Running Head: EXPEDITIONARY LEARNING IMPLEMENTATION REVIEW

Expeditionary Learning Implementation Review: Instrument Development

Sue Leibowitz
University of Massachusetts
Donahue Institute

Larry H. Ludlow
Boston College

Tom Van Winkle
Expeditionary Learning

American Educational Research Association Annual Meeting
Paper Presentation
Philadelphia
April 2014

ABSTRACT

Expeditionary Learning (EL) and UMass Donahue Institute (UMDI) worked together to develop a new instrument to assess the extent to which EL core practices are implemented. The instrument was constructed using the principles of the Rasch rating scale measurement model. Three pilot tests were conducted. Full-scale implementation and validation data were collected from 147 schools across the United States. Factor analyses indicate one dominant factor. The items themselves are internally consistent within each of the separate dimensions and overall. The Rasch model rating scale analysis of the data supports the hypothesized conceptualization and development of a meaningful continuum for both schools regarding their overall level of implementation of the EL practices, and the degree of difficulty in implementing the practices.

PURPOSE OF THE STUDY

Expeditionary Learning (EL) is a national network that partners with schools, districts, and charter boards around the country to open new schools and transform existing schools. The network provides school leaders and teachers with professional development, curriculum planning resources, and new school structures to boost student engagement, character, and achievement. Each year EL conducts a review of each school to determine the extent to which the core practices are being implemented. The Implementation Review (IR) process is a key input to EL's national, regional, and school-level planning.

The purpose of this project was to develop a new IR instrument to assess the extent to which core practices are implemented within EL schools and return actionable knowledge to EL's national and regional offices, school designers (coaches), and individual schools. The previous IR model—consisting of assessments completed by principals, teachers, and school designers (coaches), with designers responsible for articulating a final summative assessment—had been in place for several years. It was in many regards a robust system, but needed to be aligned with recently revised core practice standards and updated to address EL's concerns about aspects of its conceptualization, approach, and analysis. Key leaders of Expeditionary Learning (EL) and researchers from the UMass Donahue Institute (UMDI) worked together to significantly revise EL's existing Implementation Review (IR) process.

THEORETICAL PERSPECTIVE

The extent of school implementation was theorized as a continuum measured by a scale designed to fit the Rasch rating scale model (Rasch, 1960; Wright & Masters, 1982). Further, EL was interested in increased inter-rater reliability from the previous instrument, the opportunity to measure change longitudinally, and the opportunity to calibrate standardized scores which could be used to conduct research in the future. This new instrument provides a unique approach to measuring the level of implementation of practices within a whole school context.

METHODOLOGY

The instrument was constructed using the principles of the Rasch rating scale measurement model (Rasch, 1960; Wright & Masters, 1982) such that it would be sufficiently sensitive to measure a variety of EL core practices and adequately measure change over time. It was hypothesized that some practices are easier to implement than others and that a hierarchy exists for practices from ones easier to accomplish, or practices that need to happen first, to more difficult practices, or ones that can happen later. Within this overall hierarchy EL core practices are organized conceptually into five dimensions: Curriculum, Instruction, Assessment, Culture & Character, and Leadership, as described below (Expeditionary Learning Outward Bound, 2011, p.5).

Curriculum. Academically rigorous learning expeditions, case studies, projects, fieldwork, and service learning inspire students to think and work as professionals do, contributing high-quality work to authentic audiences beyond the classroom. Schools ensure that all students have access to a rigorous college preparatory curriculum, and regularly analyze the curriculum to check alignment to standards and opportunities for all students to meet those standards.

Instruction. Lessons have explicit purpose, guided by learning targets for which students take ownership and responsibility. In all subject areas, teachers differentiate instruction and maintain high expectations.

Assessment. Students continually assess and improve the quality of their work through the use of models, reflection, critique, rubrics, and work with experts. Staff members engage in ongoing data inquiry and analysis, examining everything from patterns in student work to results from formal assessments, disaggregating data by groups of students to recognize and address gaps in achievement.

Culture & Character. Schools build cultures of respect, responsibility, courage, and kindness, where students and adults are committed to quality work and citizenship. School structures and traditions such as crew, community meetings, exhibitions of student work, and service learning ensure that every student is known and cared for, that student leadership is nurtured, and that contributions to the school and world are celebrated.

Leadership. School leaders build a cohesive school vision focused on student achievement and continuous improvement, and they align all activities in the school to that vision. Leaders use data wisely, boldly shape school structures to best meet student needs, celebrate joy in learning, and build a school-wide culture of trust and collaboration. Leadership goes beyond a single person or team—it is a role and expectation for all.

