. (2014). Design decisions in developing learning trajectories–based assessments in mathematics: A case study. *Journal of the Learning Sciences, 23(1),* 47–95. https://doi.org/10.1080/10508406.2013.866118.

Saxe, G. B. (1988a). Candy selling and math learning. *Educational Researcher, 17(6),* 14–21. https://doi.org/10.3102/0013189X017006014.

Saxe, G. B. (1988b). The mathematics of child street vendors. *Child Development, 59(5),* 1415–25. https://doi.org/10.2307/1130503.

Sebrechts, M. M., Enright, M., Bennett, R. E. & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction, 14(3)*, 285–343. https://doi.org/10.1207/s1532690xci1403_2.

Simpson, M. A., Kosh, A. E., Bickel, L., Elmore, J., Sanford-Moore, E., Koons, H. & Enoch-Marx, M. (2015, April). The effects of varying grain size on the exchangeability of item isomorphs. *Paper presented at the 2015 National Council on Measurement in Education Conference*, Chicago, IL.

Singley, M. K. & Bennett, R. E. (2002). Item generation and beyond: Applications to schema theory to mathematics assessment. In: Irvine, S. H. & Kyllonen, P. C. (Eds.), *Item Generation for Test Development*. Mahwah, NJ: Erlbaum. p. 361–384.

Sinharay, S. & Johnson, M. (2013). Statistical modeling of automatically generated items. In: Gierl, M. J. & Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York: Routledge. p. 183–195

Star, J. R., Foegen, A., Larson, M. R., McCallum, W. G., Porath, J. & Zbiek, R. M. (2015). *Teaching strategies for improving algebra knowledge in middle and high school students.* U.S. Department of Education.

Taylor, E. V. (2009). The purchasing practice of low-income students: The relationship to mathematical development. *Journal of the Learning Sciences, 18(3),* 370–415. https://doi.org/10.1080/10508400903013462.

U.S. Department of Education (2015). *Fact sheet: Testing action plan*. Retrieved from https://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20(2)*, 79–87. https://doi.org/10.1016/j.pse.2014.11.003.