**Online Supplement for paper "Changing between representations of elementary functions: students' competencies and differences with a specific perspective on school track and gender"**

## Dimensionality of the competency construct

To verify the fit of the four-dimensional Rasch model underlying the data analysis in our article, we compared it with two other theoretically plausible Rasch models. All models were specified with between-item multidimensionality and compared using $\chi^2$-difference tests as they provide a nested structure.

- First, the data were modelled using a one-dimensional model that assigned all items to the same competency dimension. The deviance amounted to 20,563.5 for this one-dimensional model.
- Second, we specified a two-dimensional model distinguishing between tasks with a situational context and purely mathematical tasks. The deviance of this two-dimensional model was 19,894.9, indicating a significantly better model fit ($\chi^2(2) = 668.6$, $p < .001$).
- Third, this two-dimensional model was compared to the four-dimensional model which included the representational changes that were focused on by the items. The deviance of the four-dimensional model amounted to 19,601.1, again showing a highly significant advantage for the higher-ordered model ($\chi^2(7) = 293.8$, $p < .001$).

The finding that the model distinguishing between tasks with and without a situational context exhibited a significantly better fit than the one-dimensional model is in line with the results of Nitsch (2015). Moreover, the modelling cycle (Blum & Leiß, 2005) indicates that other competencies than purely mathematical ones might also be necessary when working on tasks with a situational context, for instance, verbal competencies or critical thinking. Furthermore, the advantage of the four-dimensional model is that the distinction between tasks with and without a situational context can be refined when taking into account the specific representations involved in the underlying representational changes. Altogether, these findings empirically confirm the assumption that tasks with and without a situational context as well as tasks involving different representational changes require distinct competency facets. In particular, they support the theoretically-driven decision to implement a four-dimensional modelling.

The latent correlations of the four dimensions ranged between 0.46 and 0.79 (see Table 1). These moderate correlations imply divergent validity; i.e., despite the fact that the tasks are related by the domain of functions, different competency facets are needed to perform these representational changes.

**Table 1** Latent correlations between the four dimensions

| | Graph & situation | Equation & graph | Situation & equation |
|---|---|---|---|
| Equation & graph | .46 | | |
| Situation & equation | .65 | .59 | |
| Knowledge parameters | .53 | .79 | .51 |

Scale reliability was documented using EAP/PV reliability (Bock & Aitkin, 1981), which can be interpreted as comparable to Cronbach's Alpha (Rost, 2004) and should be > .7 for each dimension (Rauch & Hartig, 2007). In the present sample, Dimension 1 (graph & situation) showed a reliability of *EAP/PV* = 0.73, Dimension 2 (equation & graph) of *EAP/PV* = 0.83, Dimension 3 (situation & equation) of *EAP/PV* = 0.72, and Dimension 4 (knowledge about the parameters) of *EAP/PV* = 0.69. Although Dimensions 3 and 4 only contained three items, their reliability proved to be satisfactory (Rauch & Hartig, 2007), indicating that our test reliably measured the intended competency as a four-dimensional construct. This finding confirms prior research by Bayrhuber-Habeck (2010) and – in a wider perspective – also research by Vollrath (1989) that characterized functional thinking as a multi-dimensional construct.

The Wright map in Figure 1 demonstrates that students' competencies (marked as X) were rather balanced with regard to the four dimensions, showing the highest variance for Dimension 3 and the lowest variance for Dimension 1. The lower part of the Wright map indicates that there were several students with poor competency scores. Most items are close to 0, revealing that they were moderately difficult, whereas very difficult or very easy items were rare. In particular, all item difficulties are in the interval $[-3;\ +3]$, as recommended by Baker (2001). However, it could make sense to provide further items of very high and low difficulty in order to reliably estimate the competency of the best-performers and the weakest learners. Nevertheless, items of medium difficulty should not be deleted as they provide the most diagnostic information (Frey, 2012).

