Creating and Interpreting the TIMSS and PIRLS 2011 Context Questionnaire Scales

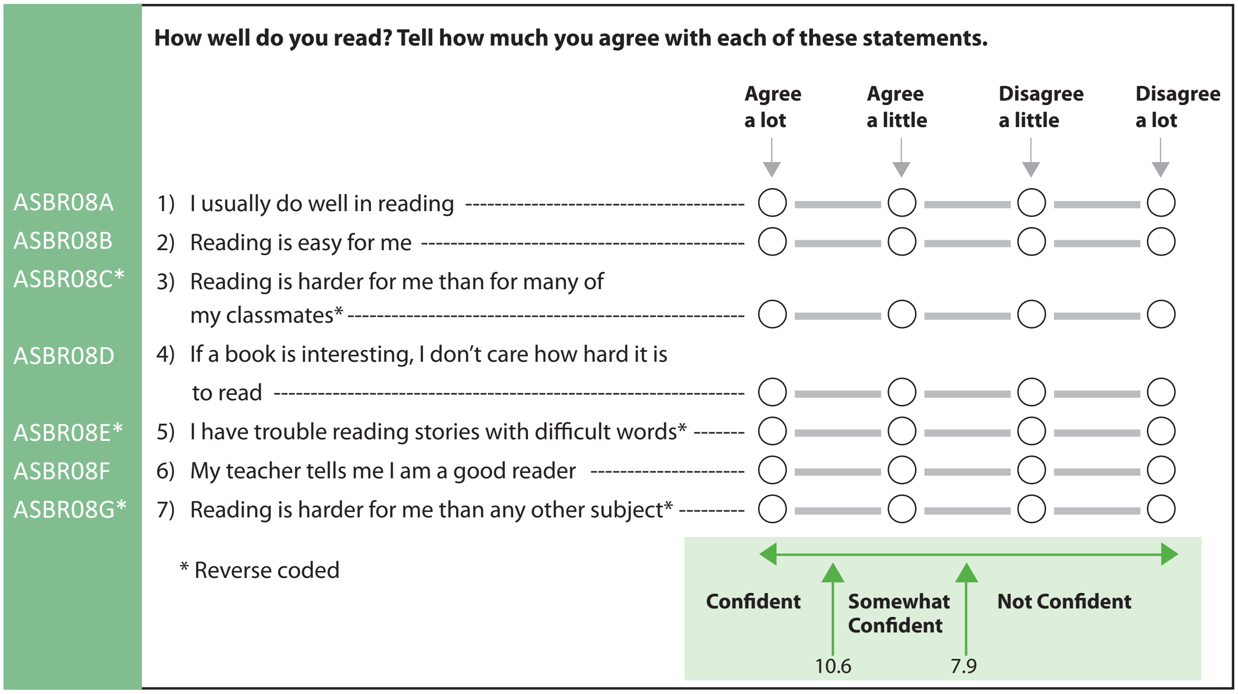
Michael O. Martin
Ina V.S. Mullis
Pierre Foy
Alka Arora

Overview

As described in the **Assessment Framework and Instrument Development** section, most context questionnaire items in TIMSS and PIRLS 2011 were designed to be combined into scales measuring a single underlying latent construct. The scales were constructed using IRT scaling methods, specifically the Rasch partial credit model (Masters and Wright, 1997). As a parallel to the International Benchmarks of achievement in TIMSS and PIRLS, each context scale was divided into regions corresponding to high, middle, and low values on the construct. To facilitate interpretation of the regions, the cutpoints delimiting the regions were defined in terms of combinations of response categories. This chapter describes the procedure for constructing, interpreting, and validating scales based on responses to student, teacher, school, and parent questionnaires.

Reporting Context Questionnaire Scales in TIMSS and PIRLS 2011

As an example illustrating the TIMSS and PIRLS approach to reporting context questionnaire data, Exhibit 1 presents the PIRLS 2011 *Students Confident in Reading* scale. As the name suggests, this scale seeks to measure how confident students feel about their ability to read, in terms of their level of agreement with seven statements about their reading. For each of the seven statements, students were asked to indicate the degree of their agreement with the statement: agree a lot, agree a little, disagree a little, or disagree a lot. Using IRT partial credit scaling, student responses were placed on a scale constructed so that the mean scale score across all PIRLS countries was 10 and the standard deviation was 2. Statements expressing negative sentiment were reverse coded during the scaling (statements 3, 5, and 7). Students **Confident** in their reading had a scale score greater than or equal to the point on the scale corresponding to agreeing a lot, on average with four of the seven statements and a little with three of the statements. Students **Not Confident** in their reading had a score no higher than the point on the scale corresponding to disagreeing a little with four of the statements, on average, and agreeing a little with three of them.



