Supplementary Material 4

Results Study 2

2 x 2 x 2 univariate analyses.

Omitted items. Both gender ($F_{1, 538} = 63.44$, $p < .001$; $\eta^2 = .105$) (girls omitted more items on both tests combined) and test order ($F_{1, 538} = 8.93$, $p = .003$; $\eta^2 = .016$) (fewer items were omitted when the power test was presented first on both tests combined) had a significant main effect on the dependent variable, but they did not interact with each other ($F_{1, 538} = 0.001$, $p = .978$). The multivariate analyses further demonstrated that fewer items were omitted under power conditions compared with speeded conditions (condition: $F_{1, 538} = 624.13$, $p < .001$; $\eta^2 = .537$). This decrease was stronger for girls than for boys (Condition x Gender: $F_{1, 538} = 40.22$, $p < .001$; $\eta^2 = .070$). Furthermore, the decrease in omitted items from the speeded to the power test was stronger when the speeded version was presented first (Condition x Test Order: $F_{1, 538} = 168.86$, $p < .001$; $\eta^2 = .239$). The decrease in the number of omitted items when the speeded version was presented first was higher for girls than for boys (Condition x Gender x Test Order: $F_{1, 538} = 7.07$, $p = .008$; $\eta^2 = .013$).

Accuracy index. The 2 x 2 x 2 univariate repeated-measures ANOVA using the accuracy index as the dependent variable yielded the following results. Both gender ($F_{1, 538} = 7.82$, $p = .005$; $\eta^2 = .014$) (girls worked less accurately on both tests combined) and test order ($F_{1, 538} = 15.72$, $p < .001$; $\eta^2 = .028$) (accuracy was higher on both tests combined when the power test was presented first) had a significant main effect on the dependent variable, but they did not interact with each other ($F_{1, 538} = 2.25$, $p = .134$). The multivariate analyses further demonstrated that accuracy was higher under power conditions compared with speeded conditions (condition: $F_{1, 538} = 7.33$, $p = .007$; $\eta^2 = .013$). This increase was the same for both genders (Condition x Gender: $F_{1, 538} = 0.53$, $p = .467$) and was not influenced by test order (Condition x Test Order: $F_{1, 538} = 0.07$, $p = .793$). Furthermore, the interaction between

the three variables was not significant either (Condition x Gender x Test Order: $F_{1, 538}$ = 0.003, p = .955).

2 x 2 x 2 univariate analyses with covariates. The scores on the two intelligence tests, number of omitted items, and accuracy index, respectively, presented in the speed and power conditions constituted the within-subjects variable (condition). Gender and test order served as independent variables. Performance in numerical reasoning, number of omitted items, and the accuracy index, respectively, served as the dependent variables. The Positive/Approach Attitude and Negative Emotionality/Avoidance variables served as covariates.

Test score. Both gender ($F_{1, 534}$ = 12.77, p < .001; η^2 = .024) and test order ($F_{1, 534}$ = 20.52, p < .001; η^2 = .038) still had a significant main effect on the dependent variable and still did not interact with each other ($F_{1, 534}$ < 0.001, p = .993). The two covariates also had a significant main effect on the averaged test performance (Positive/Approach Attitude: $F_{1, 534}$ = 73.00, p < .001, η^2 = .122; Negative Emotionality/Avoidance: $F_{1, 534}$ = 24.00, p < .001, η^2 = .044).

Students still performed better in the power than in the speed condition (Condition: $F_{1, 534}$ = 579.26, p < .001; η^2 = .520). However, there was a significant interaction between the variables Positive/Approach Attitude and condition ($F_{1, 534}$ = 25.65, p < .001; η^2 = .046). A comparable interaction was not found for Negative Emotionality/Avoidance and test condition ($F_{1, 534}$ = 0.22, p = .639). Furthermore, the analysis with the covariates revealed that the increase in test performance was still stronger for girls than for boys (Condition x Gender: $F_{1, 534}$ = 15.97, p < .001; η^2 = .029). Still, the increase in performance from the speeded to the power test was stronger when the speeded version was presented first after controlling for the different covariates (Condition x Test Order: $F_{1, 534}$ = 114.73, p < .001, η^2 = .177). This effect was barely affected by the considered covariates. Last but not least, the marginally significant three-way interaction (Condition x Gender x Test Order: $F_{1, 534}$ = 4.29, p = .039, η^2 = .008)

was also not affected by additionally considering the emotional and motivational covariates. Still, the effect that girls benefitted more from the change in test condition than boys was stronger when the speeded test was presented first, thus indicating an additional stronger training effect for girls than for boys.