The research team condensed the large number of original EL core practices into a smaller number that were determined to have high impact on student achievement either directly or indirectly. This new set came to be referred to as *power practices*. Teams consisting of researchers and EL field staff were assigned to determine power practices within each EL dimension. Once 33 power practices were refined, EL field staff members (regional directors and school designers) discussed, proposed, critiqued and agreed upon a theory-and-practice based hierarchy of implementation difficulty within each dimension. This process yielded not only a hypothesized scale continuum cutting across the dimensions but also reduced the practices, or items, to 26.

A novel scoring rubric was adapted from a classroom observation instrument (Smith & Dickinson, 2002). For each power practice, a rating scale is used with five levels of implementation. The general framework is 1 to 5 with three anchor points defined with descriptive statements: at Initial (1), Moderate (3), and Exemplary (5) levels. The adjoining points, (4) and (2), differentiate gradations between the defined levels. The specific wording for the rubric was tailored to each power practice (see, for example, Figure 1).

Figure 1: Sample Rubric with Scoring Instructions

How to Score the Rubrics

- Read column 5 for exemplary practice.
- To score, start in the middle column; read all statements in column 3.
- If **ALL** criteria in column 3 are **met or exceeded**, go to column 5.
    - If none in column 5 are met, the score = 3.
    - If some in column 5 are met, the score = 4.
    - If all in column 5 are met, the score = 5.
- If **SOME or NONE** of the criteria in column 3 are met, go to column 1.
    - If all criteria in column 1 are true, the score = 1.
    - If some criteria in column 1 are true, the score = 2.

I. Curriculum	**Exemplary**		**Moderate**		**Initial**
	5	**4**	**3**	**2**	**1**
#1 Mapping Skills & Content	Teachers and school leaders map skill standards across the year (horizontally) and from grade to grade (vertically). Curriculum content is based on required standards and sequenced to maximize opportunities for interdisciplinary connections when appropriate. Teachers and school leaders regularly create, analyze, and adjust schoolwide content maps to ensure that standards are addressed, students are engaged and challenged, and repetition is minimized. Teachers and school leaders ensure that all students have access to a rigorous curriculum with appropriate complexity of text and tasks.		Teachers and school leaders have at least begun to map skill standards across the year (horizontally) and from grade to grade (vertically). Curriculum content is sometimes based on required standards and sequenced to maximize opportunities for interdisciplinary connections when appropriate. The school has created and implemented schoolwide content maps. Teachers and school leaders sometimes ensure that students have access to a high-level curriculum with appropriate complexity of text and tasks.		There has been no work done at the school level to map skill standards horizontally and/or vertically. Curriculum content is not yet based on required standards and sequenced to maximize opportunities for interdisciplinary connections when appropriate. The school has not created schoolwide content maps. School structures or barriers exist that prevent most students from having equal access to high-level curriculum.

The power practices, or items, that constitute the instrument are shown in Table 1:

Table 1. Power Practices

I. Curriculum Dimension:
Mapping Skills & Content
Case Studies
Projects and Products
Learning Expeditions
II. Instruction Dimension:
Effective Lessons
Supporting All Students
Reflecting & Structuring Revision
Culture of Reading
Culture of Writing
Culture of Mathematics
Integrating the Arts
III. Assessment Dimension:
Assessment For Learning (AFL)
Learning Targets
Quality Assessments
Analyzing Assessment Data
Communicating Student Achievement
IV. Culture & Character Dimension:
Engaging Families
Learning Community
Crew
Beautiful Spaces
Fostering Character
V. Leadership Dimension:
School Vision
Using Data
Supporting Planning, Assessment, & Instruction
Positive School Culture
Professional Learning

These 26 items assess a complex system of education (i.e., EL) utilizing a measurement model that reduced the complexity to a meaningful construct valid description of a school in terms of a single score: “overall level of implementation”. This variable, constructed from five dimensions of highly correlated educational practices, was designed to provide a single hierarchical continuum of the difficulty, and ultimate success, in implementing the EL power practices.

Three pilot tests were conducted. Pilot 1 occurred in March 2012 with 19 schools in the Midwest and New York regions. School designers were trained by phone conference in how to use the new IR instrument. Pilot 2 occurred in May 2012 with 123 schools across all regions that did not participate in the first pilot. Regional directors and school designers received two hours of in-person training. Pilot 3 was implemented in December 2012/January 2013 with 20 schools representing all regions. The instrument was revised considerably, with changes related to language and descriptions within the rubrics rather than substantive content changes.