How well do you read? Tell how much you agree with each of these statements.

		Agree a lot	Agree a little	Disagree a little	Disagree a lot
ASBR08A	1) I usually do well in reading	○	○	○	○
ASBR08B	2) Reading is easy for me	○	○	○	○
ASBR08C*	3) Reading is harder for me than for many of my classmates*	○	○	○	○
ASBR08D	4) If a book is interesting, I don't care how hard it is to read	○	○	○	○
ASBR08E*	5) I have trouble reading stories with difficult words*	○	○	○	○
ASBR08F	6) My teacher tells me I am a good reader	○	○	○	○
ASBR08G*	7) Reading is harder for me than any other subject*	○	○	○	○

* Reverse coded

Exhibit 1: Items in PIRLS 2011 Students Confident in Reading Scale

Scaling Procedure

Partial credit IRT scaling is based on a statistical model that relates the probability that a person will choose a particular response to an item to that person's location on the underlying construct. In the *Students Confident in Reading* example scale, the underlying construct is confidence in reading, and students who agree in general with the seven statements are assumed to be more confident in their reading ability and students who disagree with the statements are assumed to be less confident.

The partial credit model is shown below:

$$P_{x_i}(\theta_n) = \frac{e^{\sum_{k=0}^{x}\left(\theta_n - \delta_i + \tau_{ij}\right)}}{\sum_{h=0}^{m} e^{\sum_{k=0}^{h}\left(\theta_n - \delta_i + \tau_{ik}\right)}} \qquad x_i = 0, 1, \ldots, m_i$$

where $P_{x_i}\left(\theta_n\right)$ denotes the probability that person n with location θ_n on the latent construct would choose response level x to item i out of the m_i possible response levels for the item. The item parameter δ_i gives the location of the item on the latent construct and τ_{ij} denotes step parameters for the response levels. For each scale, the scaling procedure involves first estimating the δ_i and τ_{ij} item parameters, and then using the model with these parameters to estimate θ_n, the score on the latent construct, for each on the n respondents. Depending on the scale, respondents may be students, parents, teachers, or school principals.

The TIMSS and PIRLS 2011 context questionnaire scaling was conducted using the ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007).

In preparation for the context questionnaire scaling effort, the TIMSS & PIRLS International Study Center developed a system of production programs that could effectively calibrate the items on each scale using ConQuest and produce scale scores for each scale respondent. Each assessment population (TIMSS fourth grade, TIMSS eighth grade, and PIRLS fourth grade) consisted of approximately 300,000 students, as well as their parents, teachers, and school principals. The estimation of the item parameters, a procedure also known as item calibration, was conducted on the combined data from all countries, with each country contributing equally to the calibration. This was achieved by weighting each country's student data to sum to 500. Exhibit 2 shows the international item parameters for the *Students Confident in Reading* scale. For each item, the delta parameter δ_i shows the estimated overall location of the item on the scale, and the tau parameters τ_{ij} show the location of the steps, expressed as deviations from delta[1].

[1] Although typically the values of the item step parameters estimated from the data are in the same order as the response categories of the item, this is not the case with the tau parameters for the *Students Confident in Reading* scale. However, as described in Adams, Wu, & Wilson (2012), this does not imply that the data do not fit the scaling model, but rather reflect the distribution of respondents across the response categories of the items. As with many of the TIMSS and PIRLS context questionnaire scales, students were very positive in their responses to the items on this scale, with more than half the students reporting that they "agree a lot" to almost all of the items. This does not prevent the scale from effectively summarizing the item responses, but does result in the disordered tau parameters.

Exhibit 2: Item Parameters for Students Confident in Reading Scale

Item	delta	tau_1	tau_2	tau_3
ASBR08A	-1.38588	-0.06832	-0.67050	0.73882
ASBR08B	-1.41524	-0.12326	-0.41288	0.53614
ASBR08C*	-0.70497	-0.15123	0.25172	-0.10049
ASBR08D	-1.01512	0.44367	-0.39057	-0.05310
ASBR08E*	-0.13353	-0.54074	0.47905	0.06169
ASBR08F	-0.85116	-0.29990	-0.45939	0.75929
ASBR08G*	-0.81967	0.16729	0.25788	-0.42517

Once the calibration was complete and international item parameters had been estimated, individual scores for each respondent (students, teachers, principals, or parents) were generated using weighted maximum likelihood estimation (Warm, 1989). All cases with valid responses to at least two items on a scale were included in the calibration and scoring processes.

