Supplementary Materials

Supplementary Method

Exploratory analyses: Personality measures as additional predictors of the training effects

As in the main hypothesis testing, PP and OP groups were analysed separately, and all variables were converted to within-group z-scores.

To perform these exploratory analyses, we applied an 'all-subsets regression analysis' with 14 predictor variables: baseline performance, fluid intelligence, working memory, perspective taking, fantasy, empathic concern, personal distress, interdependence, narcissism, openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

All-subset regression is an alternative to stepwise regression methods for finding the "best" model. Unlike the stepwise approach, all-subsets regression does not presume that a single "best" model exists. Instead, it provides all possible "equal best" models. All-subsets regression avoids premature termination, which is a limitation of stepwise approaches where some combinations of variables may be missed completely (Brown, 2005; Kuk, 1984). Instead, all-subsets regression uses all possible subsets and combinations of predictor variables and compares the regression models to a chosen statistical criterion, e.g., the Schwarz's Bayesian information criterion (BIC; Schwarz, 1978). BIC is based, in part, on the likelihood function and uses penalised sum of squares criteria. A model with the lowest BIC is the model with an optimal combination of predictor variables that best explain the variance of the outcome variable.

The all-subsets regression analysis was implemented using the R *regsubsets* function in the *leaps* package, which uses a branch-and-bound algorithm (Furnival & Wilson, 1974; Miller, 2002). The predictor variable subset with the minimum BIC was chosen as the one best explaining the variance in the training effect.

In terms of sensitivity to detect these exploratory effects, given the sample size of 92 in each group, the analysis had 80% power to detect predictor effects that are conventionally considered medium to large ($f^2 = 0.23$; Cohen, 1988). The effect size was estimated with a *pwr.f2.test* function in R for a linear regression model with 14 predictor variables and sample size 92.

Supplementary Results

To further investigate what other variables could explain the variance in the physical and observational practice effects, we performed exploratory analyses with 11 additional predictor variables. Consistent with analytical approach taken to address the main hypothesis, PP and OP groups were analysed separately, and all variables were converted to within-group z-scores. Some of the 14 predictor variables were intercorrelated, but not so high as to suggest multicollinearity (Supplementary Figure 1A). In addition, in Supplementary Figure 1B, simple correlations between the training effects and each predictor variable are reported.

Sequence-specific learning

The all-subsets regression analysis, which used BIC for model ranking, returned only one subset of predictors that best explains the variance in the training effect on sequence-specific learning. For the PP group, the winning subset included fluid intelligence, working memory and agreeableness, explaining 18.5% of the variance. The winning model and all three predictor variables were significant (Supplementary Table 1). The result shows that in addition to higher fluid intelligence and lower working memory, higher agreeableness (and not the baseline performance as was reasoned in the primary analysis) is related to better sequence-specific training effects in the PP group.

For the OP group, the winning subset included only one predictor variable: openness to experience. However, the winning predictor did not significantly explain the variance of sequence-specific training effect in the OP group (Supplementary Table 1). All rankings of the two best models for each number of predictors (1-14) for each group are presented Supplementary Figure 2A.

To investigate the possibility that individual differences in agreeableness might reflect gender differences (Schmitt, Realo, Allik, & Voracek, 2008), we repeated the all-subsets analyses including gender as an additional predictor variable. The repeated analyses did not change the results for either PP or OP group, indicating that there were no significant gender differences in the training effects.

General skill learning

For the training effects on general skill learning, none of the personality measures helped further explain the variance in the PP group. The all-subsets regression analysis with BIC for model ranking returned baseline performance and working memory as the best predictors of the physical practice effects (Supplementary Table 2).

In the OP group, baseline performance and agreeableness best explained the variance of the training effect on general skill learning. Both lower baseline performance and lower agreeableness predicted higher general skill learning, however, agreeableness did not reach statistical significance (Supplementary Table 2). All rankings of the two best models for each number of predictors (1-14) for each group are presented Supplementary Figure 2 B. Adding gender as an additional predictor variable did not change the results for either the PP or OP group.

