Pearson

Test Design Overview

TRANSCEND

Transcend™ Test Design Overview

Interim (or benchmark) assessments are an integral part of any comprehensive, learning-based assessment system. They fit between classroom assessments that are used to inform day-to-day instruction and end-of-year assessments that are used for federal accountability. Depending upon the specific design, interim assessments are often administered quarterly. This means they judge students' accumulated knowledge and skills between 6 to 10 week intervals of classroom instruction. Interim assessments produce scores to describe the *status* (description at a specific point in time), and/or the *growth* (descriptions of change over time) of students' accumulated knowledge and skills throughout the academic year. These scores are intended to support insights into short- and long-term learning goals, as well as predict how students are likely to perform on the end-of-year state test.

Figure 1 illustrates the hierarchy of the Grade 3 Common Core State Standards for Mathematics. At the top of the hierarchy is the subject, in this case, *mathematics*. Next, the subject splits into five domains: *operations & algebraic thinking*, *numbers and operations in base ten*, *numbers & operations—fractions*, *measures & data*, and *geometry*. Each domain divides into one or more clusters, which further divide into standards. Standards outline what every student should know and be able to do at the end of each grade, and therefore, are fundamental in the development of curriculum. Thus, it only makes sense that the standards hierarchy serve as a framework for relating the *construct* targeted for measurement by an assessment within a comprehensive, learning-based assessment system.

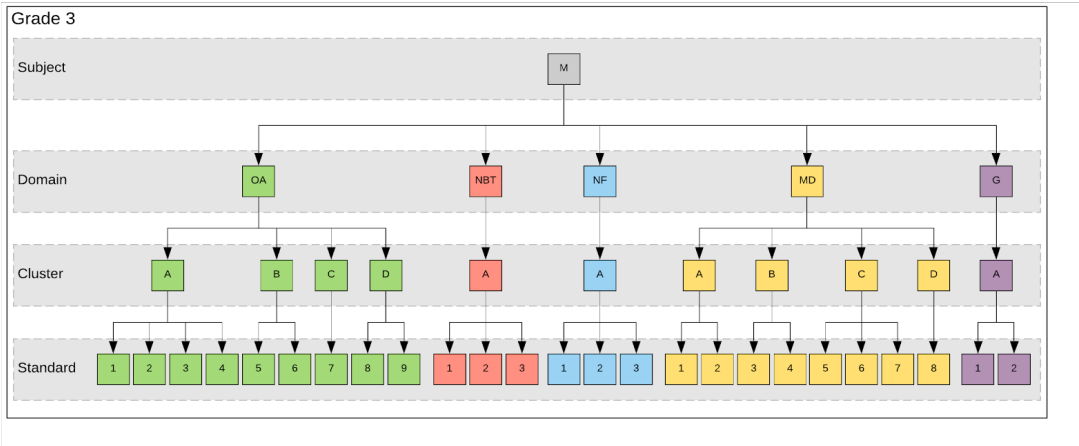

Table 1. Standards hierarchy for Grade 3 Common Core Mathematics

Classroom assessments are intended to inform day-to-day instruction. They are designed to target a construct at the smallest grain-size, at the standard-level, or possibly only a component of the standard[1].

An end-of-year state assessment, on the other hand, is designed to assess how well students understand all grade-level standards. Due to the amount of information covered in an academic year and time constraints, the target construct to be measured by an end-of-year assessment is of a large grain-size—the subject-level and possibly the domain-level of the hierarchy. The amount of time and instruction between interim administrations requires the construct targeted by the assessment to be of a grain-size similar to an end-of-year test. That is, the amount of instruction covered between interim administrations would require an exceptionally long test to ensure sufficient coverage of each standard-level construct targeted by the assessment and to report associated scores with sufficient reliability. Likewise, too much time passes between interim administrations for the scores to be useful for driving day-to-day instructional decisions. To learn that a student struggled to understand a specific concept addressed six weeks earlier is too long.

Interim assessments have the potential to provide educators with a critical data source for describing how well students are learning; the quality of curriculum or programs; and identifying those teachers and schools who need additional support, or those who may have promising practices to share. Interim assessments can provide data to drive immediate action to improve subsequent teaching and learning.

