



IRT scales for Self-reported Test-taking Motivation of Swedish students in International Surveys

Denise Reis Costa¹ and Hanna Eklöf²

¹ University of Oslo,

 Centre for Educational Measurement, Gaustadalleen, 30, 0373 Oslo, Norway

 d.r.costa@cemo.uio.no

² Department of Applied Educational Science, Umeå University, Umeå, Sweden

hanna.eklof@umu.se

Abstract. This study aims at modeling the self-reported test-taking motivation items in PISA and TIMSS Advanced studies for Swedish students using an IRT approach. In the last two cycles of the assessments, six test-specific items were included in the Swedish student questionnaires to evaluate pupil's effort, motivation and how they perceived the importance of the tests. Using a Multiple-Group Generalized Partial Credit model (MG-GPCM), we created an IRT motivation scale for each assessment. We also investigated measurement invariance for the two cycles of PISA (i.e., 2012 and 2015) and of TIMSS Advanced (i.e., 2008 and 2015). Results indicated that the proposed scales refer to unidimensional constructs and measure reliably students' motivation (Cronbach's alpha above 0.78). Differential item functioning across assessment cycles was restricted to two criteria (RMSD and DSF) and had more impact on the latent motivation scale for PISA than for TIMSS Advanced. Overall, the test-taking motivation items fit well the purpose of a diagnostic of test-taking motivation in these two surveys and the proposed scales highlighted the slight increase of pupils' motivation across the assessment cycles.

Keywords: test-taking motivation, PISA, TIMSS, IRT

1 Introduction

Regarded as a regular feature of many educational assessment systems, international surveys, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), have a major impact on the discussions about educational quality in many countries around the world (Wagemaker, 2013).

Created in 2000 by the Organisation for Economic Co-operation and Development (OECD), PISA assesses 15-year-old student's literacy in science, mathematics, and reading. First conducted in 1995 by the International Association

for the Evaluation of Educational Achievement (IEA), TIMSS Advanced assesses students in the final year of secondary school enrolled in special advanced mathematics and physics programs or tracks.

As there are no personal benefits related to students' performance on the test, PISA and TIMSS Advanced are usually low-stakes tests for participating students, but high-stakes for other stakeholders. In this scenario, some pupils may lack the motivation to do their best on the test and the results, therefore, can be an underestimation of their knowledge (Eklöf and Nyroos, 2013).

In this study, we created a test-taking motivation scale for the PISA and TIMSS Advanced assessments in Sweden and we investigated the quality of these measures. These scales were built using six test-taking motivation items created specifically for each assessment and based on the expectancy-value model (Wigfield and Eccles, 2000). In particular, item response theory (IRT) analysis was used to examine the psychometric properties of the test-taking motivation items and their measurement invariance over the last two cycles of each assessment. Moreover, the differences in test motivation were studied across the different test administrations.

2 Methods

2.1 Data

A total of 4736 Swedish students participated in the PISA 2012 cycle and 5458 in 2015. In TIMSS Advanced, the 2008 assessment counted with 2303 Swedish students and in the 2015 cycle, 3937. In their questionnaires, six national items (Table 1) referring to their effort, motivation and how they perceived the importance of that specific test were presented. All items use four-point Likert-type scales and, except for negative items, they were reversed so that score categories are in increasing order with respect to the target trait, test-taking motivation.

The number of students who answered the test-taking motivation item varied by assessment and cycle. The percentage of students who had at least one missing response is larger for PISA 2012 (11%), followed by PISA 2015 (5%), and TIMSS Advanced 2008 (2%) and 2015 (1%). One possible reason for the missing data is that the student did not reach the end of the questionnaire where the motivation items were located. In PISA 2012, for example, about 80% of the students with missing responses omitted all the motivation items.

To handle this, we imputed the cases using the proportional odds model with students' background information. This analysis was conducted using the *mice* package (Buuren and Groothuis-Oudshoorn, 2010). As background variables, we used immigrant status, language most often spoken at home and gender for both assessments and for PISA the students' economical and socio-cultural status on top of that. A sensitivity analysis comparing the IRT item parameter estimates with imputed data and with listwise case deletion was carried out and no substantial difference on the estimates and respective interval confidences were found. Thus, we proceeded the analyses using the complete cases.

