PATIENT SAFETY CLIMATE STRENGTH – A CONCEPT THAT REQUIRES ATTENTION

***BMJQS* ON-LINE ONLY TECHNICAL APPENDIX**

1. Rwg(j) Distribution

- Rwg (James, Demaree, & Wolf, 1984) represents the observed variance (a reflection of disagreement) in ratings compared to the variance of a theoretical distribution representing no agreement (i.e., the null distribution)
- Rwg(j) is calculated based on a theoretical expectation that responses to safety climate items may be uniform or may be upwardly biased (e.g. slightly skewed). While standard practice is to report Rwg's based on a uniform distribution, this often yields the largest Rwg values (LeBreton and Senter, 2008).
- We report Rwg(j) for both the uniform and slightly skewed distributions (in column C of table 1). The values for the slightly skewed distribution represent the lower bound of the true agreement (Smith-Crowe, et al., 2014). The Rwg(j) calculated based on a uniform (i.e. rectangular) distribution represents the upper bound of the likely true agreement (Smith-Crowe, et al., 2014).

2. Measures of Interrater Reliability (IRR) versus Interrater Agreement (IRA)

- Interrater Reliability (IRR) refers to a correspondence or correlation between the rank-ordered ratings of PS survey respondents within a unit. In multilevel climate research (in which staff are "nested" within their work units), IRR is typically measured by the ICC(2) which gives a measure of how reliable the means of a group are (how much each staff member's survey responses correlate with others in the group; see LeBreton & Senter, 2008). In contrast, Interrater Agreement (IRA) refers to the extent to which the staff surveyed on a unit exactly agree in their answers or ratings to any particular item on a PS scale. Rwg(j) is the most commonly reported measure of IRA. The ICC(1) indicates the proportion of variance in scale scores, across all units measured, which is attributable to group membership, and is interpreted as providing information on both IRR and IRA (LeBreton & Senter, 2008).
- It is important to note that while most commonly reported, Rwg(j) is only one of several measures researchers and practitioners can employ to calculate units' IRA for multi-item PS climate surveys. Statistical criticisms of Rwg(j) prompted the development of several corrected or improved Rwg(j) scales (such as R*wg(j) and $Rwg_p(j)$, see summary in LeBreton & Senter, 2008), and alternatively calculated IRA scales using squared deviations from the mean (such as the simple standard deviation, SDx) and absolute deviations from the mean (absolute deviation or $AD_M(j)$) or median ($AD_{MED}(j)$). Finally, there is an index of IRA called Awg(j) which intends to correct problems of Rwg(j) being dependent on the number of anchors and respondents, and the null distribution chosen (see point #3 below for more information). LeBreton and Senter (2008) note that all of these indices have strengths, but they also tend to correlate highly with one another and tell a similar story. Further, they support our proposed practice in this paper that SDx, the simple standard deviation, is a simple and best metric to use for looking at variability in agreement among units in terms of their PS climate.

3. Statistical Significance in Agreement Indices

- Using the critical values reported in a recent paper on statistical significance in agreement indices (including Rwg) for different sample sizes (Smith–Crowe, 2014) we can infer the statistical significance of the Rwg(j) values based on the uniform distribution reported in column C of table 1.
- In addition, one-way ANOVA is typically used to test between-group variance when reporting ICCs. Using unaggregated data and ED membership as the independent variable, results indicated that four of the six safety climate subscales showed significant between group variance: *Organizational Leadership Support for Safety* scale, $F(23,418) = 4.70$, $p < .001$; *Incident follow-up* scale, $F(23,418) = 3.80$, $p < .001$; *Unit learning* scale, $F(23,418) = 2.49$, $p < .001$; *Supervisory leadership for safety* scale, $F(23,418) = 2.66$, $p < .001$. Differences between EDs on the *Judgement-free environment* and *Job repercussions of error* scales were not significant (*Judgement-free environment* scale , $F(23,418) = 1.50$, $p = .065$; *Job repercussions of error* scale, $F(23,418) = 1.2$, $p = .241$).

4. Comparison of ICC and Rwg Results

- Our Rwg and ICC(1) results are inconsistent for two of the PSC dimensions we examined: *judgement-free environment* and *job repercussions of error*. There are several possible explanations for this and consideration of these contributes to understanding of what agreement metrics provide. It is important to recall that Rwg assesses the extent of agreement within a single unit (in this case, an ED)—"a construct by group approach" while ICC(1) contrasts within-unit and between-unit variability across a sample of units—"a construct by sample approach" (Klein & Kozlowski, 2000). Therefore, while between-group variance is a factor in the calculation of ICCs, it is not taken into account in the calculation of Rwgs. This means that the lack of significant ANOVA for the Promo and Reputation scales (see section 2 above) may account for the low ICC(1)s for these two dimensions and may explain the inconsistency between the Rwg and ICC results for these two dimensions of PSC. In other words, it may be that within-group agreement is similarly high among all six dimensions, but that between-group variance is much lower for the *judgement-free environment* and *job repercussions of error* dimensions – scores on these dimensions do not vary much across settings.

REFERENCES

James RL, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology* 1984;(69):85-98.

Klein KJ, Kozlowski SWJ. From Micro to Meso: Critical Steps in Conceptualizing and Conducting Multilevel Research. *Organizational Research Methods* 2000; 3(3):211-36.

LeBreton JM, Senter JL. Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods* 2008; 11(4):815-52.

Smith-Crowe K, Burke MJ, Cohen A, et al. Statistical significance criteria for the rWG and average deviation interrater agreement indices. *J Appl Psychol* 2014; 99(2):239-61.