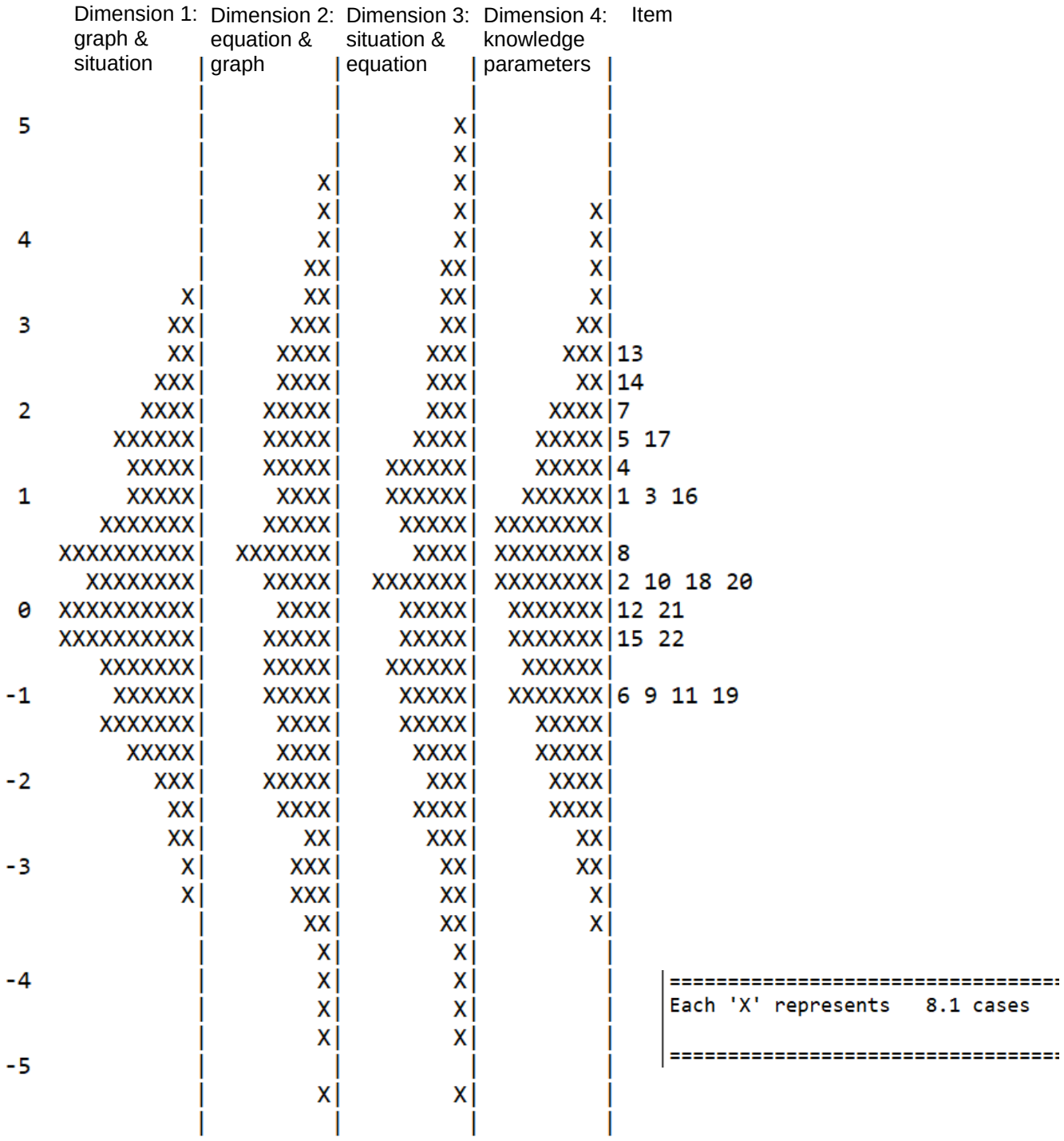


**Figure 1** Wright map of the 4-dimensional competency model

### Item-related test quality

In the following, we will report different quality indices in order to document the item-related test quality. The traditional discrimination as point biserial correlation should be > 0.3 for every item (Lienert & Raatz, 1998). In our sample, item discrimination can be considered satisfactory with values between .33 and .67.