Table 1: Item description and percentage of students with missing data by assessment and cycle

Item Code	Description	Percentage of missing response		
		PISA 2012	PISA 2015	
MOTIV_R	I felt motivated to do my best on the PISA test	10.6%	4.4%	
GODEFF_R	I engaged in good effort throughout the PISA test	10.8%	4.7%	
DIDBES_R	I did my best on the PISA test	11.1%	4.7%	
WORKIT_R	I worked on the tasks in the test without giving up even if some tasks felt difficult	10.9%	4.5%	
IMPWEL_R	Doing well on the PISA test was important to me	10.7%	4.7%	
IMP2_R	Doing well on the PISA test meant a lot to me	10.8%	4.8%	
		TIMSS 2008	Adv. TIMSS 2015	Adv.
MOTIV_R	I felt motivated to do my best on this test	2.3%	1.0%	
DIDBE_R	I gave my best effort on this test	1.8%	1.2%	
WORKLR	I worked on each item in the test and persisted even when the task seemed difficult	2.0%	1.1%	
NOCONC	I did not give this test my full attention while completing it	1.6%	0.9%	
NOEFF	I tried less hard on this test as I do on other tests we have at school	1.9%	1.8%	
NOWORK	While taking this test, I could have worked harder on it	1.7%	2.1%	

Note: The suffix ”_R” refers to the reversing of the items. All PISA items were reversed from the original response scale: Strongly disagree (4), Disagree (3), Agree (2) and Strongly agree (1). Items were scored so that a low value is always indicative of a more negative attitude towards the test in terms of perceived importance and reported invested effort. Likewise, three items in the TIMSS Advanced assessments were reversed: MOTIV_R, DIDBE_R and WORKLR. In TIMSS Advanced, the original response scale was: Disagree a lot (4), Disagree (3), Agree (2) and Agree a lot (1).

2.2 Statistical analyses

Descriptive analysis. The percentage of students' agreement with the test-taking motivation items was illustrated in a radar plot. For this analysis, negative items were reversed and the response options dichotomized (with the highest value referring to the two more positive response categories of the attitudes scale).

Reliability. We computed the Cronbach's alpha reliability coefficient. It ranges between 0 and 1, with higher values indicating higher internal consistency of the scale. Commonly accepted cut-off values are 0.9 to signify excellent, 0.8 for good, and 0.7 for acceptable internal consistency (OECD, 2017).

Dimensionality. An analysis of the eigenvalues was done. Using the polychoric correlation matrix, the principal axis factor analysis and the minimum residual solution to estimate the communalities, eigenvalues were calculated using the psych package (Revelle, 2014). The eigenvalues communicate variance and guide the factor selection process by conveying whether a given factor explains a considerable portion of the total variance of the observed measures (Brown, 2014). We used the Kaiser criterion, where eigenvalues above 1.0 provide an indication of unidimensionality of the latent structure.

IRT analyses. The analyses were conducted in four steps. The first step was related to the analysis of the item parameters through the Multiple-Group Generalized Partial Credit model (MG-GPCM) approach, considering each assessments cycles as a group and the estimated item parameter equal (invariant) across groups (Model 1). In the second step, an analysis of the differential item functioning (DIF) over time was carried out. For those items that did not present an indication of DIF, their parameter estimates were fixed across the groups (anchor items) in the third step of the analysis (Model 2). Finally, we estimated the individual scores for each assessment.

The MG-GPCM is based on the assumption that the two-parameter dichotomous response model governs the probability of selecting the k -th category over the $(k-1)$ category by (Muraki, 1999):

$$P_{gjk}(\theta_g) = \frac{\exp[\sum_{r=1}^k Z_{gjr}(\theta_g)]}{\sum_{m=1}^{K_j} \exp[\sum_{r=1}^m Z_{gjr}(\theta_g)]}, \quad (1)$$

where: $Z_{gjr}(\theta_g) = Da_{gj}(\theta_g - b_{gjr})$, a_{gj} is the slope parameter for group g and item j , b_{gjr} is the item category parameter for group g , item j , and category r , D is equal to 1.7, generally inserted to make the logit scale comparable to a normal metric. The latent trait, θ_g , is generally assumed to be normally distributed for each group ($g = 1, \dots, G$). In this study, we use the slope-intercept parameterization implemented in the mirt package (Chalmers, 2012), where

$Z_{gjr}(\theta_g) = a_{gj}\theta_g + d_{gjr}$, where a_{gj} is the slope parameter for group g and item j , d_{gjr} is the intercept parameter for group g , item j , and category r .