Full-scale implementation and validation data were collected during Spring 2013. The sample consists of 147 schools across the United States. They represent varying levels of poverty (as defined by Free/Reduced Lunch),

minority status, governance (district, charter, independent, other), and years in the EL partnership. School designers (coaches) provided the "official" scores. However, administrators and teachers also participated in rating their school, so these ratings were used in the validation too. Spring 2013 data were collected using Qualtrics. Data were analyzed using SPSS and WINSTEPS (Linacre, 2009).

RESULTS

Classical Measurement

Cronbach's alpha was .968 for all items combined and ranged from .845 to .920 on the individual dimensions. For all items combined, the average corrected item-total correlation was .722 with the dimension values ranging from .656 to .772. These statistics support the reliability of the total score and the individual dimensions, and highlight the highly correlated nature of the items within each dimension.

Dimension scores were created by adding individual item raw scores within each dimension and a total score was created by adding the raw scores of all items. Their correlation matrix shows very high correlations among the five conceptualized dimensions (.655 to .883) and between each dimension and the total score (.838 to .956).

Principal axis factoring tested the extent to which the dimensions define five orthogonal factors. Initial analysis showed good evidence that their item-level correlation matrix was appropriate for factoring: the determinant was not zero, the Kaiser-Mayer-Olkin Measure was high at .948, and Bartlett's test value was statistically significant. The first eigenvalue was 14.608, accounting for 56.186% of the total variance, while the second eigenvalue was 1.478 accounting for only 5.686% of the variance. Item communalities for the single factor were all high, e.g., .789. The factor loadings for factors 2, 3, 4, and 5 were weak and the factors were not interpretable. These results, along with the dramatic "elbow" pattern seen in their scree plot provide strong evidence for one common factor.

A follow-up analysis tested for two independent factors. With a Varimax rotation most items loaded on one factor with five (out of 26) items loading onto a weak second factor. These five items comprise the Leadership dimension. With an oblique rotation (using the oblimin procedure) the correlation between the two factors was .702. This finding again supports the earlier results that one dominant factor unites the items of the five dimensions, and even when a second factor is extracted, it is highly correlated with the first.

Overall, the various factor analyses indicate one dominant factor, not five. The items themselves are internally consistent within each of the separate dimensions (and overall, using all items to form a single score) and yet are so highly inter-correlated that the five dimensions can be reduced to a single factor called "overall level of implementation." From a statistical perspective the dimensions may be interchangeable in terms of their predictive strength (and individually they may not even be as strong as using a single total score created from all the items).

Rasch Rating Scale Measurement

The Rasch model rating scale analysis of the data supports the hypothesized conceptualization and development of a meaningful continuum for both schools regarding their overall level of implementation of the EL practices, and the degree of difficulty in implementing the practices. Figure 2 shows the variable map for the Spring 2013 IR administration. The implementation continuum consists of easy practices on the lower end – functions that are easier to accomplish earlier in the process or require accomplishment early on – and more difficult practices on the upper end – functions that are more difficult to accomplish and occur later.

At the easier end of the continuum are leadership and school culture practices such as having a school vision, creating a positive school culture, creating a community of learning, and engaging families. On the difficult end of the continuum are multiple assessment practices and more complex instructional practices like creating effective lessons, having a culture of mathematics, and communicating student achievement. The "easier" practices may not mean they are easy to accomplish, but rather that they must be done before moving to more difficult curriculum, instruction, and assessment practices.

Since administrators and teachers also participated in rating their school, these data were used in validating the scale. Table 2 shows item and school separation values for designers, administrators, and teachers. Designers and administrators had similar separation statistics for both items and schools. Teachers were similar to administrators in rating their own school. However, the statistics were not comparable for item separation due to the large difference in N between teachers and both designers and administrators.

Table 2. Item and School Separation for Designers, Administrators, and Teachers

	N	Item	School
Designers	147	4.53	5.27
Administrators	166	4.33	4.03
Teachers	2318	--	4.07

The correlations for item difficulty measures across the three groups are generally strong, with designers and administrators the strongest (r=.896). Administrators and teachers were also robust (r=.748), with the least agreement between designers and teachers (r=.591). Although in general, teachers tended to rate easier, on some items they rated harder (Figure 3).

The category characteristic curves provide evidence that the unique rubric scoring was effective (Figure 4). All categories are used and the scoring categories are being responded to similarly across the three groups. The designer and administrator use of categories 2 and 4 is nearly identical, while teachers tended to use category 3 and 5 relatively more frequently.