A fundamental requirement for interim scores to serve this capacity is that the interim assessment design must match the richness and depth of the goals for student learning and be well aligned to the curriculum. This document walks through many features of commercial interim assessment designs, specifically those that employ adaptive test delivery, and describes the extent to which these features fulfill such requirements.

Unidimensional Adaptive Test with Grade Band Item Bank

Common commercial adaptive interim assessments employ an adaptive algorithm that selects items from an item bank that includes items aligned to standards across multiple grades, referred to as a *grade band*. For example, the bottom of Figure 2 illustrates a Common Core Mathematics 3-5 grade band item bank, which means the items in the bank are aligned to Grades 3-5 Common Core Mathematics standards. The item bank representation includes grey dashed boxes to illustrate the conceptual grade-level borders.

[1] Note, that the standard may not be the specific driver of the classroom assessment. In the case of mathematics, levels in a specified learning trajectory may provide a more suitable framework for designing classroom assessments.

This representation includes one box for each mathematics standard[2]. Notice, the standards adopt *domain*-specific colors where, within the Grade 3 partition, the nine green "OA" boxes represent the third grade *operations & algebraic thinking* standards; the three red "NBT" boxes represent the three third grade *numbers & operations in base ten* standards; the three blue "NF" boxes represent the three third grade *numbers & operations—fractions* standards; the eight yellow "MD" boxes represent the eight third grade *measures & data* standards, and the two purple "G" boxes represent the two third grade *geometry* standards. Because there are no actual grade-level barriers in the grade band item bank, the adaptive algorithm is free to select any item from the pool for a student's test.

The actual test for Student *j* is represented in Figure 2 by the box labeled "Test." The boxes within the test labeled *i* with a subscript represents the items on Student *j*'s test. For example i_1 represents the first item, i_2 represents the second item, and i_n represents the last item. The dotted arrows from the item bank to the items represents the adaptive algorithm selecting items from the item bank.

At the top of Figure 2, the grey node labeled "M" represents the construct model. The construct model defines what is explicitly measured by the test. When the construct model is represented with a single node, it is referred to as *unidimensional.* For the current example, the "M" is generically defined as *Grades 3-5 Mathematics*. The solid arrows from the construct to the items represent the interaction between examinee *j* and the items. That is, it represents the conceptual underpinnings of the measurement model: that Student *j*'s *Grades 3-5 Mathematics* achievement causes his or her response to the selected mathematics items.

[2] The diagram adopts a *domain.standard* format (e.g., NF.2) which differs from the official *grade.domain.cluster.standard.sub-standard* format (e.g., 3.NF.A.2.A).

