Appendix 4: Additional Statistical Analyses

Part A: Results Multiple Regression Analysis

Tab.1 Resulting models from multiple regression analysis. Steps show the inclusion (forced entry) of grade level and crystallized intelligence as the two major predictors for energy test scores. See 'Note' for model summary. *B* = regression coefficients, *SE B* = standard error of regression coefficient, β = standardized regression coefficients

		B	***SE B***	***β***
Step 1				
	Constant	-.16	.02	
	Grade level	.06	.00	.60**
Step 2				
	Constant	-.20	.02	
	Grade level	.05	.00	.48**
	Crystallized intelligence	.40	.04	.28**
Note: R^2 = .36 (Step 1), $p < .001$; R^2 = .44, *adj*. R^2 = .43 (Step 2); ** $p < .001$				

Part B: Methods and Results from Rasch analysis

1. Methods

1*pl* Rasch models (Wu, Adams, and Wilson, 2007) of students' energy understanding were computed using AcerConquest 2.0. Importantly, a 1*pl* Rasch model was used (e.g., instead of a three-dimensional model for energy understanding in biology, chemistry, and physics) because the scope of this study was the development of the new instrument and we thus regard the new test as one. However, we present a separate research article with a detailed analysis of underlying dimensions of students' energy understanding, as well as changes in these potential dimensions at different grade levels – please see Authors (2016).

This model was conducted both across grade levels and for grade levels 6, 8, and 10 separately. The instrument adequacy in terms of the fit between item difficulty and students' abilities was analysed using a Wright map. Reliability was computed using WLE person separation indices, which are comparable to Cronbach's α values (Wright & Stone, 1979; Wu et al., 2007, p. 25, 160). Differential item functioning (DIF, Wu et al., 2007, Ch.8) was used to determine if the developed items were answered differently by male and female students of similar abilities. DIF effects were analysed in relation to commonly used benchmarks (Wetzel, Böhnke, Carstensen, Ziegler, and Ostendorf, 2013).

2. Results

2.1 Wright map

Figure 1 shows a Wright map with the distribution of item difficulties (right) and students' abilities (left) in the reference group grade 10. The two parameters (student abilities and items difficulties) are presented on a logit scale with high-performing students, respective difficult items at the top, and low-performing students, respective easy items at the bottom. Ideally, item difficulties are spread widely to cover all abilities of the test takers. The distribution of item difficulties should furthermore reflect the distribution of students' abilities. Figure 1 shows a sufficiently wide distribution of item difficulties and a reasonable fit between the distributions of item difficulties and student abilities. Thus, the developed instrument fitted students' performance in the reference group well. As the instrument was

initially designed to assess energy understanding until high school (grade 12), the resulting instrument was slightly more difficult than grade 10 students' abilities.