Moreover, the weighted mean squares (MNSQ) indicate the extent to which items deviate from the estimated model; they have an expected value of 1.0 and are displayed with a confidence interval and corresponding t-values by Conquest. According to Bond and Fox (2015), MNSQ values between 0.7 and 1.3 can be considered satisfactory. The confidence interval and t-value enable the evaluation of the probability of a possible misfit. In particular, MNSQ below the confidence interval (t-value < -1.96) indicate a so-called overfit; i.e., the item fits the model better than expected. Whereas overfit is not generally considered problematic, items showing a so-called underfit, i.e., MNSQ above the confidence interval (t-value > 1.96), should be evaluated with regard to their possibly low discriminatory power. In our sample, a possible misfit was investigated based on weighted mean squares (MNSQ), revealing values within the recommended interval of 0.7 and 1.3. However, three items (see Table 2) deviated significantly from the estimated model in the sense of an overfit (t-value < −1.96), and three other items exhibited a significant underfit (t-value > 1.96).

**Table 1** Items with t-values > |1.96|

| **Items with overfit** | | | **Items with underfit** | | |
|---|---|---|---|---|---|
| **(#)** Description | MNSQ | T | **(#)** Description | MNSQ | T |
| **(19)** Graph →situation (SC) | 0.91 | −2.2 | **(14)** Graph →equation (Slope as fraction; open) | 1.12 | 2.1 |
| **(15)** Equation →graph (SC) | 0.89 | −2.3 | **(22)** Graph →situation (SC) | 1.12 | 3.2 |
| **(21)** Equation →graph (SC) | 0.85 | −3.2 | **(8)** Equation →graph (Slope as fraction; open) | 1.2 | 3.8 |

*Note:* Open and SC (Single-Choice) refer to the item format

As t-values get significant even for a little misfit in large samples, Bond and Fox (2015) recommend to evaluate the discriminatory power of items not only based on MNSQ and t-values but to assess the discriminatory power of items with underfit based on their item characteristic curves (ICC). Taking a look at the ICC of the items with underfit (see Figure 2) reveals that their discriminatory power appears to be partly reduced in the upper and lower periphery of the latent trait. These items are easier for weak learners and more difficult for strong learners than expected. Comparing these ICCs with the ICC for Item 3, which showed a perfect fit (MNSQ = 1.0; t = 0.1), confirms that the ICCs of the items with underfit are relatively inconspicuous.