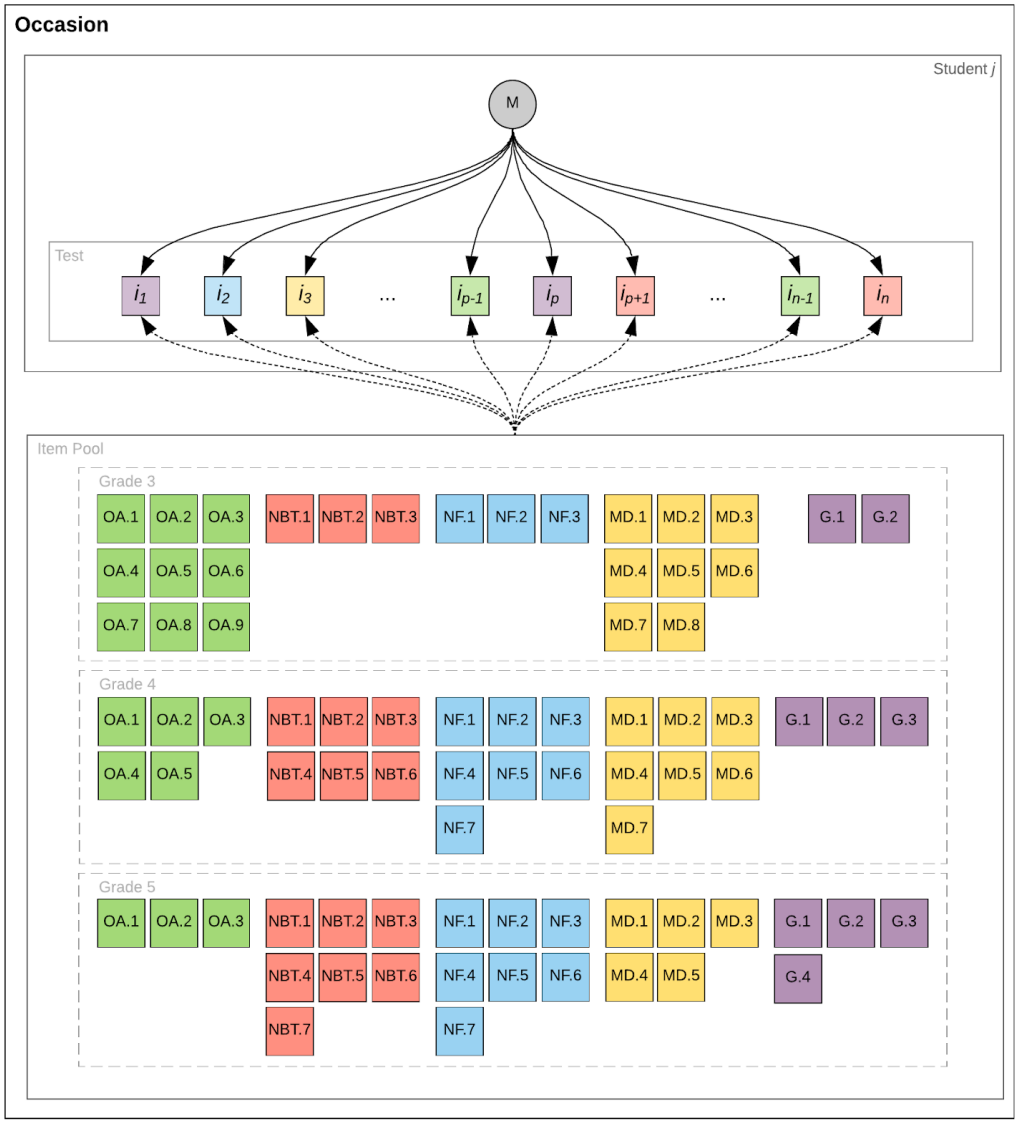

Figure 2. Diagram of a unidimensional construct model assessed by a unidimensional adaptive test using a 3-5 grade band item bank aligned to Common Core mathematics standards.

Although the unidimensional adaptive test with a grade band item bank is the most common commercial interim assessment design, it carries with it many practical issues.

Using a grade band item bank means weakening the instructional validity of the test scores. Consider a high achieving third grade student who encounters items aligned to Grade 4 standards. Allowing this student to encounter items aligned to off-grade standards may be seen as a good thing—the assessment does not place a ceiling on high achieving students. However, a well-designed grade-specific item bank should include items aligned to the on-grade standards that range in difficulty. For example, most math teachers can imagine

an item that requires a student to demonstrate they understand "a fraction 1/*b* as the quantity formed by 1 part when a whole is partitioned into *b* equal parts; understand a fraction *a*/*b* as the quantity formed by *a* parts of size 1/*b*" (3.NF.A.1) that challenges even the highest achieving third grade student. Using a grade band item bank, however, might lead the adaptive algorithm to instead show this student an item that requires they "[e]xplain why a fraction *a*/*b* is equivalent to a fraction (*n* × *a*)/(*n* × *b*) by using visual fraction models, with attention to how the number and size of the parts differ even though the two fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions" (4.NF.A.1), which is a grade four standard. Consequences of assessing students with items aligned to off-grade standards are many.

First, the construct targeted for measurement does not map to the construct addressed in the classroom, or targeted by the end-of-year summative assessment. As mentioned above, the score resulting from such a test potentially offers insights into *Grades 3-5 Mathematics* achievement when *Grade 3 Mathematics*, as defined by the Grade 3 standards, is the construct targeted for third grade instruction. While the item bank may claim to be aligned to state standards, the construct being measured by the test is not aligned to grade-level instruction.