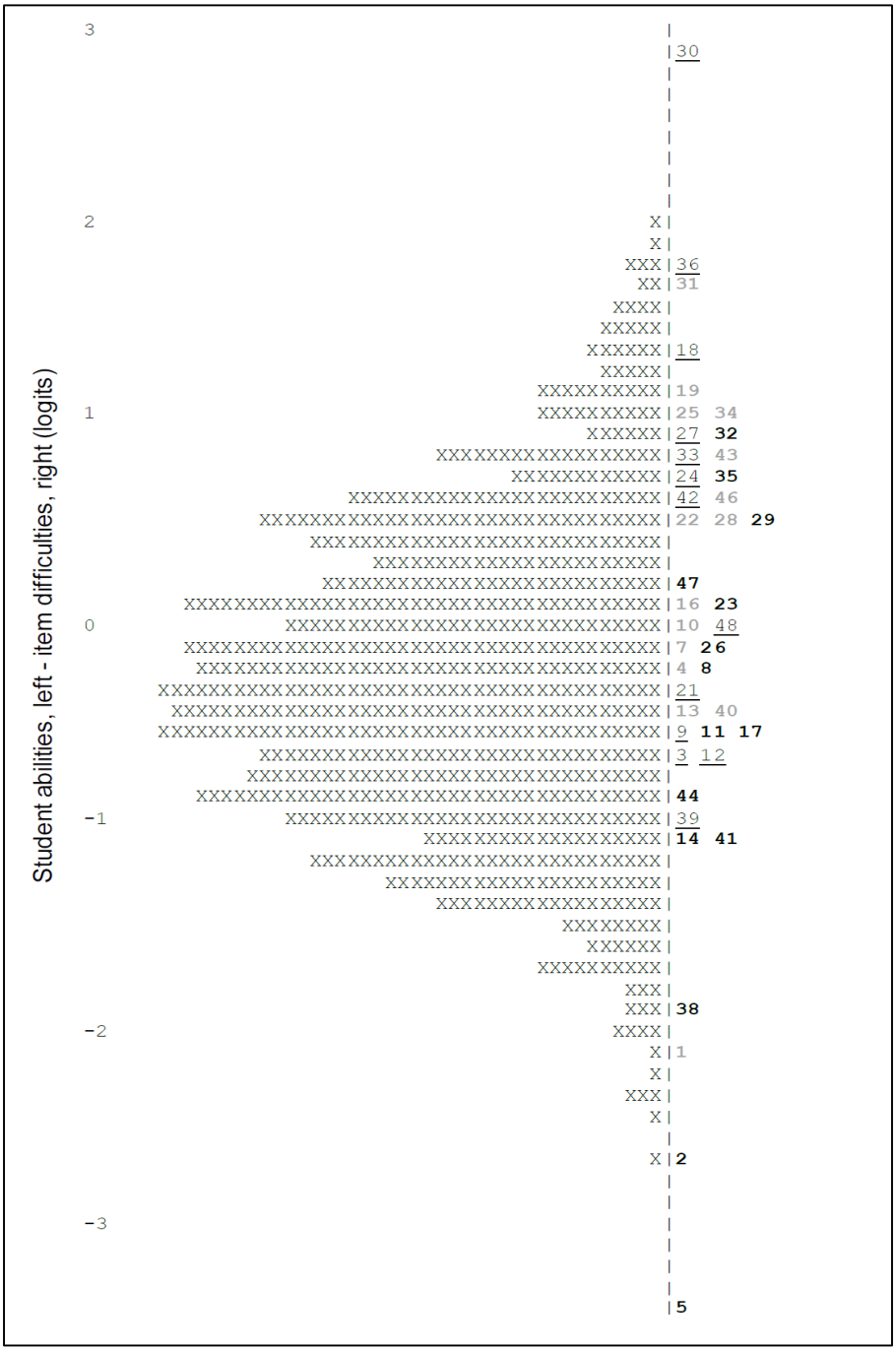

Figure 1. Wright map of grade 10 student abilities (left) and item difficulties (right). Item numbers refer to numbers in the test booklet (see online material 1). Formats of item numbers indicate disciplines: shaded – biology, underscored – chemistry; **bold – physics**.

2.2 WLE person separation reliabilities

Across grades 6–10, WLE person separation reliability was high (.82) for the full test (grade 6: .57, grade 8: .76, grade 10: .83). The respective reliabilities of the disciplinary subtests were sufficient to satisfactory (.58–.62).

2.3 DIF analysis

The overall DIF effect of gender was significant and effected 15 items ($\chi^2 = 13$, $df = 1$, $p < .001$). However, only three items showed 'slight to moderate' DIF effects ($m = .32$ – range of DIF estimators = .25 – .38), while the remaining 12 effects proved 'negligible' ($m = .18$; estimators < .20, see benchmarks by Wetzel et al., 2013). Thus, DIF effects for gender can be considered unproblematic for the items of the developed instrument.

3. References

Authors, 2016

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do Individual Response Styles Matter? *Journal of Individual Differences*, 34(2), 69-81. doi: 10.1027/1614-0001/a000102

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press/University of Chicago.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0: generalized item response modelling software*. Camberwell, VIC: Acer Press.