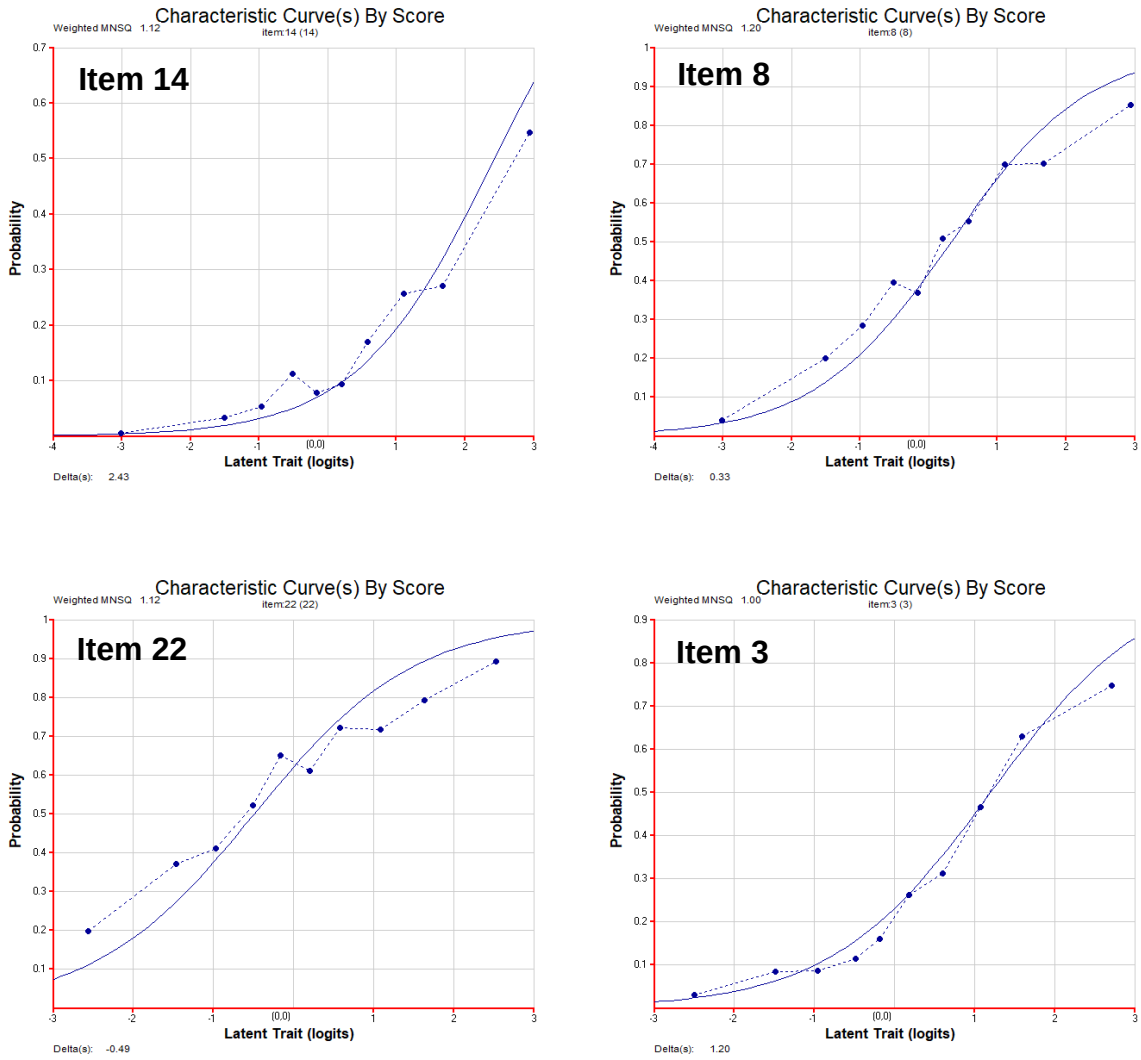


**Figure 2** The ICCS of Items 8, 14, and 22 showing an underfit and of Item 3 showing a perfect fit

The investigation of the test quality indices revealed a satisfactory item-related test quality with regard to traditional item discrimination and possible misfits. Although the t-values of three items showed a significant underfit, an inspection of the corresponding ICCs indicated that there is no need to delete these items when using the instrument in the future (cf. Bond & Fox, 2015). Altogether, the reported global and item-related fit indices show that the four-dimensional modelling fits the data to a satisfactory extent. Therefore, the instrument can be used in future research.

## References

Baker, F. (2001). *The Basics of Item Response Theory*. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bayrhuber-Habeck, M. (2010). *Konstruktion und Evaluation eines Kompetenzstrukturmodells im Bereich mathematischer Repräsentationen.* [Construction and evaluation of a competency model in the field of mathematical representations.] Retrieved from https://phfr.bsz-bw.de/frontdoor/index/index/docId/355 [9.8.2021].

Blum, W., & Leiss, D. (2005). Modellieren im Unterricht mit der „Tanken"-Aufgabe. [Modelling in mathematics classroom with the "refueling" task.] *mathematik lehren, 128,* 18–21.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika 46*, 179–197.

Bond, T.G., & Fox, C.M. (2015). *Applying the Rasch model. Fundamental measurement in the human sciences.* New York: Routledge.

Frey, A. (2012). Adaptives Testen. [Adaptive testing.] In H. Moosbrugger, & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 7–26). Berlin: Springer.

Lichti, M., & Roth, J. (2019). Functional thinking – a three-dimensional construct? *Journal für Mathematik-Didaktik 40*(2), 169–195.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. [Test construction and analysis.] Weinheim: Beltz.

Nitsch, R. (2015). *Diagnose von Lernschwierigkeiten im Bereich funktionaler Zusammenhänge.* [Diagnosing learning difficulties in the field of functional relationships.] Wiesbaden: Springer Spektrum.

Rauch, D., & Hartig, J. (2007). Interpretation von Testwerten in der IRT. [Interpretation of test values in the IRT framework.] In H. Moosbrugger, & A. Kelava (Eds.), *Test- und Fragebogenkonstruktion* (pp. 240–250). Berlin: Springer.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. [Textbook test theory – test construction.] Bern: Huber.

Vollrath, H.-J. (1989). Funktionales Denken. [Functional thinking.] *Journal für Mathematikdidaktik 10,* 3–37.