The scale scores produced by the weighted likelihood estimation are in the logit metric and range from approximately -5 to +5. To convert to a more convenient reporting metric, a linear transformation was applied to the international distribution of logit scores for each scale, so that the resulting distribution across all countries had a mean of 10 and a standard deviation of 2. Exhibit 3 presents the scale transformation constants applied to the international distribution of logit scores for the *Students Confident in Reading* scale to transform them to the (10, 2) reporting metric.

Exhibit 3: Scale Transformation Constants

Scale Transformation Constants	
A = 9.96677	Transformed Scale Score = 9.96677 + 2.18490 • Logit Scale Score
B = 2.18490	

On the TIMSS and PIRLS achievement scales in mathematics, science, and reading, the Low, Intermediate, High, and Advanced International Benchmarks of achievement are specific reference points on the scale that can be used to monitor progress in student achievement. Using a scale anchoring procedure (see **Using Scale Anchoring to Interpret the TIMSS and PIRLS 2011 Achievement Scales**), student performance at each Benchmark is described in terms of the mathematics, science, or reading (depending on the subject) that students reaching that Benchmark know and can do. The percentage of students reaching each of these International Benchmarks can serve as a profile of student achievement in a country.

To provide an analogous approach to reporting the context questionnaire scales, a method was developed to divide each scale into high, middle, and low regions and provide a content-referenced interpretation for these regions. The interpretation is content referenced to the extent that the boundaries of the regions were defined in terms of identifiable combinations of response categories. The particular response combinations that defined the regions boundaries, or cutpoints, were based on a judgment of what constituted a high or low region on each individual scale. For example, based on a consideration of the questions making up the *Students Confident in Reading* scale, it was determined that in order to be in the high region of the scale and labeled "Confident," a student would have to agree a lot, on average, to at least four of the seven statements and agree a little to the other three. Similarly, it was determined that a student who, on average, at most agreed a little with four of the statements and disagreed with the other three would be labeled "Not Confident."

The scale region cutpoints were quantified by assigning a numeric value to each response category, such that each respondent's responses to the scale's questions could be expressed as a "raw score." Assigning 0 to "Disagree a lot," 1 to "Disagree a little," 2 to "Agree a little," and 3 to "Agree a lot" results in raw scores on the *Students Confident in Reading* scale ranging from 0 (disagree a lot with all seven statements) to 21 (agree a lot to all seven). A student who agreed a lot with four statements and agreed a little with the other three would have a raw score of 18 (4×3 + 3×2). Following this approach, a student with a raw score of 18 or more would be in the "Confident" region of the scale. Similarly, agreeing a little with three statements and disagreeing a little with four statements would result in a raw score of 10 (3×2 + 4×1), so that a student with a raw score less than or equal to 10 would be in the "Not confident" region.

A property of a Rasch scale is that each raw score has a unique scale score associated with it. Exhibit 4, presents a raw score-scale score equivalence table for the *Students Confident in Reading* scale. From this table, it can be seen that a raw score of 10 corresponds to a scale score of 7.9 and a raw score of 18 corresponds to a scale score of 10.6. These scale scores were the cutpoints used to divide the scale into the three regions.

Exhibit 4: Equivalence table of the raw score and the transformed scale score

Raw Score	Transformed Scale Score	Cutpoint
0	1.97908	
1	3.87424	
2	4.75552	
3	5.36024	
4	5.83733	
5	6.23940	
6	6.59960	
7	6.92758	
8	7.23353	
9	7.52532	
10	7.80940	7.9
11	8.09145	
12	8.37394	
13	8.67056	
14	8.98289	
15	9.32033	
16	9.69670	
17	10.12500	
18	10.64147	10.6
19	11.29688	
20	12.25876	
21	14.35923	

Validating the TIMSS & PIRLS 2011 Context Questionnaire Scales

As evidence that the context questionnaire scales provide comparable measurement across countries, reliability coefficients were computed for each scale for every country and benchmarking participant and a principal components analysis of the scale items was conducted. Exhibit 5 presents the results of this analysis for the *Students Confident in Reading* scale. The Cronbach's Alpha reliability coefficients generally were at an acceptable level, with most above 0.6 or 0.7, although in a few countries the value was below 0.5. The exhibit also shows the percentage of variance among the scale items accounted for by the first principal component in each country. In most cases this was acceptably high, indicating that the items could be adequately represented by a single scale. The factor loadings of each questionnaire item from the principal components analysis are positive and substantial, indicating a strong correlation between each item and the scale in every country.