Misfit is minimal. Item A4-Communicating Student Achievement is somewhat misfitting for all three groups. Some people score it higher than expected; others scored it lower. Item CC2-Crew also misfit for teachers with some higher and some lower than expected.

Analyses of Rasch measures using demographic variables were performed using ANOVA. There were no statistically significant differences across the seven EL regions or in governance (Charter, District, and Independent). Statistically significant differences ($p<.01$) exist between high poverty schools (> 75% FRL; lower implementation estimates) and both middle (40%-75% FRL) and low (<40% FRL) poverty schools, but not between middle and low poverty schools. Analysis of years in the EL partnership in relation to level of implementation of EL practices indicates a positive correlation (r=.512) with years of partnership accounting for 26% of the variance in implementation.

"Curriculum maps" to explore current level of implementation and expected level by each of the five EL dimensions have been developed and will be illustrated. See, for example, Figure 5. In this type of map we show the actual responses designers provided, the expected responses for a school at that level of implementation, and, most important, the next set of practices to work towards for a school.

Significance of the Study

A Rand Corporation study (Vernez, Karam, Mariano, & DeMartini, 2006) concluded that few schools that adopt Comprehensive School Reform Models fully implement them for various reasons, including the challenges associated with teachers actually implementing new instructional approaches and associated measures of change. This new instrument provides a powerful diagnostic approach to measure the level of implementation of practices within a whole school context and, when coupled with sufficient technical support, provides clear measures of instructional change school wide. Further, it provides useful longitudinal implementation data within a large school network and can provide valuable overall and dimension-level information for national, regional, and school-level planning and assessment.

Figure 2. Expeditionary Learning Implementation Review – Spring 2013
System Level Variable Map: Level of Implementation Relative to the Average Practice [N=147]

