

Appendix D: Pilot Reliability

Table S1. *Statistics from the pilot reliability study*

Rater configuration	Items included (trials)	Cronbach's α	ICC(3,1) single measures (95% CI)	ICC(3,k) average measures (95% CI)
All raters, both sessions (Monica, Francie, Josiah)	6	0.922	0.640 (0.585–0.692)	0.914 (0.894–0.931)
Francie & Josiah, both sessions	4	0.921	0.746 (0.705–0.784)	0.922 (0.905–0.935)
Francie intra-rater (Trial 1 vs 2)	2	0.863	0.759 (0.705–0.805)	0.863 (0.827–0.892)
Josiah intra-rater (Trial 1 vs 2)	2	0.907	0.831 (0.790–0.864)	0.907 (0.883–0.927)
Monica intra-rater (Trial 1 vs 2)	2	0.860	0.752 (0.696–0.799)	0.859 (0.821–0.888)

Clarification: These reliability estimates were calculated for the preliminary 32-criterion pilot rubric. Following the pilot, criterion wording was clarified and two criteria were removed, yielding the 30-criterion final rubric used for the primary analysis. Reliability was not re-estimated after these modifications.