TIMSS and PIRLS 2011
Field Test Analysis of Background Questionnaire Scale Items

Using the data from the field test, the measurement properties of each of the items intended to form a background scale were analyzed to ensure that items would be suitable for scaling with the 1-Parameter IRT (Rasch) measurement model following the 2011 data collection. This involved checking the assumption of an underlying unidimensional construct and estimating the reliability of the resulting scale. As valid indicators of effective environments for learning, the background scales should be related to student achievement. Thus, the relationship between the item sets and student achievement in reading, mathematics, and science, also was investigated. The analysis steps taken to determine the measurement properties of each set of field test items are described in the following sections.

Checking for Unidimensionality

There is no absolute criterion for a set of items to constitute a unidimensional scale, but a scale typically may be considered "sufficiently unidimensional" if a single underlying construct is the dominant influence on the item responses (Reckase, 1979; Embretson and Yang, 2006). In factor analytic terms, this implies the existence of a single large factor accounting for most of the covariance among the items.

In the TIMSS and PIRLS field test, the dimensionality of the items for each scale was evaluated using Principal Components Analysis.[1] Field test data from all participants were combined for these analyses. In the analyses all countries were treated equally regardless of their sample size. To evaluate the credibility of a unidimensional underlying construct, an analysis based on the first principal component was conducted. If there was evidence of more than one component, the item set was revised to better conform to a single dimension. That is, items not loading on the first component were flagged for elimination from the scales. Consistent with Comrey and Lee's (1992) rule of thumb, items with factor loadings (correlations between each item and the overall scale) less than 0.3 were eliminated from a scale unless they were considered to have crucial conceptual importance for measuring the construct.

The item sets for most scales were found to be sufficiently unidimensional, although in a few instances, items were identified that did not contribute to the measurement of the construct. Such items were eliminated.

Estimating Reliability (Internal Consistency)

To provide an indicator of the reliability of the proposed scales, Cronbach's Alpha, a measure of internal consistency, was computed for each scale. A scale was considered sufficiently reliable if Cronbach's Alpha

---

1 Principal Components Analyses were conducted using SPSS 16.0.

was at least 0.7 (Nunally & Bernstein, 1991). Most of the proposed scales had Cronbach's Alpha coefficients that exceeded 0.7. A few individual items, however, were identified that did not increase the reliability of the scale and thus were considered for elimination.

Evaluating the Relationship Between the Proposed Scales and Student Achievement

For indicators of effective learning environments, a positive relationship with student achievement is an important aspect of validity. To examine the relationship between the proposed scales and student achievement, a preliminary score was constructed for each scale. This was done by assigning a numerical value to each item response category and summing across the items in the scale. For example, if the response options for a set of 5 questions were *disagree a lot* (coded 1), *disagree a little* (coded 2), *agree a little* (coded 3), and *agree a lot* (coded 4), the maximum score was 20 and the minimum score was 5. The responses were coded so that a high score indicated a supportive learning environment and a lower score a less supportive learning environment.

Although questionnaire responses were provided by students, their parents, teachers, and school principals depending on who completed the questionnaire, the information pertained to the learning environment of the student, and so the field test analysis was conducted with the student as the unit of analysis, consistent with the TIMSS and PIRLS reporting approach. That is, parents', teachers', and school principals' responses were attached to student records, so that scales describing learning environments could be related to student achievement results.

For the field test analysis the score distribution of each scale was divided into three categories: above the 75th percentile (top 25%), above the 25th percentile but not above the 75th percentile (middle 50%), and below the 25th percentile (bottom 25%). Average achievement in reading, mathematics, and science was computed for the students in each scale category. Those students in the most supportive learning environments (top 25%) were expected to have higher achievement than those in the middle category and those in the middle category higher than those in the bottom category. Most of the proposed scales had such a positive relationship with achievement within and across countries and thus were valid indicators of effective learning environments. A few scales, however, lacked the expected relationship with student achievement and were eliminated.

Reviewing the TIMSS and PIRLS 2011 Background Questionnaires

During the summer of 2010, the results of the field test analysis of the questionnaire items were reviewed by the TIMSS Questionnaire Item Review Committee (QIRC), the PIRLS Questionnaire Development Group (QDG), the TIMSS National Research Coordinators, and the PIRLS National Research Coordinators. Most of the field test scale items were found to have good measurement properties and were retained for the main data collection. To minimize response burden, items that did not make a contribution to construct measurement or lacked a relationship with student achievement were not included in the final questionnaires. The field test analysis ensured that the proposed contextual scales were likely to provide countries with valid and reliable indicators of effective environments for learning with the least possible response burden.

The field test analysis resulted in the adoption of items for the final background questionnaires for 37 background scales for TIMSS and PIRLS at the fourth grade and 33 for TIMSS at the eighth grade as summarized below:

Number of Context Questionnaire Scales for the TIMSS and PIRLS 2011 Data Collection

	Grade 4	Grade 8
	Number of Scales	
Home Questionnaire	8	-
Student Questionnaire	11	9
School Questionnaire	8	6
Teacher Questionnaire	10	18*

* 9 scales were based on mathematics teachers' responses and 9 scales were based on science teachers' responses.

References

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis.* Hillsdale, NJ: Lawrence & Erlbaum Associates.

Embretson, S., & Yang, X. (2006). Item Response Theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research.* Hillsdale, NJ: Lawrence & Erlbaum Associates.

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics,* 4, 207-230.