Exhibit 5: Cronbach Alpha Reliability Coefficient and Principal Component Analysis of the Items in the PIRLS 2011 Students Confident in Reading Scale

Country Name	Cronbach Alpha Reliability Coefficient	Percent of Variance Explained	Factor Loadings for Each Item						
			ASBR08A	ASBR08B	ASBR08C*	ASBR08D	ASBR08E*	ASBR08F	ASBR08G*
Australia	0.72	40	0.73	0.78	0.71	0.46	0.61	0.34	0.69
Austria	0.75	43	0.78	0.81	0.68	0.39	0.51	0.63	0.71
Azerbaijan	0.58	29	0.39	0.53	0.69	0.28	0.61	0.49	0.68
Belgium (French)	0.66	35	0.71	0.71	0.64	0.32	0.55	0.47	0.66
Bulgaria	0.78	46	0.77	0.83	0.70	0.48	0.51	0.68	0.71
Canada	0.71	40	0.72	0.77	0.73	0.34	0.60	0.39	0.73
Chinese Taipei	0.72	39	0.79	0.81	0.56	0.60	0.27	0.62	0.57
Colombia	0.48	26	0.55	0.57	0.62	0.24	0.45	0.42	0.61
Croatia	0.75	44	0.79	0.78	0.72	0.33	0.44	0.74	0.69
Czech Republic	0.77	44	0.73	0.82	0.72	0.38	0.56	0.64	0.71
Denmark	0.73	41	0.76	0.79	0.75	0.38	0.69	0.41	0.59
England	0.73	42	0.74	0.80	0.75	0.38	0.66	0.26	0.73
Finland	0.69	40	0.76	0.75	0.71	0.24	0.58	0.46	0.75
France	0.69	38	0.73	0.71	0.71	0.29	0.55	0.53	0.66
Georgia	0.63	35	0.69	0.70	0.63	0.20	0.51	0.63	0.61
Germany	0.76	44	0.78	0.80	0.74	0.33	0.54	0.64	0.70
Hong Kong SAR	0.69	36	0.77	0.75	0.56	0.60	0.41	0.48	0.57
Hungary	0.77	46	0.76	0.81	0.73	0.31	0.52	0.76	0.71
Indonesia	0.53	27	0.21	0.28	0.82	0.08	0.65	0.24	0.77
Iran, Islamic Rep. of	0.54	29	0.56	0.64	0.65	0.17	0.52	0.48	0.62
Ireland	0.71	40	0.73	0.78	0.72	0.37	0.62	0.34	0.71
Israel	0.66	35	0.52	0.70	0.69	0.38	0.55	0.57	0.66
Italy	0.66	35	0.69	0.73	0.64	0.27	0.54	0.61	0.58
Lithuania	0.74	43	0.80	0.82	0.70	0.25	0.49	0.67	0.66
Malta	0.71	39	0.68	0.76	0.67	0.39	0.53	0.58	0.66
Morocco	0.38	24	0.04	0.61	0.62	0.25	0.32	0.59	0.62
Netherlands	0.78	46	0.77	0.82	0.76	0.24	0.61	0.60	0.77
New Zealand	0.67	36	0.69	0.74	0.65	0.39	0.53	0.44	0.64
Northern Ireland	0.71	39	0.71	0.75	0.70	0.39	0.64	0.31	0.70
Norway	0.67	37	0.73	0.75	0.64	0.37	0.59	0.42	0.62
Oman	0.54	29	0.60	0.69	0.59	0.20	0.39	0.58	0.57
Poland	0.76	44	0.76	0.79	0.71	0.23	0.66	0.63	0.70
Portugal	0.73	41	0.77	0.77	0.62	0.37	0.58	0.70	0.54
Qatar	0.52	30	0.34	0.56	0.75	-0.07	0.61	0.41	0.74