Supplementary References

Brown, P. J. (2005). Variable Selection. In *Encyclopedia of Biostatistics* (pp. 205–232). Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/0470011815.b2a09055

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J. : L. Erlbaum Associates.

Furnival, G. M., & Wilson, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics*, *16*(4), 499–511. doi:10.1080/00401706.1974.10489231

Kuk, A. Y. C. (1984). All subsets regression in a proportional hazards model. *Biometrika*, *71*(3), 587–592. doi:10.1093/biomet/71.3.587

Miller, A. (2002). *Subset Selection in Regression*. Chapman and Hall/CRC.

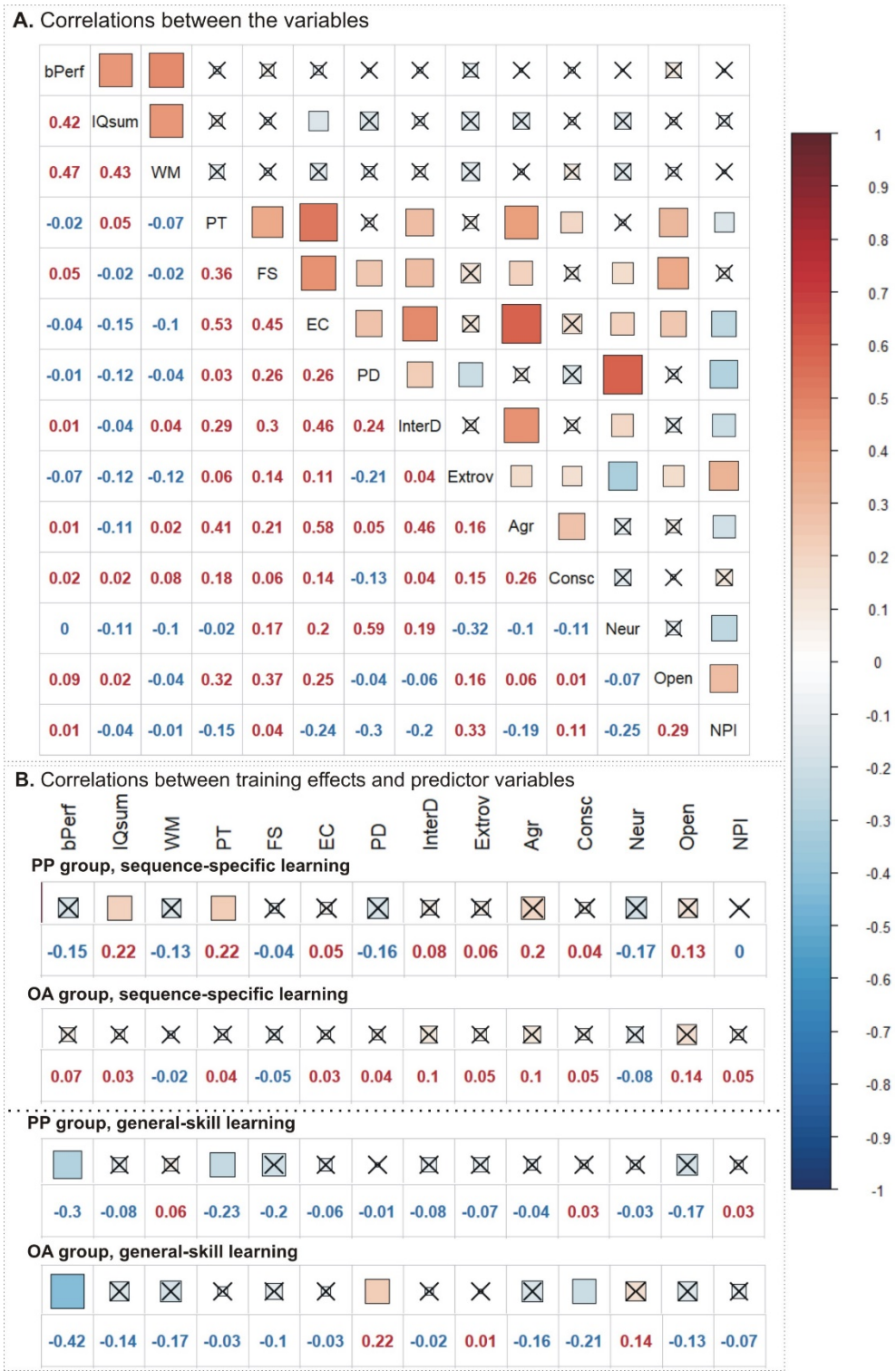
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Supplementary Table 1. Exploratory analyses including personality -- The winning models for sequence-specific learning.