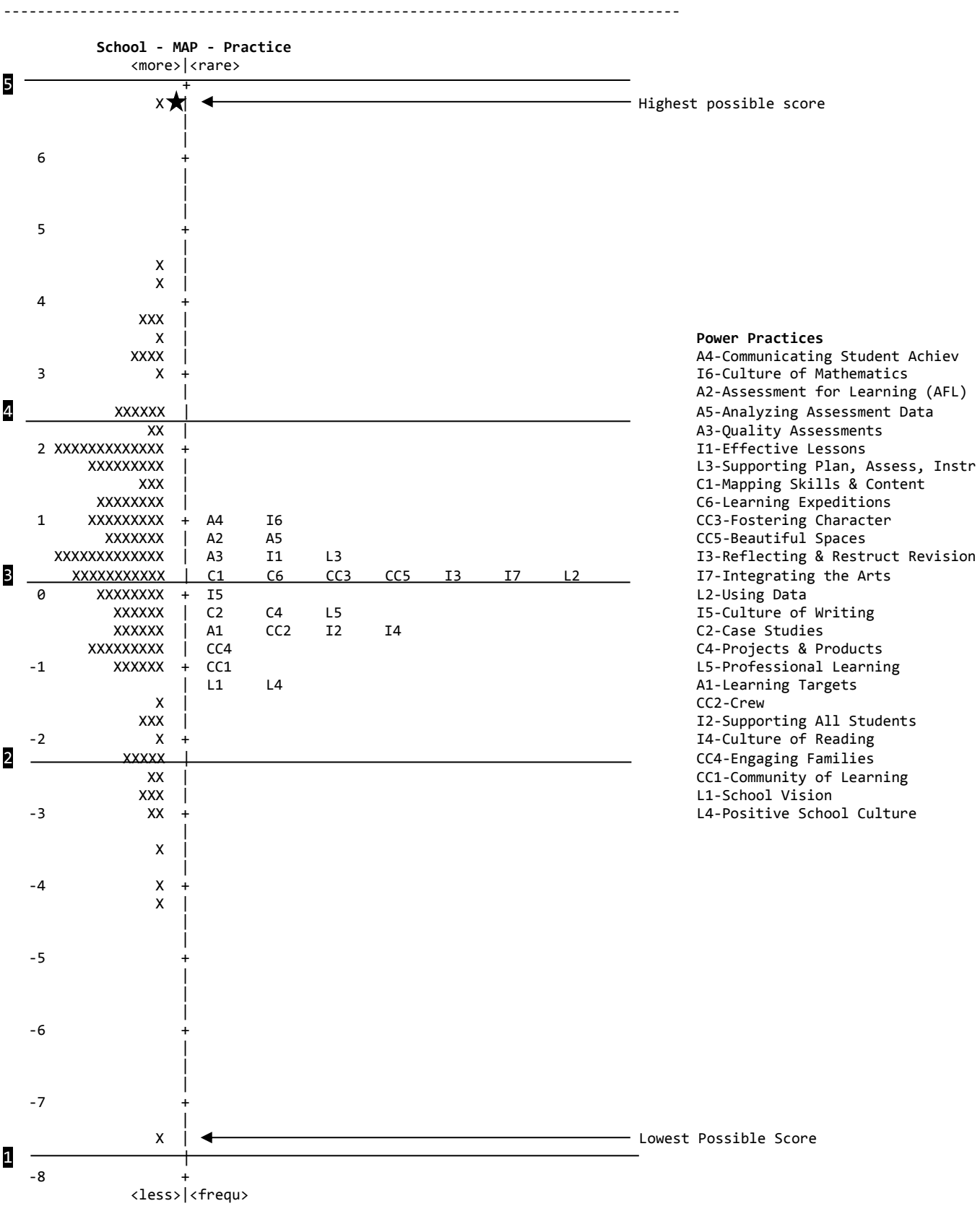
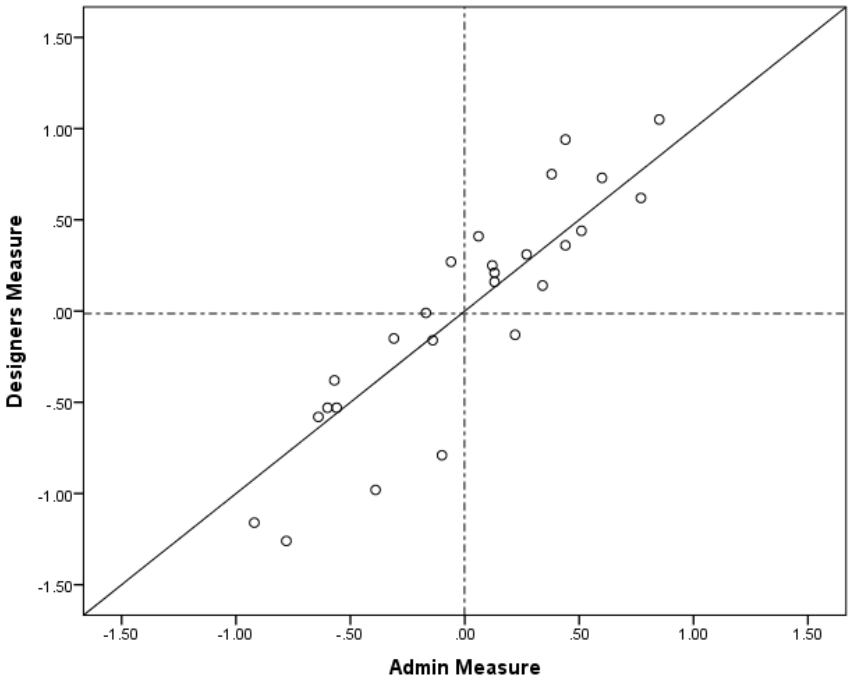
