Supplementary Material 1

Method Study 1

Sample. The gender differences found at a Gymnasium are not representative of gender differences in the general student population in Germany. About 38% of German students were attending this kind of school at the time of our investigation in Baden-Württemberg, and the students participating in the present study could be considered to be representative of this kind of population (mostly white students without immigration backgrounds from families with higher socioeconomic status). Students were selected for the Gymnasium on the basis of their grades and, strongly related to grades, on their teacher-rated achievement behavior. At the end of elementary school, in the fourth grade, the teacher gave a mandatory recommendation either for the Gymnasium (academic track) or for another kind of school (different kinds of vocational tracks). As girls tend to have better grades than boys in elementary school, more girls than boys get a recommendation for the Gymnasium and attend a Gymnasium. Thus, in the academic track of the German school system, there are more females (about 51%) than males (about 49%), similar to most universities in Western countries (see Voyer & Voyer, 2014). This general gender distribution did not significantly differ from the gender distribution in our sample ($X^2 = 0.147$; $p = .701$). At the time when we conducted the study, students attending a Gymnasium in Baden-Württemberg were not allowed to choose different levels in math classes or other main subjects (e.g., German) until graduating from this kind of school. Thus, girls and boys had been taught the same curriculum and had spent the same amount of time in math classes.

Intelligence measure. The basic module of the IST 2000 R includes nine reasoning tasks. The reasoning tasks are subdivided into three verbal, three numerical, and three figural reasoning tasks intended to measure verbal, numerical, and figural reasoning (Liepmann et al., 2007). Examples of the tasks used in the IST-2000 R can be found in Schulze et al. (2005). Each verbal, numerical, and figural reasoning task consists of 20 items. In the first verbal task,

"Sentence Completion" (SC), participants have to choose one out of five words in order to complete a sentence correctly. In the second verbal task, "Verbal Analogies" (VA), participants have to choose one out of five words in order to complete a verbal analogy correctly. In the "Verbal Similarities" (VS) task, the participants are requested to choose the two words that are most similar out of six. The first numerical task, "Calculations" (CA), consists of simple arithmetic operations, and the solution has to be written down. In the second numerical task, "Number Series" (NS), participants have to write down the next number corresponding to a rule, which has to be identified. In the "Signs" task (SI), participants have to insert arithmetic signs into equations in order to complete them correctly. "Abstract Pieces" (AP) is the first figural reasoning task. It requires participants to choose one out of five geometric figures that can be composed of smaller pieces. The task called "Cubes" (CU) requires the participants to choose the one cube that represents a rotated target cue out of five. The third figural reasoning task, "Matrices" (MA), corresponds to abstract figural matrices as they are often used to assess fluid intelligence. Verbal, numerical, and figural intelligence scores were computed as sums of the corresponding reasoning tasks. Maximum task score in all three domains was 60. The construct validity of the test has been demonstrated extensively (e.g., Liepmann et al., 2007; Steinmayr et al., 2010). Measurement properties in our study were satisfactory as well. In the group that we tested according to the manual (Condition 1), the internal consistencies were the following: verbal reasoning $\alpha = .77$; numerical reasoning $\alpha = .90$; figural reasoning $\alpha = .80$. In the group that was tested with deviant testing times (Condition 2), the coefficients were the following: verbal reasoning $\alpha = .78$; numerical reasoning $\alpha = .88$; figural reasoning $\alpha = .79$.