Physical practice				Observational practice			
Model	$F_{3,88} = 6.635$, $p < 0.001$, $R^2 = 0.185$			**Model**	$F_{1,90} = 1.81$, $p = 0.182$, $R^2 = 0.020$		
Coefficients	**β [95% CI]**	**t**	**p**	**Coefficients**	**β [95% CI]**	**t**	**p**
Intercept	0	0	1	**Intercept**	0	0	1
Fluid intelligence	0.395 [0.180, 0.611]	3.641	0.0005	**Openness**	0.140 [-0.067, 0.348]	1.345	0.182
Working memory	-0.329 [-0.545, -0.113]	-3.037	0.003				
Agreeableness	0.286 [0.089, 0.482]	2.888	0.005				

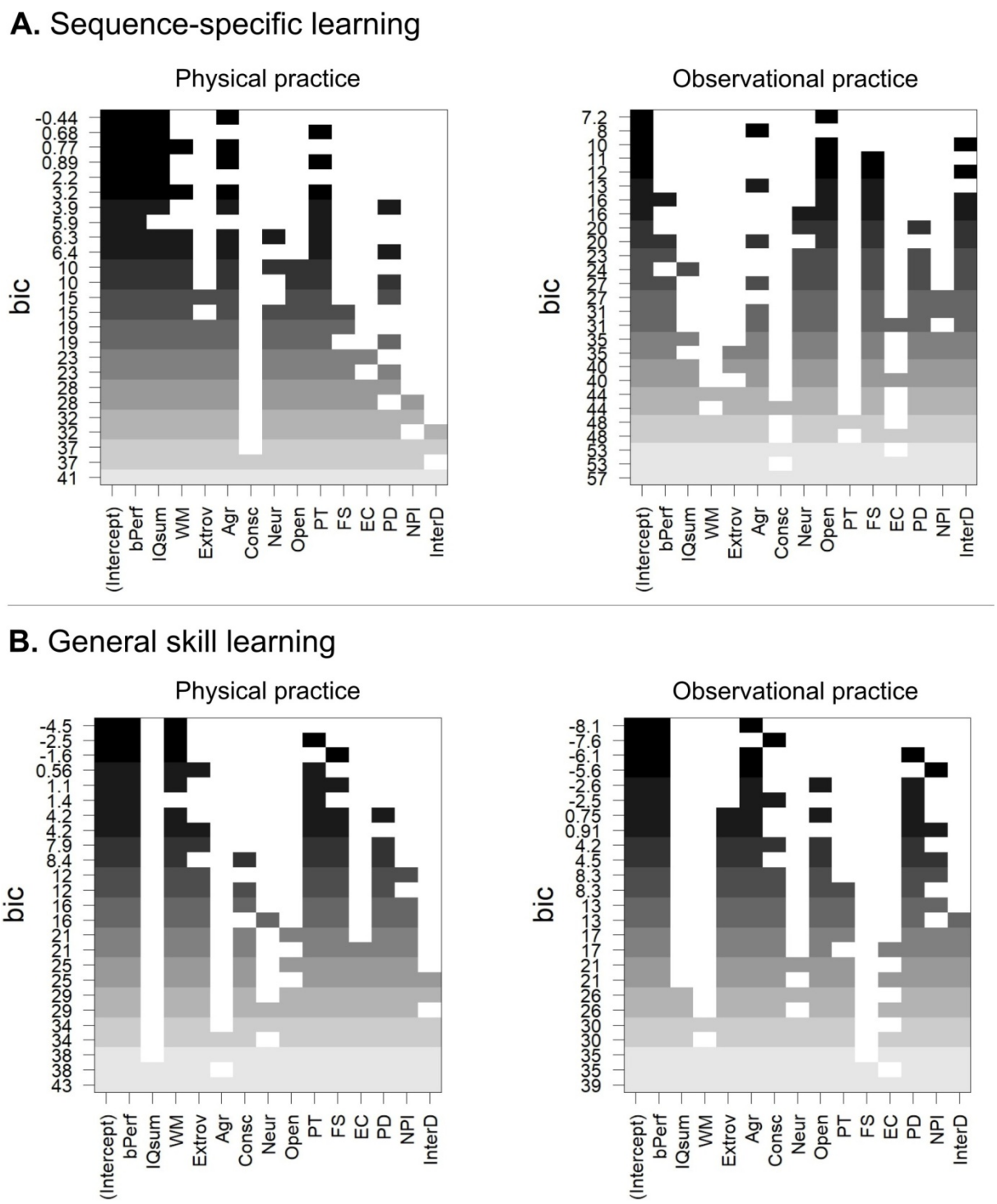
Supplementary Table 2. Exploratory analyses including personality -- The winning models for general skill learning.

