# ODU

## **Reliability and Consistency of Standardized Patient Ratings Using** the Master Interview Rating Scale: What do the Measures Mean?

### Introduction

The Master Interview Rating Scale (MIRS) is an instrument used by many standardized patients (SPs) to assess learners. It has 27 items that can be selected to match clinical cases and has been shown to be a reliable measure of communication skills.2,3 However, little is known about the reliability and consistency of SP use of MIRS items.

The goal was to examine the reliability and consistency of two SP raters on their assessments for MIRS. Data were pulled from a larger set as an example of anomalies that can occur when reviewing. For these data, t, two SPs rating the same learners on items from a sample of scenarios were analyzed. Correlations and intraclass correlations (ICCs) were calculated. Correlations range from -1 to +1 4,5 and typical ICCs range from 0 to 1 with higher values indicating more consistency among the ratings. Thus, it was hypothesized that two SPs rating the same learners with standard MIRS items would demonstrate positive correlation and high ICCs.

#### Mark W. Scerbo Kimberly Perry **Old Dominion University**

#### **Amelia Wallace Bruce Britton** Matt Cole Mary Rubino Eastern Virginia Medical School, Norfolk, VA

### Method

Two SP raters scored 10 videos depicting a learner interacting with another (SP). The SP raters scored the learner on 7 MIRS items: narrative, impact of illness, verbal facilitation, nonverbal facilitation, patient education, achieving a shared plan, and overall interview technique items. Each item was scored on a five-point scale with verbal anchors (higher values reflecting better performance). We used the MIRS version modified at Eastern Virginia Medical School in 2005. A small dataset was specifically selected to illustrate both typical and unusual values.

#### Results

The positive Spearman Rho values indicate agreement between raters that learners performed high or low on the item. Negative values show that raters disagreed as to whether learners were high or low on the item. Typical values for the ICC are interpreted as follows: scores less than 0.40 are poor, 0.40 to 0.59 is fair, 0.60 to 0.74 is good, and 0.75 to 1.00 is excellent.5 However, the ICC can be negative

(Achieve Shared Plan). In this case, the variability of scores for one rater exceeded the variability between raters.6 Last, the Rho and ICC could not be calculated for the Verbal item because one rater assigned the same score to every learner.

#### **Descriptive Statistics for MIRS Items Aggregated Across Learners**

Stand Narrative Impact of Nonverbal Patient Edu Achieve a S **Overall Inte** Verbal Mean (SD)

Our goal was to gain a better understanding of SP use of MIRS items by focusing on the nature of the values and not the absolute values. The correlations vary and show fair agreement on some items as hypothesized, but also some unexpected differences. Negative correlations indicate raters viewed learners very differently and may be an area requiring some recalibration. This is also true for the

lard Instrument	Spearman Rho	ICC (Averaged)
	.499	.277
llness	.542	.300
	.385	.259
ucation	.443	.624
Shared Plan	186	240
erview Technique	.415	.631
	Not computed	Not computed
	.463 (.069)	.449 (.183)

Conclusion

Verbal item in this dataset suggesting this rater may not be sensitive to differences among learners. The ICCs lower than the Rho values indicate that raters tend to agree but assign different scores. The ICCs higher than the Rho values indicate that raters tend to agree, and often assign same score. This analysis was based on a subset of MIRS items but could be expanded to the full set. Further, by examining two SP ratings of the same learner it enables a finer tuning of scoring not possible under conditions where SPs rate different learners across the different cases.

#### References

- 1984; 6(4): 284-290.



1.Pfeiffer C. The Master Interview Rating Scale (MIRS): An Instrument for Faculty Students and Standardized Patients to Use for Teaching and Evaluation. Available from: https://icollaborative.aamc.org/resource/835/ 2.Stillman PL, et al. Construct validation of the Arizona Clinical Interview Rating Scale. Educ Psychol Measure 1977;37(4):1031-8.

3.Schirmer JM, Mauksch L, Lang F, Marvel KM, Zoppi K, Epstein RM, Brock D, Pryzbylski M. Assessing Communication Competence: A Review of Current Tools. Fam Med 2005;37(3):184-92.

4.Gwet, K.L. Handbook of inter-rater reliability, 4th ed. Gaithersburg, MD: Advanced Analytics, LLC. 2014.

5.Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psych Bull 1979; 86(2): 420-8.

6.Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psych Assess