**SUPPLEMENTARY APPENDIX A**

**Example of EPA F1:**

A screenshot of a cell phone

Description automatically generatedA close up of a logo

Description automatically generated

**SUPPLEMENTARY APPENDIX B**

**Simulation Cases**

|  |
| --- |
| Anaphylaxis |
| Atrial Fibrillation |
| Upper GI Bleed |
| Status Asthmaticus |
| Sepsis |
| AAA |
| Isolated Head Injury |
| Status Epilepticus |
| Multi-system Trauma |
| Difficult Airway |
| Congestive Heart Failure |
| Excited Delirium |
| Hyperkalemia |
| Pre-eclampsia |
| Pediatric Arrest |
| Pediatric Opioid Toxicity |
| Burn with Carbon Monoxide |
| Code Blue in Waiting Room |

**SUPPLEMENTARY APPENDIX C**

Generalizability Theory:

A strength of G-theory as a measure of reliability is its ability to deal with unbalanced designs such as the present study. That is, because each resident had a different number of EPA assessments completed in both workplace and simulation contexts, G-theory allowed us to obtain reliability coefficients by treating EPA assessments as a facet nested within resident. G-coefficients are interpreted in the same way as other reliability metrics, with values ranging from 0-1.[[1]](#footnote-1) A G-coefficient of 0.80 is the generally accepted threshold for high stakes assessments, while lower coefficients of 0.70 and 0.60 are often accepted for intermediate and low stakes assessments respectively.[[2]](#footnote-2) Given the low-stakes, formative nature of EPAs, a dependability analysis was conducted to determine the number of assessments per resident needed to obtain a reliability of 0.6.

Lin’s Concordance Correlation Coefficient:

To examine the relationship between EPA ratings obtained in the workplace and those obtained in simulation environments, mean simulation EPA ratings were calculated by averaging ratings from both raters for each observation and then calculating an overall mean for each resident. Individual ratings were collapsed into mean ratings per resident and compared using Lin’s concordance correlation coefficient (CCC).[[3]](#footnote-3) To account for the potential impact of measurement error, we also calculated a disattenuated correlation coefficient,[[4]](#footnote-4) using the reliability estimates obtained from the aforementioned G-theory analyses.

Within-subjects Analysis of Variance:

To determine whether mean EPA ratings from the simulated and workplace environments differed as residents progressed through their training, we collapsed data into three 4-month blocks: months 1-4, months 5-8 and months 9-12. A within-subjects analysis of variance (ANOVA) was conducted using the mean ratings as the dependent variable and environment (2 levels: simulation vs. workplace) and training month (3 levels: months 1-4, months 5-8, months 9-12) as independent variables. This factorial design allowed us to examine: (1) the effect of training environment on EPA ratings, irrespective of month of training (i.e., main effect of environment), (2) the effect of month of training on EPA ratings, irrespective of training environment (i.e., main effect of month), and (3) the interaction of these two factors on mean EPA effects (i.e., does the difference between EPAs ratings for simulation and workplace training environments differ depending on month of training). Bonferroni corrections were applied to all multiple pairwise comparisons. Effect sizes were calculated using partial eta-squared (η2) for ANOVAs and Cohen d for *t* tests. The magnitude of these effect sizes was interpreted using classifications proposed by Cohen: small effect sizes ≈ η2<0.02 and Cohen d<0.2; medium effect sizes ≈ 0.02<η2<0.13 and 0.2<d<0.8; and large effect sizes ≈ η2>0.14 and d>0.8.[[5]](#footnote-5),[[6]](#footnote-6)

1. Brennan R. Generalizability theory. Educ Meas Issues Pract. 1992;11(4):27–34. [↑](#footnote-ref-1)
2. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004 Sep;38(9):1006–12. [↑](#footnote-ref-2)
3. Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45:255–68. [↑](#footnote-ref-3)
4. Muchinsky P. The correction for attenuation. Educ Psychol Meas. 1996;56(1):63–75. [↑](#footnote-ref-4)
5. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Revised Edition. Hillsdale, NJ: Laurence Earlbaum Associates Inc.; 1988. [↑](#footnote-ref-5)
6. Cohen J. Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs. Educ Psychol Meas. 1973;33(1):107–12. [↑](#footnote-ref-6)