PAPER 3:

Title: Developments for a Diagnostic System to Assess Sources of Mathematical Difficulties

Author(s): Susan E. Embretson
Background:

Although state accountability tests for assessing mathematical achievement have important consequences, they are unable to pinpoint specific sources of deficits for differentiated instruction. That is, the state achievement tests yield a single proficiency score based on large numbers of more specific indicators of competency. Also, the difficulty of items classified as the same indicator of a competency standard can vary broadly which further complicates diagnosis. Validly, reliably and efficiently diagnosing specific deficits requires not only a large item bank with well-specified psychometric and substantive properties, but also adaptive item administration tailored to each skill pattern. Formative evaluation tests that are available typically lack this foundation.

Purpose:

The current study is part of a project that is building an adaptive assessment system to diagnose sources of mathematics deficits for students in Grade 7 and 8. The diagnostic test will be directly linked to state accountability tests at grade level and to student tutorials and lesson plans that are coordinated with the state achievement tests. An alternative set of tutorials also will be available with reduced cognitive complexity level from mathematically irrelevant sources of item difficulty to more clearly focus on mathematical knowledge, especially for special needs students.

The project will provide validation for two assessment systems for mathematical difficulties: 1) the state accountability tests, which are valid for overall assessment of competency, will be validated for differential diagnosis of more global mathematical deficits and 2) an adaptive diagnostic test will be developed and validated for assessing appropriate lesson plans and tutorials for specific skills at varying cognitive levels. The adaptive system provides a more precise diagnosis of instructional appropriateness than the repurposed state accountability test. Nonetheless, the state accountability test may provide adequate diagnosis for collections of mathematical skills, which could be most useful as a prior for the adaptive test. To accomplish these overall goals, several basic research developments are necessary including: 1) analyzing existing operational test items for multiple sources of knowledge difficulty and cognitive complexity, 2) developing appropriate new items for the adaptive test and 3) examining the appropriateness of alternative psychometric models for diagnosis.

The current study concerns the first goal and hence contributes to the content, response process and internal structure aspects of test validity for the state accountability test providing diagnostic information.

Setting:

The research location is classrooms or school laboratory settings throughout the state of Kansas. Assessments are 100% online for state accountability testing in mathematical achievement. Further, lesson plans and student tutorial are available online and they are coordinated with the blueprints for the state accountability tests.
Population:

The students will broadly represent the population of Kansas, but a special focus will be on students with competencies below curriculum-based performance standards, many of whom will be from low income or ethnic minority groups.

Intervention/Practice/Program:

The purpose of the study is to provide initial stages of validating tests for diagnosing sources of mathematical difficulties.

Research Design:

The focus of the current study is examining aspects of validity for applying the state accountability test in mathematical achievement to the diagnosis of global sources of mathematical difficulties. The content and response process aspects of construct validity are studied by analyzing items in an operational test form for 8th grade mathematical achievement. That is, the items will be analyzed for cognitive and knowledge complexity. The test blueprint for 8th grade mathematics is a hierarchical system. Specific skill indicators (25) are nested within benchmarks (10), which in turn are nested within standards (4). At the highest level (standards), the blueprint specifies four areas; Number and Number Sense, Algebra, Geometry and Data. Although each of the 86 items on a test form were constructed for a specific indicator, the presence of multiple indicators, sometimes from different standards, is possible and thus leading to multiple sources of knowledge complexity. The items also may vary widely in cognitive complexity, since items within an indicator vary in difficulty. A cognitive model with five processing stages (Embretson & Daniel, 2008), Translation, Integration, Solution Planning, Solution Execution and Decision, was used to score item cognitive complexity. Each stage contained two or more variables that were postulated to impact complexity,

The content aspect of validity was assessed by determining the degree to which items reflect multiple sources of mathematical knowledge, where mathematical knowledge is defined by the test blueprint indicators. The response process aspect of validity was assessed by the processing complexity involved in item solution. These two aspects of validity are important for diagnostic assessment for two reasons: 1) precision of diagnosis depends on isolating separate sources of knowledge for remediation and 2) cognitive complexity level can represent an auxiliary source of problem complexity for many students. Finally, the internal structure aspect of validity was studied by fitting a multidimensional IRT model for diagnosis to actual test data and examining the internal properties of the assessments.

Data Collection and Analysis:

The operational 8th grade test form for mathematical achievement with the largest volume was obtained. The form was analyzed for the presence of multiple indicators by a team of raters with expertise in mathematical curriculum. The forms were also analyzed for cognitive complexity by a team with raters with advanced knowledge in cognitive psychology. Thus, a team of raters
with an advanced background in cognitive psychology scored the items according to the multi-stage model of cognitive processing variables. A random sample of 2,993 student records from 8th grade was selected to relate item performance to item knowledge indicators and cognitive complexity.

Results:

Rater reliability was assessed for both the knowledge classifications and the cognitive complexity scores. For knowledge classifications, the Fleiss index of rater reliability was .65. For the cognitive complexity, rater reliability was assessed for each variable that impacted processing complexity. Across variables, the mean rater reliability was .87. Thus, overall, rater reliability was satisfactory.

The final knowledge classification results indicated that 30% of the items represented more than one indicator of mathematical knowledge and that 22% of the items involved multiple standards. The final cognitive model scores indicated that while all items involved Translation (i.e., to encode the meaning of the item), there were nine patterns of involvement of the other four components. That is, 14% of the items involved just one additional processing stage, 60% of the items involved two processing stages, and 26% of the items involved three or more processing stages. Thus, processing component complexity varied substantially across items. Furthermore, within processing stages, the descriptive statistics showed substantial item variability in complexity.

The implications of the two systems of variables, knowledge-based and cognitive complexity, were analyzed by a stepwise regression of IRT item difficulties on the two sets of scored variables. First, the four standards were entered (R = .283, p = .144) and significant prediction was not obtained. Thus, the standards had items of similar difficulty levels. Next, the cognitive variables scored for the processing stages were entered and prediction increased significantly (r = .683, p < .001).

A diagnostic item response theory model was fit to the item response data to study the internal structure aspect of validity. The multicomponent latent trait model for diagnosis (Embretson & Yang, 2010) was fit to the data, using the four standards to define the non-compensatory dimensions. It was found that the empirical reliability of the scores on the four dimensions ranged from .786 to .876, indicating high reliabilities for the separate scores. Further, the correlations between scores on the dimensions were moderate, ranging from .619 to .731, which is the expected level for cognitive measures. Finally, profiles of scores on the standards were identified for examinees whose overall scores fell just below the proficiency cutline. These profiles would lead to different recommended tutorials for these examinees.

Conclusions:

The content aspect of validity for using the mathematical achievement test was supported by the finding of multiple indicators and multiple standards in many items. The level of multiple knowledge requirements was not so extreme as to preclude using the items for differential diagnosis of mathematical competencies. The response process aspect of validity was supported.
by finding substantial item differences in cognitive complexity. Further, cognitive complexity was strongly related to item difficulty. This finding supports diagnosing cognitive complexity level, in addition to knowledge type, so that tutorials at the appropriate cognitive level may be recommended. Finally, the internal structure aspect of validity was supported by finding that the reliability of scores on the separate knowledge dimensions (i.e., the standards) exceeded the intercorrelations of the dimensions. This finding, coupled with finding distinct patterns of deficits for students who fall below proficiency, supports the use of the year-end test at a global level for diagnosis. The precision of more narrow diagnosis (i.e., at the indicator level) should be studied for the year-end test in future research and compared to adaptive test results.