The second issue with the inclusion of items aligned to off-grade standards is what happens when the student answers one or more of them incorrectly. There is no way to understand if the student responded incorrectly due to a misunderstanding of the concepts embedded in the standard, or simply because he or she has not yet had the opportunity to learn those concepts. The resulting score, or the student's location on the scale, can be quite nebulous, leading to unclear next steps for educators and parents alike.

Consider a group of administrators are using scores from a mathematics interim assessment to understand the effectiveness of a particular math curriculum. The data shows growth for high achieving students is slower than the growth made by low- and mid-achieving students. The administrators conclude that the curriculum is not adequate for this particular segment of students. However, the particular growth rate of the high-achieving student group may have been a consequence of the students seeing items aligned to off-grade standards of which they have not yet had the opportunity to learn, as the curriculum was not designed to address off-grade standards. The curriculum under evaluation may have improved the high-achievement students' understanding of the grade-specific standards, but interim assessing using a grade band item bank was unable to determine it.

This lack of clarity in score interpretation brings a third issue. If a third grade student and a fourth grade student receive the same test score, it does not mean that the third grader is ready for Grade 4, nor does it mean the fourth grader would be served better in Grade 3. Cross-grade comparisons must be made with great care.

Grade-specific Item Bank

To improve the instructional utility of an adaptive interim test, some commercial assessments restrict the item bank to items aligned to grade-specific standards. These tests retain the unidimensional construct model and adaptive algorithm described above, but the algorithm can only choose items that are on grade-level. Figure 3 illustrates such a design. Here, the item bank has been reduced to Grade 3 standards only. As a result, the unidimensional construct "M" is now defined as *Grade 3 Mathematics*. Although constraining the item bank to include grade-specific standards derives a construct that represents the content targeted for instruction, the unidimensional construct model is often considered too simplistic.

Note that the item bank in Figure 3 illustrates the five domains underlying *Grade 3 Mathematics.* Unless these domains are explicitly declared in the construct model, the adaptive algorithm will ignore them and deliver items at random until some termination criteria, such as a maximum number of items, is met. Oftentimes, the only content constraint placed on an adaptive test is that a minimum number of items on the test's predefined *content strands*[3] are met. This process is typically called *content balancing*. However, the selection of items to satisfy the content balancing constraints is rarely based on the student's performance on those specified content strands.

We can see that the first item delivered to Student *j* in Figure 3 is an item aligned to a geometry standard (it is purple to indicate it is aligned to a geometry standard). Consider the situation where Student *j* has struggled with geometry concepts all year and answers this question incorrectly. The algorithm will take the incorrect answer into consideration and select a less difficult item for item 2. The second, less difficult, item selected is aligned to a standard from the *number & operations—fractions* domain (again, it is blue to indicate it is aligned to an NF standard). The student happens to really understand fractions and answers the question correctly. The third item will be more difficult than the second, but which domain will the item be selected from?

What we can't see in Figure 3 is that the last 5 items of a 30 item test had to include two Geometry items to satisfy the 5 geometry item constraint. The adaptive algorithm will choose two geometry items to meet the five item requirement, but the difficulty of those items will be based on the student's estimated mathematics achievement up to that point. Student *j* has already seen three geometry items, to which they responded incorrectly, but responded correctly to many items from the other content strands (e.g., *measures & data*). When the algorithm selects the last two geometry items, it will make the selection based on the student's estimated *Grade 3 Mathematics* achievement at that time, ignoring the fact that they may struggle with geometry specifically. This means the algorithm is likely to select geometry items of greater difficulty, to which the student will likely respond incorrectly. Thus, while the algorithm satisfied the test's blueprint, the assessment was not designed to provide

[3] Note that these content strands are typically of a domain-level grain-size but rarely map one-to-one with the standard document.

precise insights into the content strands. The assessment will likely, however, report scores for each content strand.