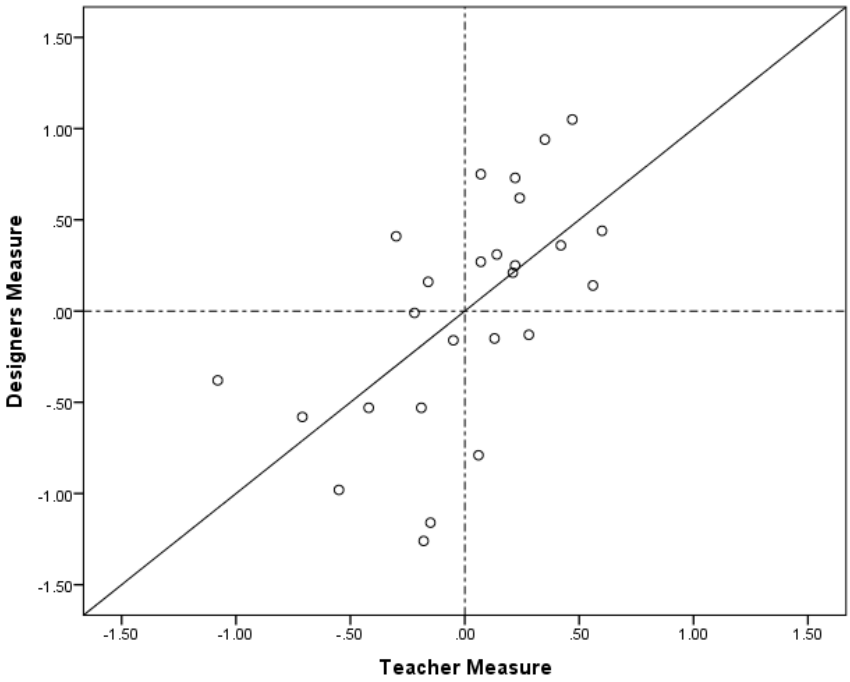