Physical practice				Observational practice			
Model	$F_{2,89} = 9.651$, $p < 0.001$, $R^2 = 0.160$			**Model**	$F_{2,89} = 11.81$, $p < 0.001$, $R^2 = 0.192$		
Coefficients	**β [95% CI]**	**t**	**p**	**Coefficients**	**β [95% CI]**	**t**	**p**
Intercept	0	0	1	**Intercept**	0	0	1
Baseline perf.	-0.521 [-0.711 -0.331]	-4.341	0.00004	**Baseline perf.**	-0.431 [-0.619, -0.243]	-4.569	0.00002
Working memory	0.377 [0.138, 0.615]	3.139	0.002	Agreeableness	-0.182 [-0.370, 0.006]	-1.929	0.057



Supplementary Figure 1. Correlations between the variables. The figure shows correlation coefficient values and their representations as squares. Positive correlations are displayed in red and negative correlations in blue colour. Colour intensity and size of the squares are proportional to the magnitude of the correlation. Crossed squares represent non-significant ($p < 0.05$) correlations. **bPerf**, baseline performance; **IQsum**, fluid intelligence; **WM**, working memory; **PT**, perspective taking; **FS**, fantasy; **EC**, empathic concern; **PD**, personal distress; **InterD**, interdependence; **Extrov**, extraversion; **Agr**, agreeableness; **Consc**, conscientiousness;

Neur, neuroticism; **Open**, openness to experience; **NPI**, narcissism. **A.** Correlations between all 14 predictor variables. **B.** Correlations between the training effects and predictor variables.



Supplementary Figure 2. Model rankings. Model rankings of maximum two best models for each number of predictors (1-14) for the physical practice and observational practice groups. The rankings were based on Bayesian information criterion (BIC). A model with the lowest BIC (displayed at the top of each figure) is the model with an optimal combination of predictor variables that best explain the variance of the sequence-specific learning (**A**) or general skill learning (**B**). Shaded fields indicate which predictors are selected for the model; **bPerf**, baseline performance; **IQsum**, fluid intelligence; **WM**, working memory; **Extrov**, extraversion; **Agr**, agreeableness; **Consc**, conscientiousness; **Neur**, neuroticism; **Open**, openness to experience; **PT**,

perspective taking; **FS**, fantasy; **EC**, empathic concern; **PD**, personal distress; **NPI**, narcissism; **InterD**, interdependence.