1. Instruments

1.1. Cognitive variables

Physics problem conceptualization, as described above, is measured on an ordinal scale (see above). The scale contains 4 items (physics problems). The interrater reliabilities are $\kappa_{\text{pre}} = 0.93$ and $\kappa_{\text{post}} = 0.87$, indicating that the coding rubric was intersubjectively understandeable. The internal reliabilities for the scored responses are $\alpha_{\text{pre}} = 0.87$ and $\alpha_{\text{post}} = 0.82$. This indicates that the scale is in and of itself consistent (i.e., measures a similar construct) and it is appropriate to sum the points for the participants. A summary of reliability and difficulty for this scale can be found in Tabel 1.

In order to raise external validity for the physics problem conceptualization scale, another established scale for generic physics problem solving abilities was utilized (Heller et al., 2007). Generic physics problem solving is measured with a binary scale (0 incorrect, and 1 correct). The scale contains 15 items (e.g., "Which wheel has the most spins per minute?"). Few missing values were imputed using random regression imputation (see below).

Quantitative, non-verbal reasoning skills were measured with several subscales of well-established instruments (Heller et al., 2007). We used three subscales (Q1, Q2, and N1). In particular, N1 involved mental rotation and abstract reasoning, while Q1 and Q2 capitalized on students' understanding of equations and numerical quantities (e.g., what is bigger: $1/2$ or $1/4 + 1/4$). The predictive validity (correlations with school grades one year later) of the N and Q scales are $.20, .07$, and $.45, .35$ with science and math, respectively. A subset of items was used such that students from all grades (10,11,12) solved the same items. Missing values were not an issue for all scales but cognitive abilities. For cognitive abilities 48 percent are missing. Consequently, we used random regression imputation (Gelman & Hill, 2007) in order to recover the missing values with expectancy of success, physics problem conceptualization abilities, and heuristic physics problem solving abilities as predictor variables. As calculated via t-test, the imputed scale did not significantly differ from the original scale, $M(\text{orig.}) = 30.18$, $sd = 8.79$; $M(\text{imp.}) = 30.48$, $sd = 9.07$; $t(149.87) = -0.24$, $p = .812$, $r = 0.02$.

Finally, achievement in round 1 is included as a cognitive variable. This is measured through 4 homework physics problems that are graded by the teachers.

1.2. Motivational and success variables

Expectancy physics competition, as a prime motivator for academic behavior, with 5 items (e.g., "I think it will be difficult to learn what is required in the physics Olympiad."). The scale shows satisfatory reliability, discriminatory power, mean and standard deviation – though the scale shows a ceiling effect which can be seen in a Shapiro-test for normality ($p < .05$). Values towards physics competition are measured with 4 items (e.g., "I is personally important for me to be good in the physics Olympiad."). While reliability and discriminatory power are satisfactory, the variable is non-normal (Shapiro-test with $p < .05$). Sense of belonging is measured with 15 items (e.g., "In the physics community I feel accepted."). The reliability and discriminatory power are good, while, similarly, the scale is non-normal as tested through Shapiro-test ($p < .05$). Finally, social support is measured with 12 items with satisfatory item characteristics.

In order to measure how successful students were, their highest qualified round (of

Table 1. Overview of utilized variables.

Scale	N^{a}	α	$\bar{r}_{i(t-i)}$	Range	M	SD
Cognitive variables:						
Problem comprehension	141	0.87	0.52	-1-2	7.35	6.8
Heuristic physics problem solving	130	0.69	0.3	0-1	0.54	0.2
Cog. abilities	73	0.89	0.37	0-47	30.18	8.79
Achievement round 1	139			0-45	26.92	10.39
Motivational variables:						
Self-efficacy physics	141	0.85	0.69	1-4^{b}	3.33	0.57
Expectancy physics competition	141	0.71	0.47	1-4	2.6	0.49
Value physics competition	141	0.77	0.57	1-4	2.85	0.64
Sense of belonging	141	0.81	0.42	1-4	3.32	0.34
Social support in physics	141	0.77	0.41	1-4	2.55	0.5
Success variables:						
Highest round	141			1-3	1.61	0.63

[a] N ... number of responses with at least one valid value, α ... internal consistency, and $\bar{r}_{i(t-i)}$ ... mean discriminatory power.

[b] All scales measured 1 to 4 have to response text "not true at all" to "true" (or closely similar).

overall 4) was tracked. This gives a proxy for success and can be regarded as an ordinal scale. Further details for the measures can be found in Tabel 1.

2. Sample

The sample was drawn from participants in the Physics Olympiad. All students that participated in the Physics Olympiad were contacted and a fraction of these students took part in the project. Overall the sample considered for this analysis comprised 141 students (female = 42, male = 99). The mean age was 16.21(1.16) years. These are 50 percent of the respective overall Olympiad population (this smaller number is due to the fact that not all students participated in the second round of the questionnaire where the cognitive measures were assessed).

In order to detect possible selection biases with regards to the participants in the questionnaire in round 1, the students who participated in round 2 in the questionnaire (not the Olympiad) will be compared with the students who participated in round 1 with regards to background variables that were administered for both groups. The groups differed with respect to achievement in round 1, M(all) = 25.19, sd = 10.92; M(sub) = 26.92, sd = 10.39; $t(290.25)$ = -1.57, p = .118, r = 0.09. The groups did not differ in terms of highest qualification, M(all) = 1.57, sd = 0.71; M(sub) = 1.61, sd = 0.63; $t(311.28)$ = -0.63, p = .531, r = 0.04, and self-efficacy in physics, M(all) = 3.3, sd = 0.39; M(sub) = 3.32, sd = 0.34; $t(319.86)$ = -0.47, p = .638, r = 0.03 – so that we are confident that it is no artificial sample that constituted itself. When compared with the overall Olympiad population, the sample of students in round 2

differed from the overall population with regards to achievement, $M(\text{sample}) = 26.92$, $sd = 10.39$; $M(\text{all}) = 25.45$, $sd = 9.05$; $t(170.76) = 1.58$, $p = .117$, $r = 0.12$, and with regards to age, $M(\text{sample}) = 16.21$, $sd = 1.16$; $M(\text{all}) = 17.1$, $sd = 1.04$; $t(175.4) = -8.61$, $p = 0$, $r = 0.54$. The students in round 2 were younger with a large effect and higher in achievement in the Olympiad with a small effect. These effects have to be kept in mind when considering implications from this study.