Figure 3. Item Difficulty Measure Plots for Designers, Administrators, & Teachers

Part A. Designers and Administrators



Part B. Designers and Teachers



Part C. Teachers and Administrators

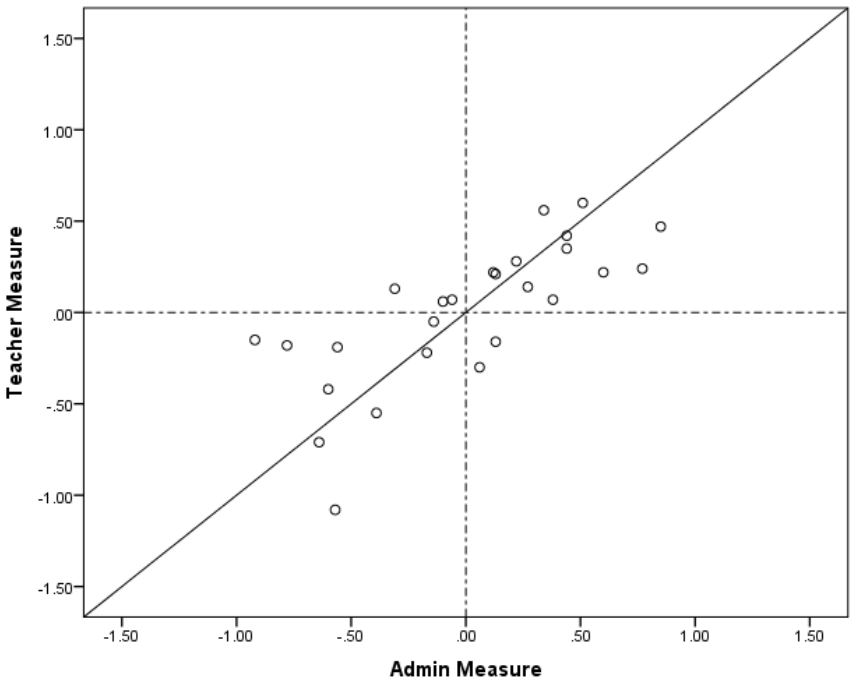
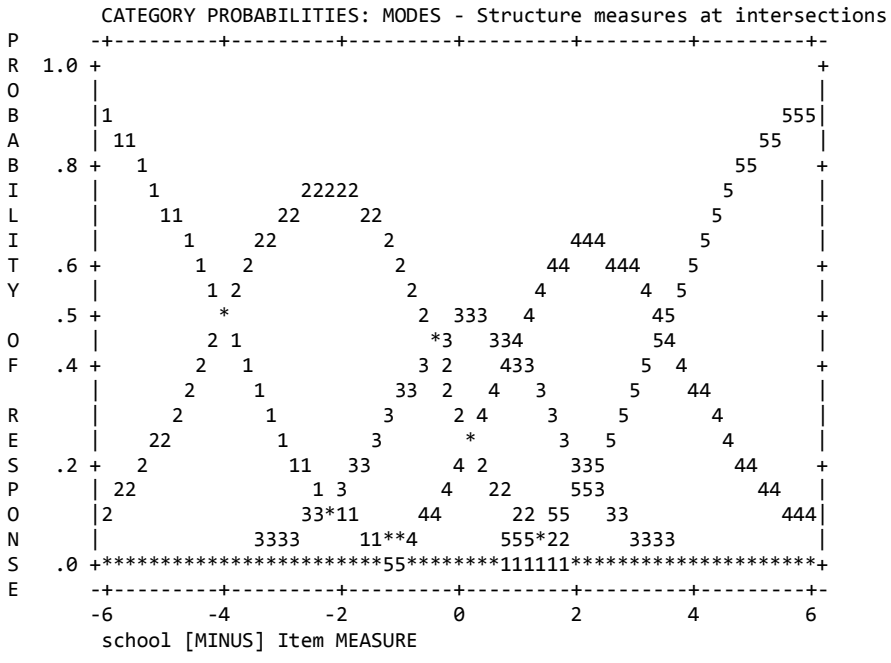
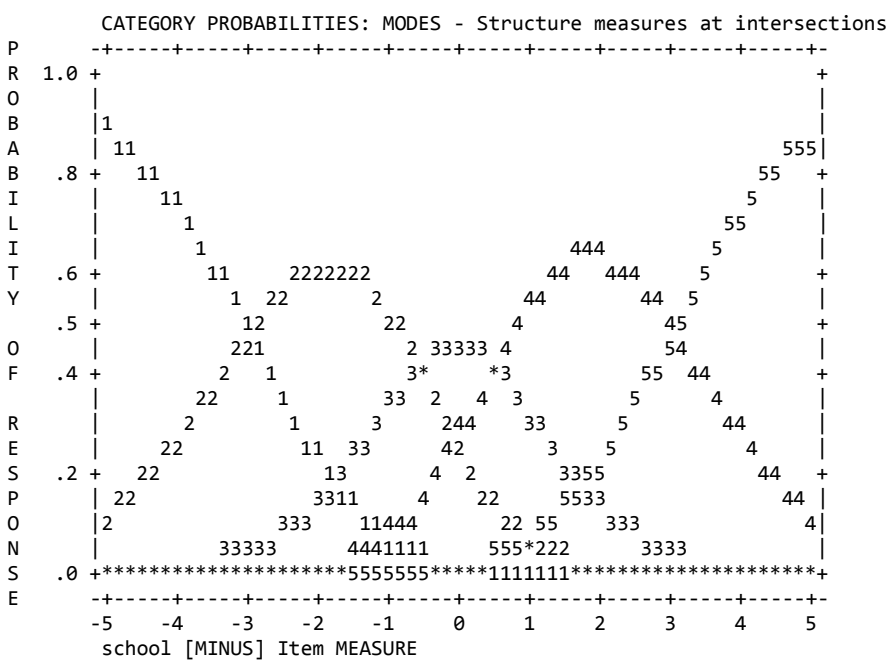


