ONPAR Mathematics

Research Design and Methodology

Four research questions are planned: 1) What aspects of the interactive and traditional items facilitate or inhibit students in the three groups from meaningfully addressing the items? 2) Given skill estimates conditional on math ability, how do the low and mid ELL students' scores on the interactive items vs. the traditional items relate to the differences in scores on the two versions as for the control group of native English speaking students? 3) Are there meaningful patterns of relationships between student ability, accessibility, reading level, group, types of mathematics skills, item versions, and within-version task characteristics? What is the convergent and discriminant validity of the testlets when selected variables are used? 4) In using multi-media, semiotic representations and constraints that follow from addressing the needs of English learners, has this project conceived of novel approaches to presenting the item and placing it within a simulated and interactive context, presenting the targeted requirements of items using other mediums, and identifying an enlarged set of response opportunities? If so, how would project findings inform the use of multi-semiotic and multi-media assessment items in academic assessments for all students?

To address the research questions a randomized controlled study preceded by iterative clinical trials is planned.

Sampling

The sampling for the clinical cognitive lab trials will be recruited from charter schools in the DC area or identified after-school and weekend program locations in the DC, mirroring methods frequently used by test publishers to recruit students for pilot testing (for instance, see Pearson and Riverside Publishing recruiting documents). Elementary and middle students from the three groups who have had the opportunity to learn the material being assessed on the items, and from whom parent consent has been completed, will be targeted. Small groups of students (n's of approximately 8) will be studied at a time, with about 20 students total taking part in the trials.

The three groups of elementary and middle school students who will participate in the controlled study are two focal groups: ELLs, mostly Spanish speakers, with low proficiency in English (levels 1 and 2), ELLs, mostly Spanish speakers, with mid proficiency in English (levels 3 and 4-4.5), and the control group made up of native English speakers. In a recently completed study, Cook (2008) analyzed how ELL students at different levels of proficiency in English performed on their state's mathematics test with accommodations. State by state, his results showed that relationships between the students' proficiency scores and the test scores seemed to approach parity with their native English speaking peers when the students received English language proficiency level scores from about the mid 4's or higher (out of 6); less proficient students scored significantly lower than their ELL peers nearing mid 4's in English acquisition. Therefore the two groups will reflect the range of ELL students identified by Cook. Native

English speakers, on the whole, should have the least trouble addressing the mathematics in both item versions; therefore, a representative group of these students will act as the control group. They will be selected based on balance of demographics and range of mathematics ability levels.

The sample for the controlled study will be elementary and middle school students from districts recruited from the WIDA consortium. To ensure adequate power, cell sizes of at least 50 students have been recruited for each of the 12 cells (3 groups x 2 versions x 2 grades) for a total of about 600 students. Of these, approximately one-third will be English learners who score within the first two levels on ACCESS for ELLs®, the English language proficiency test, approximately one-third will be English learners who score from level 3 to level 4.5 on ACCESS for ELLs®, the English language proficiency test, and one-third will be native English speakers.

Instruments

The focus of the project is the construction and research of computer interactive mathematics items and their performance relative to their traditional task counterparts. For the clinical trials approximately 8 of these tasks will be studied at a time. For the randomized study, one form for each of the two versions of testlets (interactive and traditional) will be administered by computer to students in each of the two grades. This will occur at two different times during the school year. Each of the forms will be composed of approximately 15 items which measure content taught during the period preceding the administration.

Besides the items, retrospective interview protocols have been developed for the clinical trials which will focus on exploring why students responded as they did, the sophistication of their reasoning skills, their learning experiences, and the accessibility of the item component.

Prior to each of the testlet administrations in the experimental study, teachers will complete a questionnaire per student asking teachers to rate the student's ability levels on skills measured in the testlet versions. The rating system being used will be at a similar level of detail to one used by Schmidt, McKnight, Houang, Wang, Wiley, Cogan and Wolfe (2001) in their analyses of TIMSS textbook elements and curriculum data across countries. These researchers have since used this approach in other studies where it has been found to be a replicable indicator for making differential judgments about content; and Kopriva and others have used this approach to their satisfaction in two subsequent studies (Kopriva, Wiley et al., 2008; Kopriva & Bauman, 2008). The questions ask teachers to rate each student based on whether classroom evidence suggests students have consistently acquired, sometimes acquired, or almost never acquired the skills identified as proficient in the standards. These skills correspond to the objective level target mathematics and cognitive level in each of the items on the testlet the students will take. The teacher ratings will become the independent mathematics ability indicator used in the analysis. Selected demographic data about the students will also be collected from the teacher.