Exhibit 5: Cronbach Alpha Reliability Coefficient and Principal Component Analysis of the Items in the PIRLS 2011 Students Confident in Reading Scale (Continued)

Country Name	Cronbach Alpha Reliability Coefficient	Percent of Variance Explained	Factor Loadings for Each Item						
			ASBR08A	ASBR08B	ASBR08C*	ASBR08D	ASBR08E*	ASBR08F	ASBR08G*
Romania	0.74	42	0.76	0.78	0.65	0.40	0.52	0.69	0.66
Russian Federation	0.64	40	0.76	0.71	0.71	-0.06	0.48	0.68	0.71
Saudi Arabia	0.58	29	0.38	0.58	0.68	0.34	0.59	0.48	0.65
Singapore	0.69	37	0.72	0.77	0.65	0.48	0.51	0.39	0.65
Slovak Republic	0.77	45	0.76	0.81	0.67	0.33	0.63	0.66	0.70
Slovenia	0.77	45	0.77	0.80	0.67	0.29	0.68	0.70	0.67
Spain	0.61	32	0.70	0.69	0.59	0.43	0.27	0.62	0.55
Sweden	0.74	42	0.74	0.80	0.71	0.36	0.63	0.44	0.75
Trinidad and Tobago	0.68	37	0.69	0.74	0.66	0.35	0.50	0.57	0.65
United Arab Emirates	0.57	30	0.50	0.65	0.67	0.24	0.54	0.43	0.66
United States	0.71	39	0.72	0.77	0.73	0.34	0.55	0.41	0.72
Sixth Grade Participants									
Botswana	0.53	28	0.53	0.59	0.73	0.28	0.34	0.40	0.68
Honduras	0.47	26	-0.02	0.03	0.79	-0.20	0.70	-0.21	0.79
Kuwait	0.59	30	0.15	0.59	0.71	0.35	0.63	0.43	0.73
Morocco	0.40	26	-0.19	0.58	0.69	0.22	0.45	0.50	0.69
Benchmarking Participants									
Alberta, Canada	0.71	40	0.71	0.78	0.75	0.33	0.63	0.31	0.73
Ontario, Canada	0.71	39	0.73	0.76	0.71	0.33	0.60	0.39	0.72
Quebec, Canada	0.71	40	0.72	0.78	0.74	0.30	0.58	0.41	0.72
Maltese - Malta	0.73	40	0.75	0.79	0.63	0.50	0.51	0.64	0.55
Eng/Afr (5) - RSA	0.60	30	0.50	0.59	0.67	0.36	0.58	0.38	0.67
Andalusia, Spain	0.59	31	0.71	0.69	0.55	0.43	0.25	0.62	0.53
Abu Dhabi, UAE	0.56	29	0.48	0.65	0.65	0.25	0.52	0.47	0.64
Dubai, UAE	0.61	32	0.59	0.66	0.69	0.26	0.57	0.41	0.67
Florida, US	0.72	41	0.74	0.79	0.75	0.34	0.52	0.42	0.75
prePIRLS Countries									
Colombia	0.45	24	0.57	0.60	0.47	0.36	0.19	0.65	0.41
South Africa	0.48	26	0.54	0.56	0.63	0.23	0.45	0.42	0.62
Botswana	0.45	29	0.53	0.53	-0.50	0.52	-0.55	0.56	-0.55

As indicators of effective environments for learning, a positive relationship with achievement is an important aspect of validity for the TIMSS and PIRLS context questionnaire scales. For the *Students Confident in Reading* scale, Exhibit 6 presents the Pearson correlation with reading achievement in PIRLS 2011 for each country, together with r-square, the proportion of variance in reading achievement attributable to the *Confident* scale. These figures show a moderate positive relationship in every country. Also shown is the proportion of variance in reading achievement attributable to differences between the regions of the *Confident* scale. This is very similar to the proportion of variance explained by the scale as a whole, indicating that dividing the scale into regions retains most of the relationship between the scale and achievement.