Number of omitted items. Both gender ($F_{1, 534} = 23.55$, $p < .001$; $\eta^2 = .042$) and test order ($F_{1, 534} = 9.05$, $p = .003$; $\eta^2 = .017$) still had a significant main effect on the total number of omitted items but did not interact ($F_{1, 534} = 0.003$, $p = .955$). The two covariates also had a significant main effect on the total number of omitted items (Positive/Approach Attitude: $F_{1, 534} = 57.56$, $p < .001$, $\eta^2 = .097$; Negative Emotionality/Avoidance: $F_{1, 534} = 12.29$, $p < .001$, $\eta^2 = .022$). All other effects were significant as well (Condition: $F_{1, 534} = 699.42$, $p < .001$, $\eta^2 = .567$; Condition x Positive/Approach Attitude: $F_{1, 534} = 37.85$, $p < .001$, $\eta^2 = .066$; Condition x Negative Emotionality/Avoidance: $F_{1, 534} = 4.26$, $p = .039$, $\eta^2 = .008$; Condition x Gender: $F_{1, 534} = 14.45$, $p < .001$, $\eta^2 = .026$; Condition x Test Order: $F_{1, 534} = 179.71$, $p < .001$, $\eta^2 = .252$; Condition x Test Order x Gender: $F_{1, 534} = 7.78$, $p = .005$, $\eta^2 = .014$).

Accuracy index. Only test order ($F_{1, 534} = 17.79$, $p < .001$; $\eta^2 = .032$) but not gender ($F_{1, 534} = 0.005$, $p = .942$) still had a significant main effect on the dependent variable accuracy index, and the variables also did not interact ($F_{1, 534} = 2.04$, $p = .154$). The two covariates also had a significant main effect on the number of omitted items (Positive/Approach Attitude: $F_{1, 534} = 25.78$, $p < .001$, $\eta^2 = .046$; Negative Emotionality/Avoidance: $F_{1, 534} = 17.92$, $p < .001$, $\eta^2 = .032$). Both test condition ($F_{1, 534} = 7.09$, $p = .008$, $\eta^2 = .013$) and the interaction between test condition and Negative Emotionality/Avoidance ($F_{1, 534} = 6.98$, $p = .008$, $\eta^2 = .013$) yielded a significant effect on the accuracy index. All other effects were not significant (Condition x Positive/Approach Attitude: $F_{1, 534} = 0.34$, $p = .562$; Condition x Gender: $F_{1, 534} = 1.70$, $p = .192$; Condition x Test Order: $F_{1, 534} = 0.002$, $p = .969$; Condition x Test Order x Gender: $F_{1, 534} = 0.004$, $p = .953$).

Factor analysis. The two factors explained about 66% of the variance in the analyzed variables. The following variables loaded on the first rotated factor, which explained 37% of the variance: mathematical ability self-concept (λ = .82), intrinsic motivation (λ = .79), expectation of success concerning the upcoming numerical test (λ = .73), self-estimated numerical intelligence (λ = .73), and hope for success (λ = .71). On the second factor, which explained 29% of the variance, the loadings were the following: mathematical test anxiety measured as worry (λ = .86) and emotionality (λ = .84) plus fear of failure (λ = .70). On the basis of the results of the factor analysis, we calculated two new variables.

Analysis for different ability groups. For the low-ability group, a large effect was found for the interaction between condition and test order ($F_{1, 182} = 21.19$, $p < .001$; $\eta^2 = .10$). A small effect was found for the interaction between gender and condition ($F_{1, 182} = 3.79$, $p = .053$; $\eta^2 = .01$). This effect was marginally qualified by test order and thus by a learning effect (Condition x Gender x Test Order: $F_{1, 182} = 3.55$, $p = .061$; $\eta^2 = .02$). For the medium-ability group, the learning effect was even stronger (Condition x Test Order: $F_{1, 189} = 72.26$, $p < .001$; $\eta^2 = .28$). Furthermore, the strongest interaction effect emerged for this group (Condition x Gender: $F_{1, 189} = 16.57$, $p < .001$; $\eta^2 = .08$) and this was not qualified by a learning effect (Condition x Gender x Test Order: $F_{1, 189} = 1.13$, $p = .289$; $\eta^2 = .006$). The same was true for the high-ability group. The 2 x 2 interaction effects were both also significant (Condition x Test Order: $F_{1, 159} = 21.21$, $p = .002$; $\eta^2 = .12$; Condition x Gender: $F_{1, 159} = 10.24$, $p = .002$; $\eta^2 = .06$) but not the 2 x 2 x 2 interaction ($F_{1, 159} = 0.05$, $p = .83$).

Criterion validity. First, we checked via univariate variance analysis whether math grades differed in the two experimental groups (power condition first vs. speed condition first) or between gender. The two experimental groups did not differ ($F_{1, 540} = .50$, $p = .481$). However, gender had a main effect on math grades ($F_{1, 540} = 8.76$, $p = .003$, $\eta^2 = .016$). Boys had slightly better math grades than girls which has also been found in comparable samples in previous studies (e.g. Steinmayr et al., 2014). In order to check for possible moderation

effects of time constraints or gender on the regression of grades on numerical intelligence we first inspected the correlation between grades and numerical intelligence in the two conditions when they were presented first. Here, we found the following correlations: grades and power test r = .341; grades and speeded test r = .420. Simple slope analysis revealed that the criterion validity between the two conditions did not differ (*B* = .009, *SD Error* = .009, *t* = 0.93, *p* = .35). The same was true for gender in the speeded (*B* = -.002, *SD Error* = .010, *t* = -0.20, *p* = .84) and power condition (*B* =.013, *SD Error* = .018, *t* = 0.80, *p* = .43).