When a test utilizes content balancing, it will more than likely report scores on those content strands, referred to generally as *subscores.* A subscore is intended to represent a more granular aspect of the test. For unidimensional tests, the subscore is not explicitly defined in the measurement model, but is a byproduct of the design. The estimation of subscores from a unidimensional adaptive test has two issues worth noting. First, as described above, the adaptive algorithm may not have optimally targeted the strand-level for a specific student. Second, to estimate a subscore from a unidimensional test, only the items aligned to that content strand are used in the estimation of another unidimensional score. Such a method has many flaws that have been well documented in psychometric literature. The biggest issue is that they are estimated with so much error (uncertainty or low reliability) that they often provide less clarity about the content strand than the overall score provided in the first place.

Note, the issues outlined in this section are relevant for any unidimensional test, whether or not it is a linear form, adaptive with grade-specific item bank, or adaptive utilizing a gradeband item bank.

Multidimensional Adaptive Test

Often, when a unidimensional construct model is too simplistic, psychometricians consider a multidimensional construct model. To understand the differences between a unidimensional adaptive test and multidimensional adaptive test, we introduce Figure 4. Notice that the construct model has become much more involved. In fact, we've added a construct representation for each of the five mathematics domains defined by the Common Core standards. Solid arrows connect the domain-level constructs to the items aligned to that specific domain. This model represents the idea that Student *j*'s domain-specific understanding elicits their responses to the domain-specific items. The curved arrows linking all of the nodes represent the conception that each domain is correlated with the other. The construct model in Figure 4 explicitly assumes *Operations & Algebraic Thinking*, *Numbers & Operations in Base Ten, Numbers & Operations--Fractions, Measures & Data,* and *Geometry* are all correlated with each other to some degree.

Where the subscore is not an explicit construct of the unidimensional construct model, the subscores *are* the scores for a multidimensional construct model. That is, upon completion of their test, Student *j* would receive a score representing their performance on each of the five domains. What they will not receive is a score representing their performance on *Grade 3 Mathematics*. That is because the construct model of Figure 3 is lacking an explicit *Grade 3 Mathematics* construct. A common solution is often to average the estimated scores from a multidimensional measurement model to obtain a *composite score*, but like subscores derived from a unidimensional model, this is a byproduct of the design. Composite scores derived from estimated multidimensional scores will often ignore critical properties for valid interpretation.

Higher-order Construct Model

Perhaps the optimal balance between a unidimensional construct model with ad-hoc subscores and a multidimensional construct model with an ad-hoc composite score is the *higher-order construct model*, illustrated in Figure 4. The directed graph of round nodes represents each construct being targeted by the assessment. The grey node labeled "M" represents the subject-level construct, *Grade 3 Mathematics*. Solid arrows connect the subject-level construct to the domain-level constructs. The construct model explicitly describes *Grade 3 Mathematics* as a composite construct made up of the five Grade 3 domain-level constructs. Next, solid arrows connect the domain-level constructs to the specific items on the domain-specific section of the Student *j*'s test. That is, we say Student *j*'s domain understanding elicits their responses to the domain-specific items.

Practically, the items associated with a particular domain contribute to the estimation of that domain, as they do in the multidimensional test in Figure 3. Because we also explicitly articulate how the subject-level construct aligns to the multiple domain-level constructs, the item responses from all items contribute to the estimation of the subject-level construct. Such a model improves the precision of the domain scores as there is an explicit connection between each and the subject-level score, while also improving the estimation of the composite score by taking into consideration the error and relative weighting of the domain scores.

Although the higher-order construct model provides a more realistic construct model, allowing the adaptive algorithm to randomly display items across domains potentially introduces construct-irrelevant variance. Random shifting between domains can place greater cognitive demand on a student, which is a different construct than the one defined by the construct model. As a result, a student with a stronger *cognitive load* may score better on the test than a student with a weaker cognitive load, even if their understanding of *Grade 3 Mathematics* is equal.

Adaptive Test Battery

To lower the cognitive complexity caused by random shifting between domains that can confound measures of domain understanding, the adaptive algorithm can be defined as an adaptive test battery. To reduce the cognitive load associated with *taking a test*, and to focus on the measurement of the target constructs, each domain is organized into test sections. These sections are represented by the dashed boxes labeled "S1" through "S5." For each test section, the adaptive algorithm targets only items within a single domain, meaning the Grade 3 Mathematics test is broken into a series of adaptive subtests. Such a framework should minimize the cognitive demands placed on students when required to jump back and forth between domains. Instead, they can focus on a single domain before needing to switch cognitive gears.

