# Options in Agreement Indices for Establishing Expert Consensus on Behavioral Ratings within School and Industrial/Organizational Psychology

Rose Jaffery[1], Austin H. Johnson[1], Mark C. Bowler[2], Sandra M. Chafouleas[1], & T. Chris Riley-Tillman[3]

[1]*University of Connecticut* [2]*East Carolina University* [3]*University of Missouri*

## Abstract

Various options exist for calculating agreement between raters. In the field of school psychology, correlation, percent agreement and kappa are commonly used (Watkins & Pacheco, 2001). However, if we peruse the industrial/organizational (I/O) psychology literature we find several agreement indices that are commonly used in that field (e.g., intraclass correlation coefficient or ICC, $r_{wg}$, $a_{wg}$). There are several limitations associated with each; however, trends within fields often drive agreement options, despite limitations. An example of how such considerations can be applied to an expert consensus-building procedure for establishing true score estimates of student behavior is discussed. In this example, ICC and $r_{wg}$ were deemed the most appropriate and informative indices to use.

## Introduction

**Background**

In educational behavioral assessment, systematic direct observation (SDO) and direct behavior rating (DBR) are two methods of providing an estimate of behavior duration frequently and in a standardized manner (Chafouleas, Riley-Tillman, & Christ, 2009).

• SDO typically involves marking the presence or absence of a target behavior during short pre-specified intervals (e.g., every second, or every 15 seconds) during a target activity (e.g., for 20 minutes during math).
• DBR involves making a brief rating of behavior immediately following a target activity (e.g., after 45 minutes of independent reading).

SDO has long been considered the gold standard of behavioral measurement methodologies; therefore, it may be reasonable to assume that scores derived from SDO could serve as a *true score estimate* for DBR. However, there are specific differences between SDO and DBR. Establishing expert DBR scores may be a better alternative for determining true score estimates of behavior.

**Objective**

Expert consensus-building procedures were conducted in order to establish true score estimates of the duration of student behaviors displayed in video clips of elementary classrooms. Procedures were modeled off of procedures used in I/O psychology (e.g., Borman, 1977; Murphy et al., 1982). The aim was to calculate indices of agreement on data obtained through multi-site expert consensus-building sessions. However, there are several agreement indices available, so investigations were made to determine the strengths and limitations of each index and which would be most appropriate.

Preparation of this poster was supported by a grant from the Institute for Education Sciences (IES), U.S. Department of Education (R324B060014). For additional information, please direct all correspondence to Sandra Chafouleas at sandra.chafouleas@uconn.edu

## Method

- ***Obtain expert ratings*:** Expert consensus building procedures consisted of 13 school psychology professors and advanced graduate students across two university-based sites viewing 18 one-minute video clips and rating one student on one target behavior after each clip. Initial ratings were discussed among the group and individuals were allowed to change their ratings after discussion.
- ***Determine most appropriate indices to use to calculate rater agreement***: This involved searching the literature regarding agreement indices across fields, then evaluating these various indices and how they could be applied to ratings obtained through expert consensus-building procedures in order to establish true score estimates of videotaped student behavior.
- ***Calculate indices of agreement***: Using expert ratings obtained through consensus building procedure.

## Agreement Options

- **Pearson's r** ~ -1.0 – +1.0 scale
  - Statistical relationship between two sets of data (correlation)
- **Percent Agreement** ~ 0 – 100% scale
  - # of agreements / (# of agreements + # of disagreements)
- **Kappa** ~ 0.0 – 1.0 scale
  - A measure of interrater agreement between two raters for categorical items that takes into account the agreement that occurs by chance.
  - Overly conservative measure due to minimum category frequencies needed.
  - Inappropriate for calculating extent of agreement between several raters.
- **Intraclass Correlation Coefficient (ICC)** ~ 0.0 – 1.0 scale
  - "Indication of the absolute consensus among raters in that it provides information regarding both rater consensus and relative rater consistency"
  - ICC(k) refers to an ICC that applies to average measurements, whereas ICC (1) applies to a single measurement
  - ICC(k) answers: "Do judges' mean ratings reliably distinguish among the groups/targets? Is there sufficient interrater reliability and agreement to justify aggregating the data?"
- **$r_{wg}$** ~ 0.0 – 1.0 scale
  - "Proportion of non-error variance in ratings"
  - Most popular measure of interrater agreement in I/O psychology
  - Assumes rating target has one "true score"
  - Controls for response bias' impact on scores, in part, by tailoring distribution to known rater biases.
- **$a_{wg}$** ~ 0.0 – 1.0 scale
  - Uses principles from kappa and adapts it from "two raters rating multiple stimuli on a categorical scale" to "agreement among multiple raters rating a single continuous construct," as is present within our study.
  - Has been posited as controlling for several issues present with $r_{wg}$
  - However, requires a minimum sample size based on # of points on the scale.