For identification purposes, the mean and variance of the reference group (in case of PISA, the 2012 cycle and, for TIMSS Advanced, the 2008 assessment) were fixed to 0 and 1, respectively.

Since one of the advantages of the MG-GPCM approach is its flexibility to estimate item parameters separately for each group, we detected DIF using two criteria and then evaluated a second MG-GPCM model fixing anchor items. For the DIF analysis, we calculated the root mean square deviance (RMSD) for each item using the *tam* package (Kiefer, Robitzsch, and Wu, 2015) and the differential step functioning (DSF) using the DIFAS software (Penfield, 2005). The cut-off criteria to flag the item with DIF was an RMSD value greater than 0.3 (OECD, 2017) or large levels of DSF effect (i.e., the log-odds ratio estimator is greater than or equal to 0.64 in absolute value) as suggested in the Penfield’s classification scheme (Penfield, 2008).

For the best model, individual scores were generated using weighted maximum likelihood (WLE) estimation (Warm, 1989) and were transformed to scales with a mean of 0 and a standard deviation of 1 for the reference group.

2.3 Student weights

It is usual in these assessments to use a type of student weight (called “senate weights” in PISA) such that all countries contribute equally to the estimation of the item parameters. On PISA 2015, for example, a senate weight was constructed to sum up to the target sample size of 5000 within each country (OECD, 2017). Since the focus of this study is related only to the Swedish samples, the student weights were not included in this study.

3 Results

3.1 Descriptive analysis

There was a significant increase in reported test-taking motivation between 2012 and 2015 in PISA. From Figure 1, we can see a difference of more than 15 percentage points on student’s agreement for all items in PISA, except on the GODEFF_R item (“I engaged in good effort throughout the PISA test”) where the difference was only 5. For TIMSS Advanced, on the other hand, the highest difference between cycles was on the DIDBE_R item (“I did my best”), where 9% of the Swedish students agree or agree a lot with this statement.

3.2 Reliability and dimensionality

From Table 2, we can see that the internal consistency evaluated by Cronbach’s alpha was at an acceptable level, with all values above 0.78. This measure was higher in the PISA assessments than in TIMSS Advanced. Using the Kaiser criterion, the eigenvalues suggest that the items could be adequately represented by a unidimensional scale for each assessment.

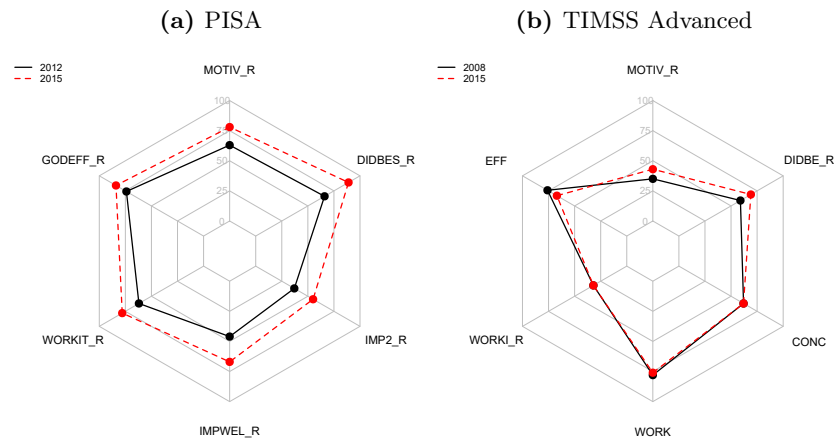


Fig. 1: Percentage of agreement by assessment and cycle. Negative items were reversed

Table 2: Cronbach's α and eigenvalues by assessment and cycle

Assessment	Cycle	α	Eigenvalue 1	Eigenvalue 2
PISA	2012	0.89	4.02	0.27
	2015	0.86	3.56	0.37
TIMSS Advanced	2008	0.79	2.78	0.23
	2015	0.82	3.09	0.27

3.3 IRT analyses

Tables 3 and 4 present the item parameter estimates through the MG-GPCM approach. By considering the item parameters invariant across the assessment cycles (Model 1), there is an improvement of half standard deviation from PISA 2012 to 2015 and about one quarter for TIMSS Advanced cycles.