Transcend Test Design

Every component of Transcend has been designed to match the richness and depth of the goals districts established for student learning:

- Transcend uses grade-specific item banks so the target constructs carry the same definition as the constructions targeted for instruction. Each grade-specific bank includes items of varying difficulty aligned to each standard.
- Transcend's subject and domain-level scores are produced from a higher-order construct model resulting in exceptional score properties that map directly to your standards document.
- Transcend employs an adaptive test battery so each student is delivered a personalized assessment that targets domain-level understanding while minimizing cognitive complexity.

As a result, Transcend is capable of producing psychometrically robust scores of high instructional validity.

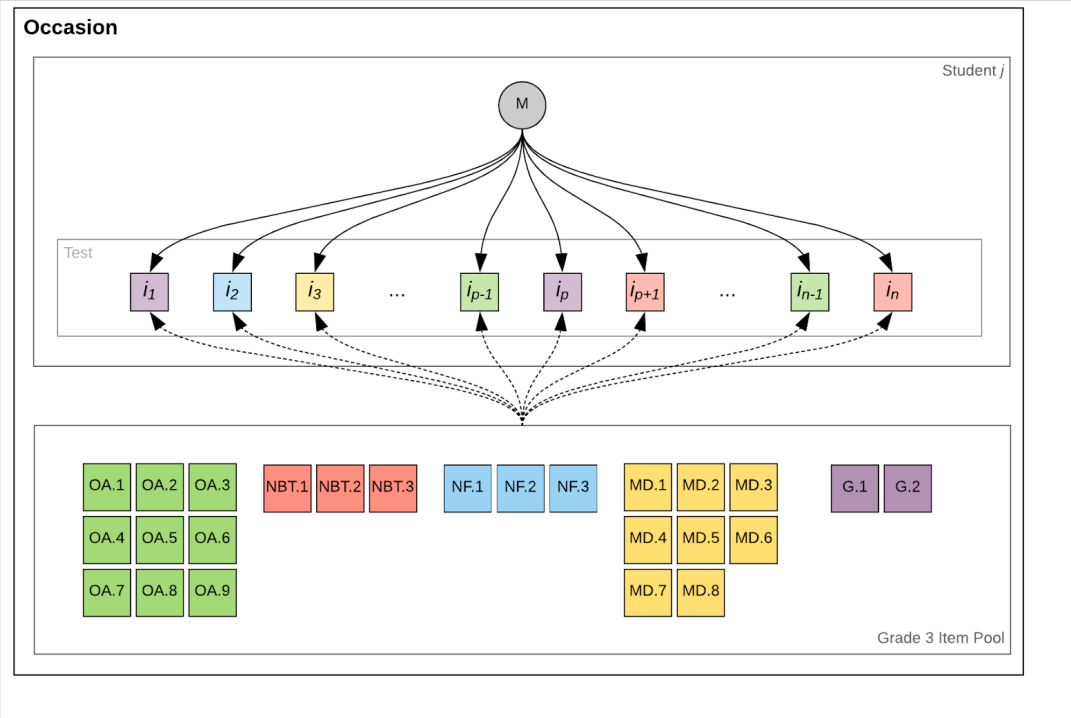

Figure 2. Diagram of a unidimensional construct model assessed by a unidimensional adaptive test using an item bank aligned to Grade 3 Common Core mathematics standards.