## Results

**ICC(k)** and **$r_{wg}$** were deemed the most appropriate for our purposes.
• **Pearson's r, percent agreement,** and **kappa** – not appropriate due to # of expert raters, & lack of detailed agreement information (only provide an overall score).
• **$a_{wg}$** – minimum sample size required based on the number of scale points, so we would need 10 raters per site for an 11-point DBR scale.
• **ICC(1)** - Large values for ICC(1) and ICC(k) would indicate that ratings were a function of the clip being rated. Of the two, ICC(k) is the most appropriate for decision-making as it represents a combination of reliability and agreement indices.

- Thus, using **ICC(k)** in combination with strong **$r_{wg}$** values indicate that it is appropriate to aggregate the ratings obtained during the expert consensus-building procedures employed in this study to create true score estimates. Indices of agreement were calculated for all clips by site using $r_{wg}$ and ICC. Across all clips, agreement in ratings improved after consensus building.

Table 1. *Expert Agreement via $r_{WG}$ and ICC(k)*

| Clip | Behavior | Site 1 | | Site 2 | | Aggregated | |
|---|---|---|---|---|---|---|---|
| | | Initial $r_{WG}$ | Final $r_{WG}$ | Initial $r_{WG}$ | Final $r_{WG}$ | Initial $r_{WG}$ | Final $r_{WG}$ |
| 1 | RS | .81 | .94 | .80 | .83 | .69 | .83 |
| 2 | AE | .92 | .99 | .97 | .97 | .94 | .98 |
| 3 | DB | .97 | .97 | .97 | .97 | .97 | .97 |
| 4 | RS | .99 | .99 | .98 | .98 | .99 | .99 |
| 5 | RS | .80 | .95 | .65 | .77 | .71 | .82 |
| 6 | DB | .94 | .97 | .95 | .95 | .94 | .96 |
| 7 | AE | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 8 | AE | .16 | .84 | .35 | .83 | .27 | .85 |
| 9 | DB | .99 | .99 | .97 | .97 | .98 | .98 |
| 10 | RS | .54 | .97 | .78 | .95 | .65 | .96 |
| 11 | AE | .42 | .94 | .97 | .97 | .64 | .96 |
| 12 | AE | .89 | .97 | .57 | .87 | .79 | .93 |
| 13 | DB | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 14 | RS | .97 | .97 | .92 | .92 | .94 | .94 |
| 15 | DB | .44 | .84 | .78 | .93 | .59 | .88 |
| 16 | DB | .89 | .97 | .90 | .90 | .90 | .94 |
| 17 | AE | .98 | .98 | .97 | .97 | .97 | .97 |
| 18 | RS | .97 | .97 | .87 | .98 | .94 | .94 |
| $M_{rWG}$ | | .82 | .96 | .86 | .93 | .83 | .94 |
| $s_{rWG}$ | | .25 | .05 | .18 | .07 | .20 | .06 |
| ICC(K) | | .984 | .997 | .978 | .989 | .990 | .997 |

*Note.* AE = Academically Engaged; DB = Disruptive; RS = Respectful.

## Summary and Conclusions

In determining a true score estimate from these expert consensus ratings, indices that provide an overall agreement score are insufficient, as they do not indicate whether the agreement value obtained was a function of the *clip* or the *rater*. ICC and $r_{wg}$ may be more appropriate in this case, yet limitations regarding these indices should be considered. ICC assumes that a random sample is used; however targets (video clips) used in this study were not randomly selected. Issues with $r_{wg}$ include (a) values are scale dependent (values derived from a 5 vs. 7 pt scale are not comparable), (b) sample size influences interpretability, and (c) the uniform null distribution assumption ("if there is no variance related to agreement, then raters disagree uniformly"; Brown & Hauenstein, 2005). Despite such limitations, these indices provide more robust information than the other available indices and thus are utilized frequently within the I/O literature for evaluating agreement between ratings obtained through expert consensus-building procedures. Such indices should be considered for use in school psychology research as well.