Figure 4. Category Characteristic Curves for Designers, Administrators, & Teachers

Part A. Designers



Part B. Administrators



Part C. Teachers

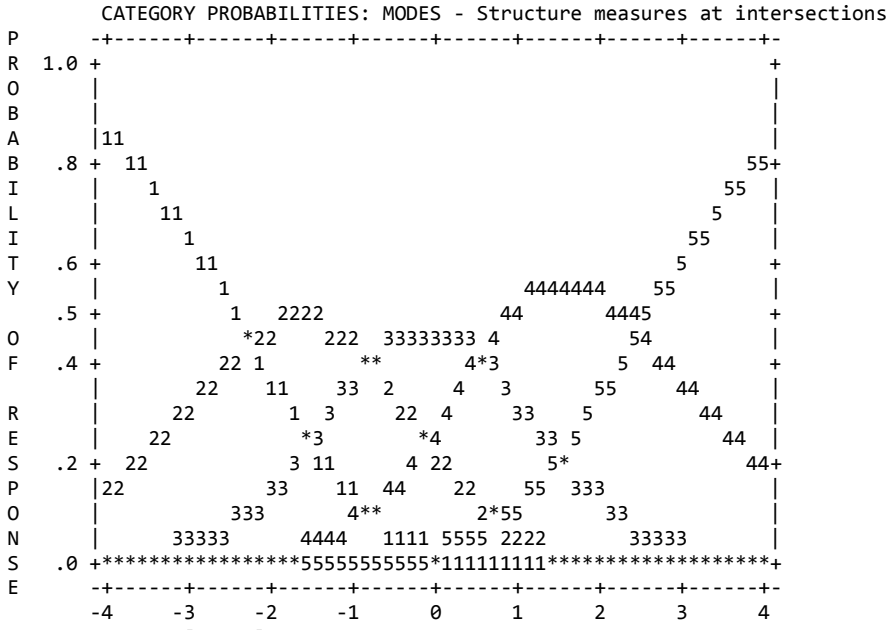
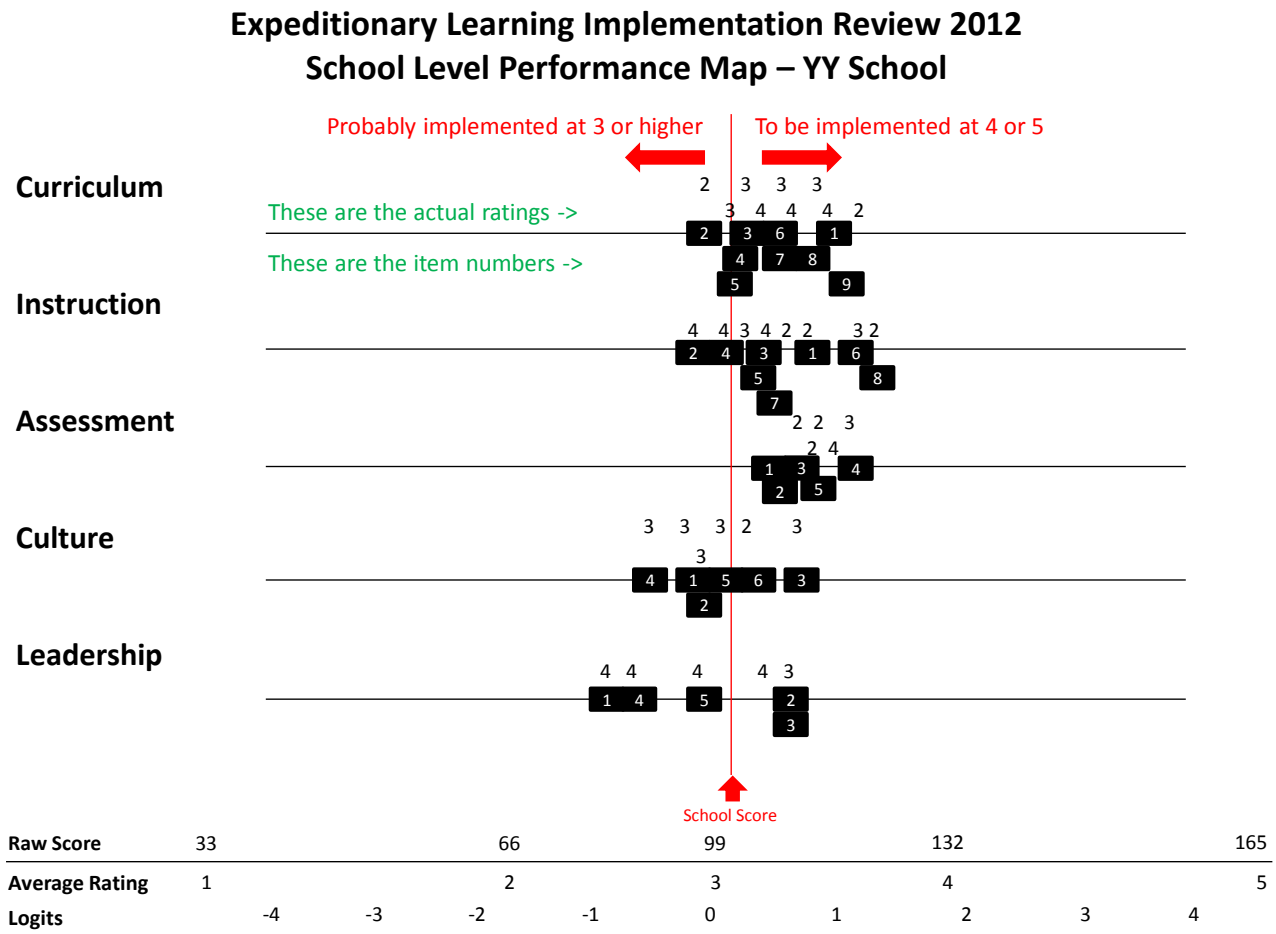


Figure 5. Dimension Map Sample



References

Expeditionary Learning Outward Bound. (2011). Expeditionary Learning Core Practices: A vision for improving schools. NYC: Expeditionary Learning.

Linacre, J.M. (2009). WINSTEPS Rasch Measurement (version 3.73) [Computer program]. Chicago, IL: MESA Press.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut. (Chicago, IL: University of Chicago Press, 1980).

Smith, M.W. & Dickinson, D.K. (2002). *Users Guide to the Early Language & Literacy Classroom Observation Toolkit – Research Edition*. Baltimore: Brookes Publishing Co.

Vernez, G., Karam, R., Mariano, L. T., & DeMartini, C. (2006). Evaluating Comprehensive School Reform Models at Scale: Focus on Implementation. Santa Monica, CA: RAND Corporation. http://www.rand.org/pubs/monographs/MG546.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago, IL: Mesa Press.