Appendix 2: Analysis of Differences in Item Type

Only multiple-choice (MC) items were employed in this study. The item design aimed to vary item difficulties by operationalising a dimension of cognitive complexity in the employed model for item development (Neumann et al., 2013). This process showed that the number of factors that need consideration varies substantially across contexts and disciplines. This variance was taken into account through three types of multiple choice items (MC items), thereby allowing a limitation of the amount of reading load per answering option in more complex items. Besides regular MC items (one correct solution), some items were formatted as two-attractor MC or complex MC items. To minimise confusion over differing items types, students were provided with a help-box that identified the correct number of answering options per item. All students were instructed about the different item types and the help-box at the beginning of testing sessions. To compare the three item types, Table 1 presents the share of students choosing an incorrect answering pattern (fewer or more options than indicated) in the three employed item types.

Table 1. Share of incorrect answering patterns (fewer or more than the correct number of options), as well as the guessing probability of three employed item types (%)

	Grade 6	**Grade 8**	**Grade 10**	**Guessing probability (%)**
Regular MC (1 attractor, 4–5 options) 32 /48 items	.13	.13	.13	25
2-attractor MC (2 attractors, 4–5 options) 10/48 items	9.3	8.8	5.4	5
Complex MC (4–6 attractors, 9–16 options) 6/48 items	11.7	13.3	4.7	≤ 1

From these data, two conclusions can be drawn:

(1) From Grade 6 to Grade 10, the share of students using incorrect answering patterns is relatively constant in all item types, thereby introducing only little variance *across grades*, which is the important dimension for the here presented study.

(2) Even though the share of students who incorrectly answered items is acceptably low, a comparison of the three item types indicates that the 2-attractor and complex MC items were more often answered incorrectly that the regular MC items. However, as the guessing probability was much

higher among the regular MC items than among the two other item types, the different item types move closer together in producing construct-irrelevant variance.

In a related analysis of occurring cognitive processes during item answering (think-aloud protocols – see discussion section), consistent differences in item types were not observed. Similar to these findings, the effect of the different item types was found to be small in an earlier study (Opitz et al., 2014). Consequently, it is concluded that the differences in answering format are small enough for the items to be applicable for the scope of this study.