Table 3: Item parameter estimates by cycle - PISA

Item code	Parameter	Model 1		Model 2	
		2012	2015	2012	2015
MOTIV_R	a1	2.304	--	2.972	2.817
	d1	2.830	--	4.265	3.309
	d2	3.713	--	0.974	0.957
	d3	0.989	--	-3.338	-3.111
GODEFF_R	a1	1.821	--	2.233	--
	d1	3.166	--	4.294	--
	d2	4.653	--	1.549	--
	d3	2.010	--	-2.994	--
IMPWEL_R	a1	2.661	--	3.239	2.957
	d1	3.109	--	4.080	3.235
	d2	2.888	--	-0.382	-0.050
	d3	-0.780	--	-4.388	-3.895
WORKIT_R	a1	1.669	--	2.145	--
	d1	2.774	--	3.752	--
	d2	3.493	--	0.798	--
	d3	1.336	--	-2.632	--
IMP2_R	a1	2.510	--	2.927	--
	d1	2.532	--	2.988	--
	d2	1.570	--	-0.994	--
	d3	-2.464	--	-4.574	--
DIDBES_R	a1	2.245	--	2.816	2.576
	d1	3.201	--	4.260	4.409
	d2	4.627	--	1.148	2.137
	d3	2.926	--	-2.505	-1.676
Group	MEAN	0	0.53	0	0.47
Group	VAR	1	0.84	1	0.79
Number of parameters		26		38	
Log-likelihood		-57971.04		-57391.65	
AIC		115994.10		114859.30	
BIC		116182.10		115134.00	

Note: The symbol "--" indicates that the estimates are equal to the 2012 column.

Table 5 indicates good item fit for all test-motivation items for the Model 1 using the RMSD criterion. Comparing the levels of DSF effect, however, three items on the PISA dataset (MOTIV_R, IMPWEL_R, and DIDBES_R) and one item on TIMSS Advance data (NOWORK) present large DIF. Thus, we estimated the item parameters using anchor items in the analysis and freely-estimated parameters for those with large DSF (Model 2). Results indicate that flagged items were more discriminative for PISA 2012 and TIMSS Advanced 2008 than the following cycles. According to the AIC and BIC criteria, Model 2 was the best model for both assessments.

Table 4: Item parameter estimates by cycle - TIMSS Advanced

Item code	Parameter	Model 1		Model 2	
		2008	2015	2008	2015
DIDBE_R	a1	2.039	--	2.391	--
	d1	3.527	--	4.175	--
	d2	4.352	--	0.840	--
	d3	1.660	--	-3.255	--
NOCONC	a1	0.671	--	1.093	--
	d1	1.002	--	1.573	--
	d2	0.307	--	-0.808	--
	d3	-0.936	--	-2.568	--
NOEFF	a1	1.363	--	1.896	--
	d1	0.456	--	0.812	--
	d2	-0.802	--	-1.734	--
	d3	-3.332	--	-3.986	--
WORKLR	a1	0.835	--	1.218	--
	d1	1.300	--	1.849	--
	d2	0.654	--	-0.806	--
	d3	-1.102	--	-2.928	--
MOTIV_R	a1	1.406	--	1.859	--
	d1	1.334	--	1.828	--
	d2	0.601	--	-0.984	--
	d3	-1.952	--	-3.625	--
NOWORK	a1	1.217	--	1.868	1.774
	d1	0.013	--	0.841	-0.046
	d2	-1.465	--	-1.968	-2.279
	d3	-4.237	--	-4.404	-4.559
Group	MEAN	0	0.22	0	0.27
Group	VAR	1	1.35	1	1.37
Number of parameters		26		30	
Log-likelihood		-41786.07		-41424.06	
AIC		83624.15		82908.12	
BIC		83799.36		83110.29	

Note: The symbol "--" indicates that the estimates are equal to the 2012 column.

