DEVELOPMENT AND TESTING OF AN OBJECTIVE STRUCTURED CLINICAL EXAM (OSCE) TO ASSESS SOCIO-CULTURAL DIMENSIONS OF PATIENT SAFETY COMPETENCY

TECHNICAL APPENDIX

Inter-rater Reliability

Inter-rater reliability was examined using the equivalent[1] of a weighted Kappa[2]– a two-way mixed, single-measures, consistency ICC. A weighted Kappa takes into account the magnitude of a disagreement between raters and is typically used for categorical data with an ordinal structure. Quadratic weights are used if the difference between the first and second category is less important than a difference between the first and third category, etc., (which was the case with our data).

Station 2 Reassessment

All data presented in the paper for station 2 are data taken from the reassessment of station 2 that was carried out (due to the fact that one assessor was in a conflict of interest position and realized it was difficult to provide objective assessments during the OSCE). However, due to the fact that we did not use student identifiers that could be linked back to the students in this study, we were unable to link the station 2 reassessment scores with students' scores on the other three stations. This did not pose a problem for our analyses because this pilot had only four stations and we were therefore not seeking to comment on the reliability of the OSCE as a whole. Instead, our analyses tend to be within station analyses. One exception are the data presented in Table 3 on the overall performance of the trainees on the group of stations. We wanted to be able to report on the proportion of trainees that scored very high (and very low) on several stations. In order to do this, we have, in table 3, used data from the initial rater in station 2 who was not in a conflict position.

Table 3 Calculation

Table 3 provides data on overall performance on the group of stations and shows that ~=25% of trainees scored greater than 3 on three or four stations. If a trainee had a mean station score >= 2.95 the score was rounded to 3 and they were included as scoring in the 3+ range for the station.

References

1. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial*. Tutor Quant Methods Psychol* 2012;8(1):23-34.

2. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit*. Psychol Bull* 1968; Oct;70(4):213-20.