Computer-based training modules for each administration and each version will be developed. The module for the traditional version will use standard instructions; the interactive module will orient them to how the screens are organized, how to use the animated icons, etc. These types of modules were used in the first ONPAR study and were completed by students in

about 12 minutes, and it is expected that administrations of the modules for this project will take about 12 minutes as well. All modules are programmed so students can listen to them in English or their home academic language.

Each native English speaking student will take the ACCESS screener which gives an estimate of their reading level for the controlled study. This score, along with the reading score on the ACCESS test for ELLs will be used in the analyses for Question 3. Additionally, an access questionnaire, developed for the controlled study will assign access levels to each item. This questionnaire will be completed by the students during testing.

For both the cognitive labs and controlled studies, parent consent forms and program, district, school and teacher information documents have been developed.

Procedures

Project staff will first identify specific constructs targeted for this investigation. Released task banks (for instance, on statewide test websites, NAEP and TIMSS) will be reviewed to identify traditional items at the appropriate level of cognitive complexity. Appropriate items will be selected and these will inform the building of the interactive items. Conceptualization of the interactive items at the targeted level of complexity will then begin with the drafting of storyboards. Subsequently, the conceptualizations will continue to be refined and tasks will be built using CAL's development and review process. This process involves four teams of experienced researchers at CAL as overseen by Dr. Kopriva from the University of Wisconsin—the math assessment team, the design and programming team, the concepts and task review team, and the research team—which operate iteratively to complete the tasks. After a set of tasks are completed and the training module for the trials is developed, the labs will begin.

It is anticipated that the labs will begin in approximately 6 months after the work begins. In all, 15-20 interactive items need to be judged viable candidates at each grade for the experimental study. The controlled study will be conducted in two stages, in the fall and in the following spring. Both data collections will be discussed in more detail below. As data are collected, they will be sorted, cleaned, scored and prepared for analyses. The data collection will be overseen by Dr. Jim Bauman of CAL.

Data Collection The focus of the cognitive lab trials is to determine how traditional and interactive items inhibit or facilitate students' understanding of the mathematics questions targeted in the items, and inhibit or facilitate their problem solving and demonstration abilities. Elementary and middle school students with the range of language and math skills will be recruited for the trials and parental consent solicited. Subsequently, small groups of students will be given selected traditional or interactive items to complete, and one-to-one retrospective interviews immediately following the group administration will be conducted by staff personnel to determine how the students interacted with the tasks. Both English and home language speaking staff will conduct the interviews using protocols developed for the project. Notes regarding the interviews will be taken and recorded on project developed forms, trials will be

videotaped, tapes and notes will be reviewed, and data from the trials will be analyzed to inform the revisions of the tasks. Documentation procedures for making changes to items on the basis of internal reviews and cognitive lab results have been developed, and these procedures will be used to keep track of changes for this project. It is anticipated that cognitive labs will occur over about 8 months, until researchers are satisfied that they understand how students are interacting with various traditional and interactive task elements under different types of problem solving situations.

It is expected that the first portion of the controlled study will be conducted in the fall for both grades 4/5 and 7/8; the second will occur the following early spring. To prepare for each administration, communication will occur with teachers and district and school personnel to confirm that students have had the opportunity to learn the material being tested, address compensation and other participation issues, and determine when administrations should occur to minimize disruption at the school sites. Students will be identified based on group characteristics and language and reading test scores, and parent consent forms will be disseminated. Successful items will be grouped into two testlet version forms (interactive or traditional) per grade. The training modules and the teacher questionnaires about each student will be developed, and teacher questionnaires disseminated and collected prior to the test administration. For each administration the forms will be randomly assigned to students from each of the three groups. Subsequently data will be sorted, scored and prepared for analyses.