Table 5: The root mean square deviation (RMSD) and the item-level log-odds ratio estimate for testing the differential step functioning for each assessment

Assessm.	Item Code	RMSD		Step 1 (SE)	Step 2 (SE)	Step 3 (SE)
PISA		2012	2015			
	MOTIV_R	0.038	0.025	0.941 (0.116)	0.111 (0.068)	0.123 (0.079)
	GODEFF_R	0.043	0.027	0.522 (0.173)	0.344 (0.069)	0.385 (0.075)
	IMPWEL_R	0.031	0.024	0.709 (0.105)	-0.195 (0.062)	-0.042 (0.092)
	WORKIT_R	0.041	0.029	0.098 (0.139)	-0.037 (0.060)	-0.083 (0.072)
	IMP2_R	0.033	0.016	0.275 (0.086)	0.078 (0.063)	0.013 (0.106)
	DIDBES_R	0.066	0.040	-0.172 (0.161)	-1.002 (0.079)	-0.500 (0.063)
TIMSS		2008	2015			
Adv.	DIDBE_R	0.028	0.021	-0.224 (0.132)	-0.442 (0.081)	-0.175 (0.096)
	NOCONC	0.054	0.044	0.424 (0.086)	0.278 (0.073)	-0.252 (0.105)
	NOEFF	0.036	0.019	0.324 (0.076)	-0.338 (0.088)	-0.196 (0.130)
	WORKLR	0.042	0.034	-0.056 (0.087)	-0.314 (0.071)	-0.259 (0.110)
	MOTIV_R	0.023	0.024	-0.270 (0.088)	-0.100 (0.077)	-0.346 (0.116)
	NOWORK	0.065	0.036	0.756 (0.073)	0.189 (0.096)	0.017 (0.157)

Figure 2 shows the distribution of the individual scores for Model 2 for each assessment in the logit metric with measured values ranging from -3 to 3. An increase in students' test motivation is observed, especially in the PISA assessment. With these measures, it is possible to monitor the test-taking motivation across assessment cycles and analyze the relationship between the reported motivation and other student measures, such as performance.

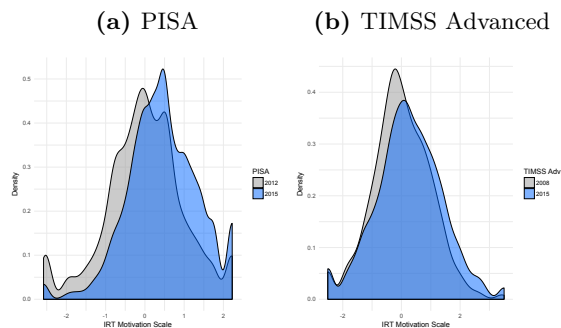


Fig. 2: Distribution of the WLE scores for Model 2 (anchor items)

4 Discussion

In this work, we evaluated two test-taking motivation scales included in the Swedish student questionnaire of two large-scale international assessments. Re-

sults indicated that both scales are unidimensional and, while the item parameters were largely stable across the two cycles of TIMSS Advanced, half of the items in PISA showed some DIF. Our findings also indicated that there was a slight increase in test-taking motivation in PISA 2015 in comparison to 2012 which can be related to the change of the test mode administration from paper and pencil to computer-based across these two PISA cycles.

For future studies, we intend to expand these analyses to carry out studies of PISA log-file data to evaluate how the self-reported test-taking motivation measures agree with student behaviors during the administration of the test. With information of response times, for example, we can further investigate the construct validity of these measures.

References

- Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–68.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.
- Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510.
- Kiefer, T., Robitzsch, A., & Wu, M. (2015). Tam: Test analysis modules. *R package*.
- Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*(3), 217–232.
- OECD. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. doi: 10.1787/9789264255425-en
- Penfield, R. D. (2005). Difas: differential item functioning analysis system. *Applied Psychological Measurement*, *29*(2), 150–151.
- Penfield, R. D. (2008). Three classes of nonparametric differential step functioning effect estimators. *Applied Psychological Measurement*, *32*(6), 480–501.
- Revelle, W. (2014). Psych: Procedures for psychological, psychometric, and personality research. *R package*.
- Wagemaker, H. (2013). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (p. 11-35). New York: Chapman Hall/CRC.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81.