ESM 2: Supplementary Material Regarding the Measures in the Study

Reading Self-Concept (Möller & Bonerad, 2007)

Table 1. Reading self-concept items (translated from German), means, standard deviations, item-total correlations, number of valid/missing cases and scale reliability (Cronbach's coefficient α).

#	Item wording	*M*	*SD*	r_{it}	N_{valid}	$N_{missing}$	α
01	Sometimes I have a hard time really understanding a text.*	3.03	0.76	.50	303	2	
02	Reading a text, I often do not know all the words.*	3.09	0.87	.41	304	1	.65
03	I understand texts quickly and very well.	3.09	0.81	.34	304	1	
04	I often have to re-read texts to fully understand them.*	3.07	0.94	.48	303	2	

Notes. Items with * were recoded for sum score calculation.

Cases with incomplete data on the scale (n = 4) were excluded from the correlational analyses.

Interest in Science

Table 2. Interest in science items (translated from German), means, standard deviations, item-total correlations, number of valid/missing cases, and scale reliability (Cronbach's coefficient α).

#	Item wording	*M*	*SD*	r_{it}	N_{valid}	$N_{missing}$	α
01	Science interests me very much.	2.81	0.96	.75	299	6	
02	In my free time I enjoy spending time on nature and science.	2.32	0.93	.60	302	3	
03	I like to read books or watch programs about science.	2.33	0.99	.55	303	2	.83
04	Science is one of my favorite subjects in school.	2.35	1.01	.71	275	30	
05	I am very good in answering science related questions.	2.59	0.74	.53	294	11	

Notes. Cases with incomplete data on the scale (n = 36) were excluded from the correlational analyses.

Items are self-worded adaptions from background questionnaires concerning students' general attitude towards science (e.g., interest, enjoyment, science-related activities) that were applied in TIMSS 2011 (cf. Martin & Mullis, 2012) and PISA 2006 (cf. OECD, 2009).

KFT N2 (Heller & Perleth, 2000)

Table 3. Scale mean, standard deviation, item-total correlation, number of valid/missing cases, and scale reliability (Cronbach's coefficient α) in the tested sample for the KFT N2 reasoning task.

Test versions	*M*	*SD*	r_{it}	N_{valid}	$N_{missing}$	α
KFT for 5th Grade students (25 items)	18.08	6.06	.29. –.73	124	0	.91
KFT for 6th Grade students (25 items)	18.02	6.38	.36. –.67	181	0	.92

Test-Taking Effort (Kunter et al., 2002)

Table 4. Mean, standard deviation, number of valid/missing cases for the 'effort thermometer' and instructions for students (item and instruction translated from German).

#	Item wording	*M*	*SD*	N_{valid}	$N_{missing}$
01	How much effort are you going to invest in the following science test?	8.08	1.91	305	0

The item was accompanied by the following introduction/instruction:

Please try to imagine an actual situation (at school or in some other context) that is highly important to you personally, so that you would try your very best and put in as much effort as you could to do well.

In this situation you would mark the highest value on the 'effort thermometer' as shown below:

→ VISUALIZATION OF THE EFFORT THERMOMETER (HIGHEST VALUE SELECTED)

Compared to the situation you have just imagined, how much effort are you willing to put into doing this test?

→ VISUALIZATION OF THE EFFORT THERMOMETER FOR THE ACTUAL RATING

Notes. The 'effort thermometer' was adapted from PISA (Kunter et al., 2002) as a measure of students' *planned effort* in the science test. To make sure that students were not influenced in their rating by the (manipulated) test items or their solution success, we administered the measure before the science test and modified the wording accordingly.

In contrast to the application of the measure in PISA 2003 (cf. Butler & Adams, 2006), because of relatively high reading demands for students in the 5th and 6th Grades, the introductory text was read out loud by the test administrator to make sure that all students understand the instructions.

Item-Solving Pleasure/ Test-Taking Pleasure

Table 5. Mean, standard deviation, and number of valid/missing cases for students' mean item-solving pleasure rating (item translated from German).