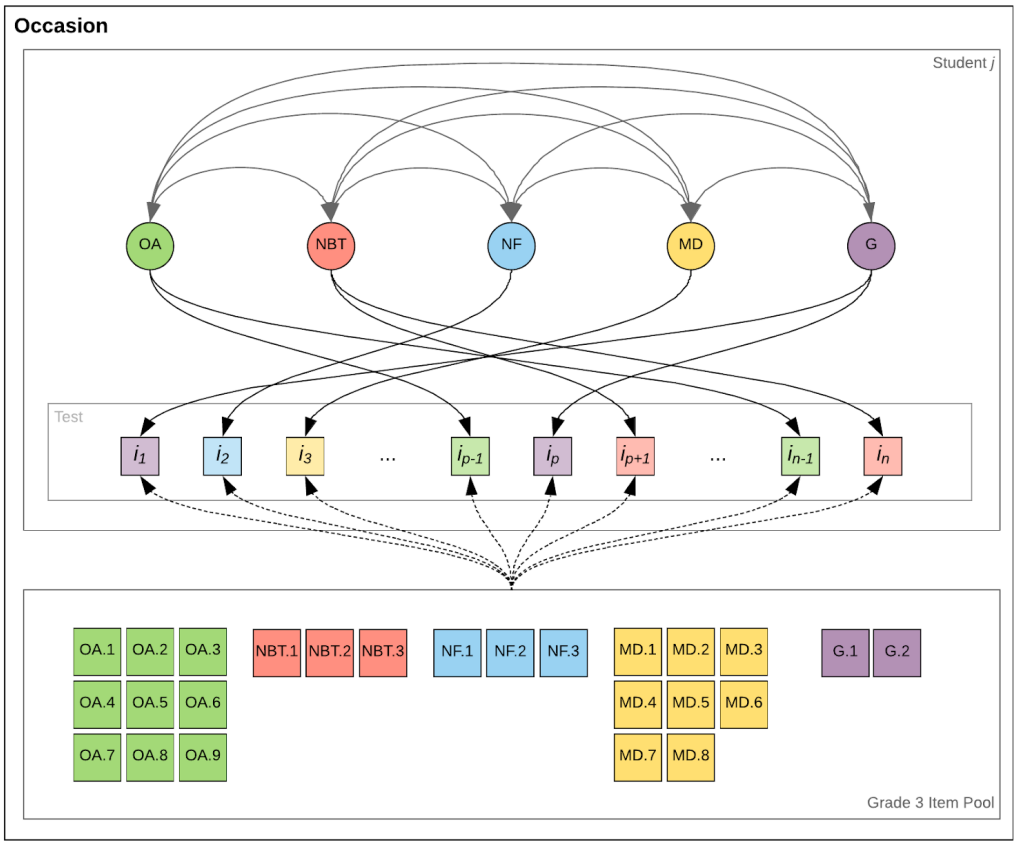

Figure 3. Diagram of a multidimensional construct model, assessed by a multidimensional adaptive test using an item bank aligned to Grade 3 Common Core mathematics standards.

