Supplemental Appendix 1: ***Negative item method effects***

Method effects are non-trait effects associated with idiosyncratic aspects of particular items or methods of data collection. Failure to incorporate method effects appropriately is likely to have substantial effects on goodness of fit and biased parameter estimates. The use of negatively worded items constitutes a potential source of construct irrelevant variance that detracts from the construct validity of interpretations of self-concept responses, particularly in young students. However, even for responses by older students and adults, factor analyses of psychological rating scales, comprising a mixture of positively and negatively worded items, typically reveal apparently distinct factors reflecting the positive and the negative items respectively (see Carmines & Zeller, 1986; Chiu, 2008, 2011; Marsh, 1986; 1996). A typical approach based on the logic of multitrait-method studies of method effects is to test for negative-item method effects by including correlated uniquenesses between negatively worded items (e.g., Benson & Hocevar, 1985; Carmines & Zeller, 1986; DiStefano & Motl, 2006; Marsh, 1986; 1996; Marsh, Scalas & Nagengast, 2010). Consistent with these expectations, Chiu (2008, 2011) reported this type of negative item method effect for the two negatively worded self-concept items for the TIMSS2003 data. She included correlated uniquenesses to control for this negative-item bias, but further noted that the factor loadings for these items were systematically lower than for the positively worded items and suggested that "items that are negatively worded appear to be unreliable in cross-cultural studies" (p. 251). Marsh, Abduljabbar et al (2013) reported similar results. This might also explain in part the extreme differences in reliability estimates between the US and Saudi responses, particularly on the math self-concept scale (2 of 4 items are negatively worded) and, to a lesser extent, the affect scale (1 of 3 items is negatively worded). In the present investigation, following recommendations by Marsh, Abduljabbar et al (2013), we compare models with and without correlated uniquenesses to test for negative item effects, and correct for them if they are shown to exist (see Appendix 2, Supplemental Materials).