#	Item wording	*M*	*SD*	N_{valid}	$N_{missing}$
01	Working on this item was fun for me.*	2.68	0.83	305	0

Note. * Reported parameters refer to the mean of person means across all presented items (the measure was repeatedly assessed for every presented item according to the multimatrix design (cf. Electronic Supplementary Material 1).

Table 6. Means, standard deviations, range of means, and range of standard deviations, mean percentage of missing values and range of missing values for students' repeated item-solving pleasure rating, reported separately for text-only and text-picture items.

Parameter range for single measures	*M (SD)**	$M_{(Range)}$	$SD_{(Range)}$	$M(\%_{missing})$	$\%_{missing}$
Rating of Text-Only Items	2.61 (1.2)	2.20–2.96	1.01–1.18	1.1 %	0–4.7 %
Rating of Text-Picture Items	2.75 (1.2)	2.32–3.06	0.97–1.32	0.5 %	0–2.9 %

Note. * Reported parameters refer to the mean of person means across all ratings for text-only and text-picture items. According to the multimatrix structure of the measures, we report only percentages of omitted missing values and not missing values by design.

All single measures of item-solving pleasure were IRT-scaled with a Rating Scale Model (Andrich, 1978); WLE estimates for every student are reported in the study as an overall measure of ***Test-Taking Pleasure***. EAP/PV reliability was estimated as .95 (the mean was fixed at $M = 0$ in the estimation procedure; $SD = 2.11$).

Overall Rating of the Pictures

Table 6. Mean, standard deviation and number of valid/missing cases for students' overall rating of the pictures (item translated from German).

#	Item wording	*M*	*SD*	N_{valid}	$N_{missing}$
01	I liked the pictures in the test.	3.12	0.96	300	5

Note. We applied this item to assess students' deliberate reflection on the pictures (in contrast to the repeated item-solving pleasure ratings during the test, which also provide insights into students' attitude towards the pictures, but without their awareness).

Science Test (cf. IEA, 2013; Martin & Mullis, 2012)

Table 7. Mean item solution frequencies and standard deviations, range of mean item solution frequencies and range of standard deviations and percentage of missing values, reported separately for text-only and text-picture items.

Solution Frequency	*M**	*SD**	$M_{(Range)}$	$SD_{(Range)}$	$\%_{missing}$
Text-Only Items	.63	.18	.18–.97	.02–.05	0 %
Text-Picture Items	.68	.18	.15–.97	.02–.05	0 %

Note. * Reported parameters refer to the mean of item means for text-only and text-picture items.

Due to the rotated multimatrix design, item solving frequencies can only provide an overview of item difficulty ranges; they are not intended for interpretation. Also, we report only omitted missing values and not missing values by design.

Table 8. Weighted mean square (WMNSQ) statistics for estimated item parameters using a Rasch model. WMNSQ means, standard deviations, WMNSQ ranges, numbers of items with a WMNSQ >1.15 and according percentage of significant *t*-values (>1.96) and EAP/PV reliability.

WMNSQ[1]	*M*	*SD*	*Range*	$N_{(\geq 1.15)}$	$N_{(t>1.96)}$	$\%_{(t>1.96)}$	*EAP/PV*
Rasch Model (M2)							
96 item parameters	0.99	0.12	0.76–1.28	5	5	5.2%	.81

Note. [1] WMNSQ values have been estimated using ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007)

References

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581–594. doi:10.1177/014662167800200413

Butler, J., & Adams, R. J. (2006). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*, 279-304.

Heller, K. A., & Perleth, C. (2000). *KFT 4-12+R – Kognitiver Fähigkeits-Test für 4. bis 12. Klassen, Revision*. Göttingen, Germany: Beltz.

International Association for the Evaluation of Educational Achievement [IEA] (2013). TIMSS 2011 assessment released science items. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://nces.ed.gov/timss/pdf/TIMSS2011 _G4_Science.pdf

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., … Weib, M. (2002). *German scale handbook for PISA 2000*. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Möller, J., & Bonerad, E. M. (2007). Fragebogen zur habituellen Lesemotivation. *Psychologie in Erziehung und Unterricht*, *54*, 259–267.

OECD. (2009). *PISA 2006 Technical Report*. Paris: OECD. doi:10.1787/9789264048096-en

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