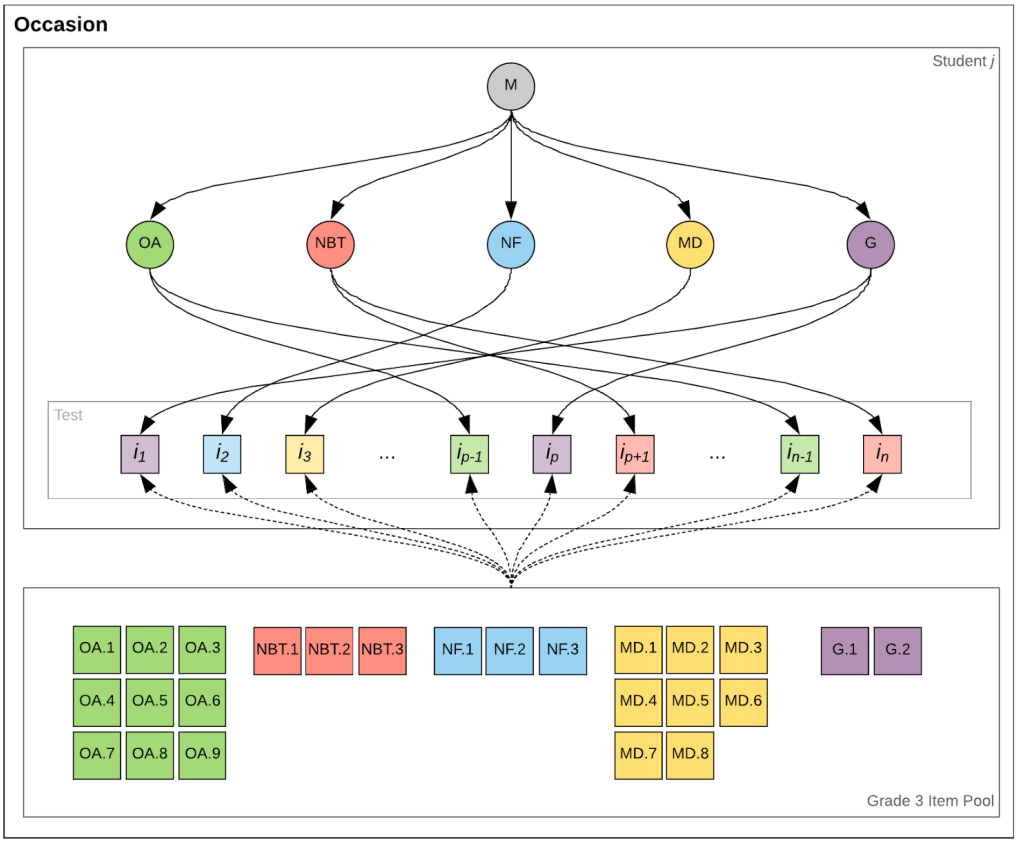

Figure 4. Diagram of a higher-order construct model, assessed by a multidimensional adaptive test using an item bank aligned to Grade 3 Common Core mathematics standards.

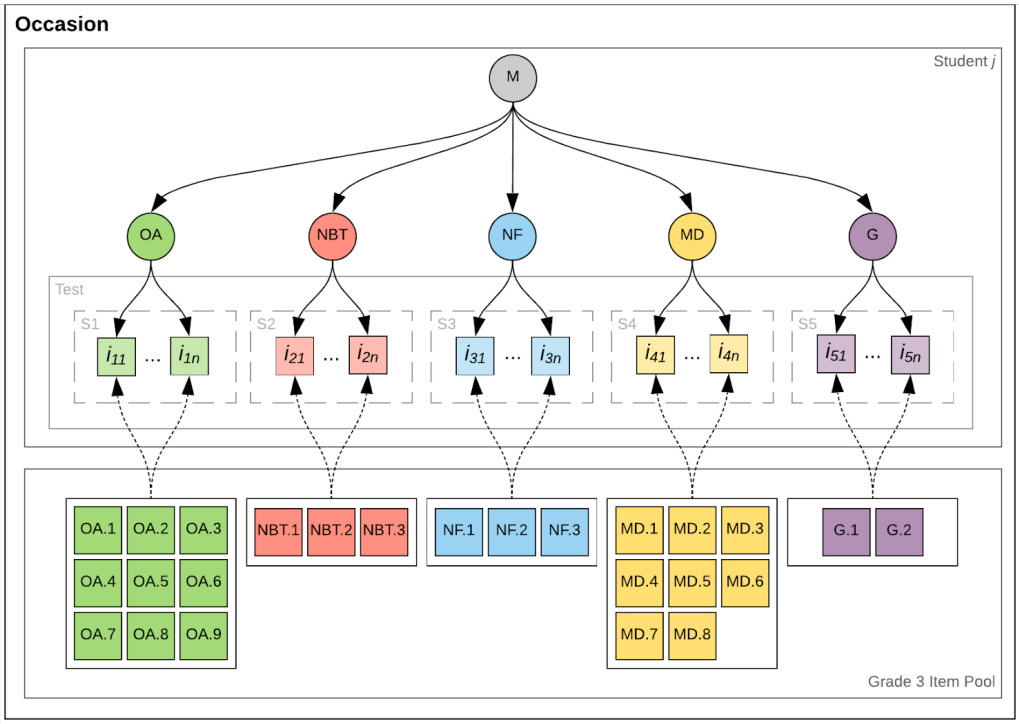

Figure 5. Diagram of higher-order construct model, assessed by an adaptive battery using an item bank aligned to Grade 3 Common Core mathematics standards.