Exhibit 6: Relationship Between Students Confident in Reading Scale and PIRLS 2011 Reading Achievement

Country	Pearson's Correlation with Reading Achievement		Variance in Reading Achievement Accounted for by Difference Between Regions of the Scale (η^2)
	(r)	(r^2)	
Australia	0.45	0.21	0.20
Austria	0.36	0.13	0.12
Azerbaijan	0.27	0.07	0.07
Belgium (French)	0.40	0.16	0.15
Bulgaria	0.41	0.17	0.17
Canada	0.40	0.16	0.14
Chinese Taipei	0.34	0.12	0.10
Colombia	0.34	0.12	0.08
Croatia	0.37	0.14	0.13
Czech Republic	0.38	0.15	0.15
Denmark	0.44	0.20	0.18
England	0.42	0.18	0.16
Finland	0.37	0.14	0.14
France	0.40	0.16	0.14
Georgia	0.35	0.12	0.10
Germany	0.39	0.15	0.14
Hong Kong SAR	0.34	0.12	0.10
Hungary	0.49	0.24	0.22
Indonesia	0.29	0.08	0.08
Iran, Islamic Rep. of	0.33	0.11	0.11
Ireland	0.37	0.14	0.13
Israel	0.41	0.17	0.16

Exhibit 6: Relationship Between Students Confident in Reading Scale and PIRLS 2011 Reading Achievement (Continued)

Country	Pearson's Correlation with Reading Achievement		Variance in Reading Achievement Accounted for by Difference Between Regions of the Scale (η^2)
	(r)	(r^2)	
Italy	0.30	0.09	0.08
Lithuania	0.43	0.19	0.17
Malta	0.46	0.21	0.21
Morocco	0.30	0.09	0.08
Netherlands	0.31	0.10	0.09
New Zealand	0.44	0.19	0.17
Northern Ireland	0.37	0.14	0.13
Norway	0.35	0.12	0.13
Oman	0.41	0.17	0.16
Poland	0.45	0.20	0.22
Portugal	0.40	0.16	0.15
Qatar	0.49	0.24	0.20
Romania	0.47	0.22	0.22
Russian Federation	0.37	0.14	0.13
Saudi Arabia	0.43	0.19	0.18
Singapore	0.38	0.15	0.14
Slovak Republic	0.41	0.17	0.15
Slovenia	0.43	0.19	0.18
Spain	0.35	0.12	0.11
Sweden	0.39	0.15	0.14
Trinidad and Tobago	0.50	0.25	0.23
United Arab Emirates	0.42	0.18	0.16
United States	0.40	0.16	0.15
International Median	**0.40**	**0.16**	**0.14**
Sixth Grade Participants			
Botswana	0.49	0.24	0.21
Honduras	0.32	0.10	0.07
Kuwait	0.39	0.15	0.13
Morocco	0.31	0.10	0.08
Benchmarking Participants			
Alberta, Canada	0.40	0.16	0.16
Ontario, Canada	0.39	0.15	0.14
Quebec, Canada	0.40	0.16	0.13
Maltese - Malta	0.36	0.13	0.12

Exhibit 6: Relationship Between Students Confident in Reading Scale and PIRLS 2011 Reading Achievement (Continued)

Country	Pearson's Correlation with Reading Achievement		Variance in Reading Achievement Accounted for by Difference Between Regions of the Scale (η^2)
	(r)	(r^2)	
Eng/Afr (5) - RSA	0.44	0.20	0.17
Andalusia, Spain	0.34	0.12	0.11
Abu Dhabi, UAE	0.45	0.20	0.18
Dubai, UAE	0.39	0.16	0.15
Florida, US	0.43	0.19	0.16

Exhibit 6: Relationship Between Students Confident in Reading Scale and prePIRLS 2011 Reading Achievement

Country	Pearson's Correlation with Reading Achievement		Variance in Reading Achievement Accounted for by Difference Between Regions of the Scale (η^2)
	(r)	(r^2)	
Colombia	0.44	0.19	0.17
South Africa	0.36	0.13	0.09
Botswana	0.43	0.19	0.16
International Median	**0.43**	**0.19**	**0.16**

The item parameter estimates and item and scale statistics presented above for the *Students Confident in Reading* scale are available for each of the TIMSS and PIRLS 2011 context questionnaire scales at **Creating and Interpreting the TIMSS and PIRLS 2011 Context Questionnaire Scales Details**.

References

Adams, R.J., Wu, M.L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. Educational and Psychological Measurement, 72(4), 547–573. Retrieved from http://epm.sagepub.com/content/72/4/547

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In: M.J. van de Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. Berlin: Springer.

Warm, T.A. (1989) Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, 54(3), 427–450.

Wu, M.L., Adams, R.J, Wilson, M.R., & Haldane, S. (2007). Conquest 2.0 [computer software]. Camberwell, Australia: Australian Council for Educational Research.