ESM 3

Data Analysis: Main RQ

To analyze the effects of both scaffolds and their combination on CL, UV, and judgment accuracy, we used descriptive statistics, a repeated measures ANOVA (modelling effects of time, condition and the interaction time x condition), and paired *t*-tests to compare the pre- and post-test measures of the conditions in the 2 × 2 design. Also, we compared pre-test values between the four conditions using ANOVAs.

Data Analysis: Exploratory RQ

To further evaluate the effects of SEp (independent variable) on the effectiveness of the scaffolds for supporting PSTs' assessments (Exploratory RQ), we estimated marginal means in a linear mixed model. The model contained the independent variables time, condition, SEp, as well as their interactions. It can be displayed as the following pseudo-algebraic expression:

$$\begin{aligned}\text{judgment accuracy} \sim\ & \text{time} + \text{SEp} + \text{condition} \\ & + \text{time} \times \text{SEp} + \text{time} \times \text{condition} + \text{condition} \times \text{SEp} \\ & + \text{condition} \times \text{time} \times \text{SEp}\end{aligned}$$

We calculated judgment accuracy in the pre- and post-test for participants with one *SD* above and below the mean of SEp, based on this model. We did this by using estimated marginal means, as these explicitly use the calculated linear model, which was fitted to all $N = 108$ participants' data (as opposed to using PST subgroups), and allows point (±1 SD) rather than interval estimates (> +1 SD and < -1 SD). When estimating the marginal mean of a certain condition in the pre- or post-test for participants one *SD* above/ below the mean on SEp, the corresponding coefficients from the linear mixed model are summed up to estimate the corresponding judgment accuracy (see ESM 4 for coefficients; note that coefficients referring to SEp were also weighted with SEp). At last, we compared these pre- and post-test accuracies using a *t*-distribution.

Power analysis

To analyze effects of the conditions over time, we used a repeated measures ANOVA, focusing on the within-between interaction (*condition x time)*. We performed an a priori power analysis for small ($f = .10$), medium ($f = .25$), and large effects ($f = .40$). The number of

https://doi.org/10.1024/1010-0652/a000362

participants required for a power of $\beta = 0.80$ can be seen in ESM3, Table 1 ($\alpha = 0.05$, correlation between measurements = 0.5).

$f = .10$	$f = .25$	$f = .40$
280 participants	48 participants	24 participants

ESM3, Table 1. Participants required for a power of $\beta = 0.80$.

The results from the power analysis indicate that our sample size of $N = 108$ is sufficient to find large, medium and small to medium effect sizes. However, small effect sizes may not be detected.

https://doi.org/10.1024/1010-0652/a000362