Additional Information on the Item Selection Process

Item-fit statistics alone (i.e., statistics that examine the fit of individual items of a test), which are often used for item selection, are not well suited for this purpose (Edelsbrunner & Dablander, 2015; Karabatsos, 2000; Smith & Plackner, 2009). They might not detect violations of the assumptions of one-dimensionality (Smith, Schumacker, & Bush, 1998) and subgroup homogeneity (Smith & Suh, 2003). Moreover, evaluating single items’ deviations from the Rasch model is considered unreasonable without conducting a global test of whether the Rasch model holds (Christensen & Kreiner, 2013). To guarantee valid item selection, we therefore applied different test statistics, starting with global and continuing with local (i.e., item-specific) analyses.

We also fitted Rasch mixture models (Rost & von Davier, 1995), which search for latent classes within a sample indicated by maximally different item parameter estimations. These latent classes represent students who answer items systematically differently, in other words, for whom the test does not measure the same underlying construct. When models with more than one class were shown to better fit the data than the regular (one-class) Rasch model did, we inspected the item difficulties estimated for the latent classes to determine which items particularly differed between the detected latent classes. We used the package Psychomix (Frick, Strobl, Leisch, & Zeileis, 2012) to test Rasch mixture models.