For all investigations informational documents will be circulated to district and school personnel, including teachers whose students are participating. Parent consent forms will be administered and collected. Project personnel will staff clinical trials and the experimental study test administrations will be conducted by both project personnel and trained test administrators.

Data Analysis Plan

Data analyses will begin after the first data collection and continue throughout the project. All analyses will be overseen by Dr. Kopriva.

Qualitative Qualitative analyses will be used in a formative manner to inform the development of the interactive tasks. They, along with the controlled study findings, will also be used to address the first research question: What aspects of the interactive and traditional tasks facilitate or inhibit students in the three groups from meaningfully addressing the tasks?

To inform the research question and the construction of tasks, data from the completed protocol response forms will be analyzed. To address the ongoing development of tasks, preliminary findings from the cognitive labs will be completed and discussed with the item development and research teams after which decisions will be made about the types of changes items might undergo. The results of the discussions will be transferred onto forms used in the change documentation procedures. To address the research question, the preliminary findings, by lab, will be collected and reviewed again along with the controlled study results. The focus of this later review is to identify patterns of characteristics over task items and task characteristics

that encourage or discourage the communication of cognitive complexity, as well as patterns that inhibit or facilitate the communication of the targeted question or the students' ability to respond according to his or her level of math skill. The results and recommendations will be documented in the final report for this project.

Quantitative Quantitative analyses will be used to address the following two research questions. Question 2 will be addressed by fitting a series of hierarchical Structural Equation Models (SEM) to assess the differences between task testlet versions within student groups, and differences between groups. The hierarchical data, students grouped by classroom and school, will be considered random effects so that parameters at the student level will be, by definition, adjusted for effects at the higher levels. In effect, the model will estimate relational parameters between ability and scores controlled for both higher level effects (classroom and school) and measurement error of the testlet. Once parameters within group or testlet version have been identified, differences between parameters of fixed groups and versions can be tested for significance. A statistical program, such as LISREL (Joreskog & Sorbum, 1999) will be used. Effect sizes of differences will also be calculated.

In this way researchers will analyze if the differences in scores between versions is greater or less for ELL students as compared to the differences between forms for native English speakers. Essentially, controlling for ability, as defined by the math ratings by the teacher on the questionnaire, allows comparison of scores of students with similar ability levels in each group. The hypothesis is that the differences will be greater for the focal groups than for the control group, which signals that the interactive tasks are better representing the skills of these students, and that there is a significant difference between the two. The relationships between teacher ratings and student performances will also be a subject of analysis.

Question 3 (Are there meaningful patterns of relationships between student ability, accessibility, reading level, group, types of math skills, task versions, and within-version task characteristics? What is the convergent and discriminant validity when selected variables are used?) will first be analyzed with an omnibus correlation matrix. Subsequently, another series of hierarchical SEM models will be fit using LISREL to incorporate the additional variables listed in question 3. Tables 1 and 2 are an example of the type of model which will be fit to analyze the convergent and discriminant validity of the testlets. Parameters from different patterns will be tested for significance.

Table 1: Convergent Validity Structural Equation Model

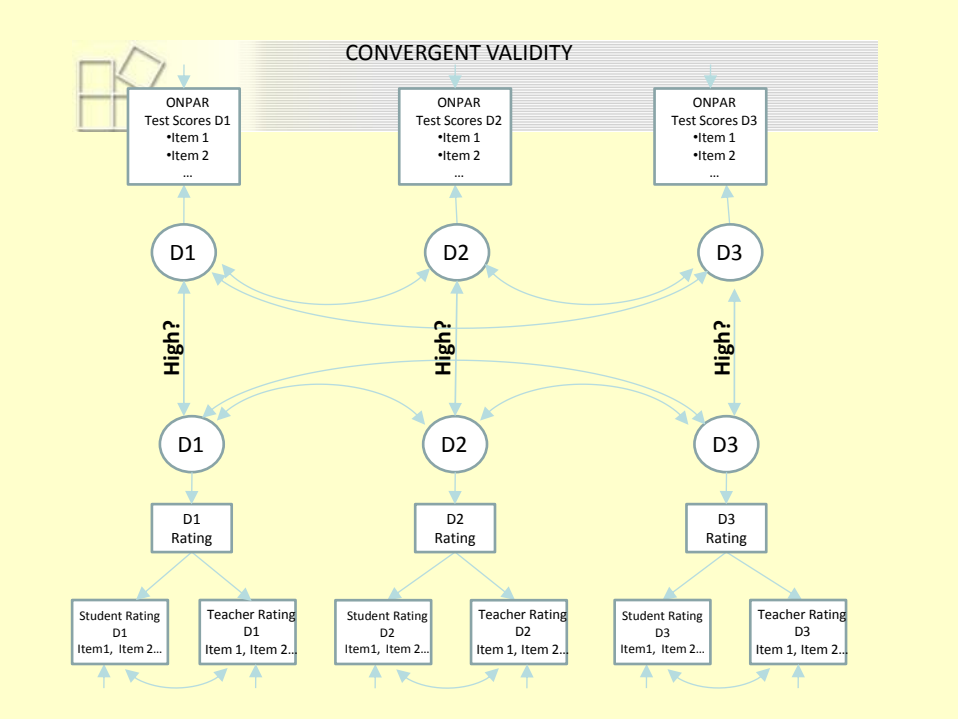
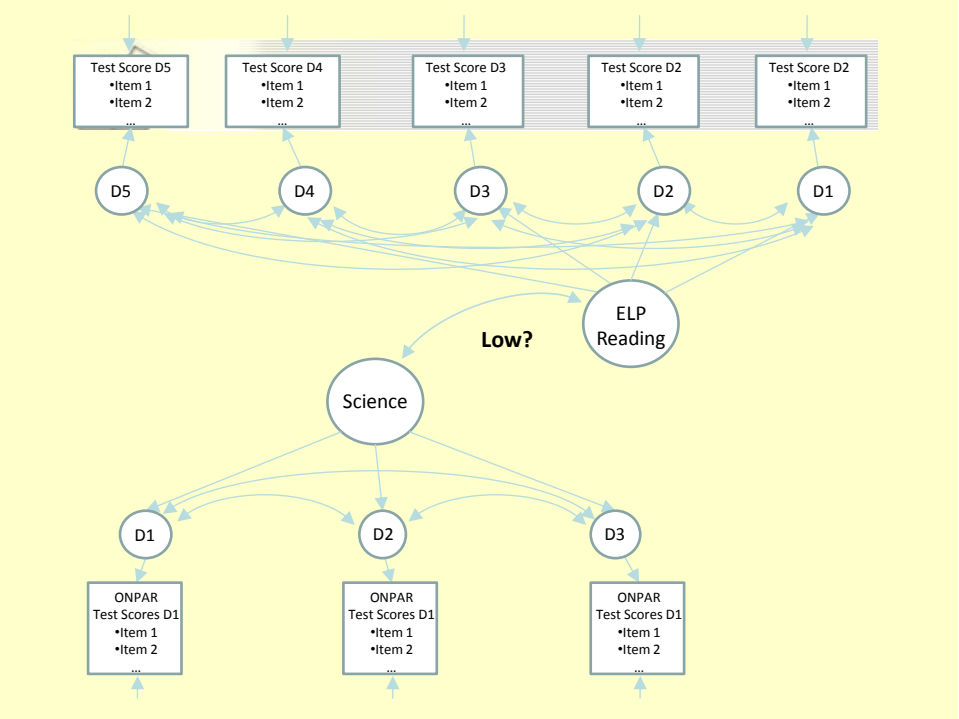


Table 2: Discriminant Validity Structural Equation Model



Question 4 asks whether novel approaches might be applicable and beneficial to a broader range of test takers. The analyses to address this question will be summary judgments based on the success of individual interactive items. Summary judgments will also be made about how well aspects of the items might effectively redirect language, how well the multi-semiotic representations reflect meaningful math, how the mathematics is conveyed, and how students perform vis-à-vis various response avenues that seem to provide students meaningful ways to demonstrate their knowledge or skills. Judgments will be based on results and implications from the analyses discussed above.