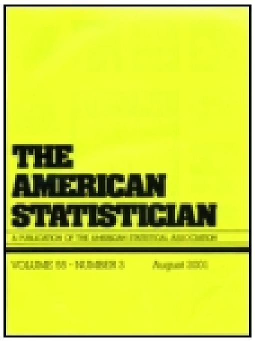


Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution?

David Quarfoot & Richard A. Levine

To cite this article: David Quarfoot & Richard A. Levine (2016) How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution?, The American Statistician, 70:4, 373-384, DOI: 10.1080/00031305.2016.1141708

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1141708

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 4, 373–384
http://dx.doi.org/10.1080/00031305.2016.1141708

How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution?

David Quarfoot[a] and Richard A. Levine[b]

[a]Center for Research in Mathematics and Science Education, San Diego State University, San Diego, CA, USA; [b]Department of Statistics, San Diego State University, San Diego, CA, USA

ABSTRACT
Interrater reliability studies are used in a diverse set of fields. Often, these investigations involve three or more raters, and thus, require the use of indices such as Fleiss's kappa, Conger's kappa, or Krippendorff's alpha. Through two motivating examples—one theoretical and one from practice—this article exposes limitations of these indices when the units to be rated are not well-distributed across the rating categories. Then, using a Monte Carlo simulation and information visualizations, we argue for the use of two alternative indices, the Brennan–Prediger coefficient and Gwet's AC2, because the agreement levels reported by these indices are more robust to variation in the distribution of units that raters encounter. The article concludes by exploring the complex, interwoven relationship between the number of levels in a rating instrument, the agreement level present among raters, and the distribution of units that are to be scored. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2015
Revised November 2015

KEYWORDS
Agreement; Conger; Fleiss; Gwet; Krippendorff; Paradox

1. Introduction

In fields as diverse as educational research, medicine, and business, a common need arises: the ability to measure how closely a group of raters agree when scoring some phenomenon. This closeness, known as interrater reliability (IR) or interrater agreement (IA), is particularly helpful in assessing the usefulness of the instrument that generated the ratings, the thoroughness of training the raters have received, and the clarity of the idea to be scored (e.g., creativity in mathematics problems, diagnostic tests in medicine, leadership in workplace bosses). The challenge in defining such IR/IA measures is that one must decide what "agreement" means, and these choices can sometimes lead to surprising results.

Before delving further, it is important to clarify how the term "interrater reliability" is used in this article. Unfortunately, this term has seen differing interpretations in the literature, depending on discipline and context. For example, in the field of organizational research, interrater reliability measures the degree of *rank-order* agreement among raters, while interrater agreement measures how close raters' *actual values* are to one another (LeBreton and Senter 2008). Thus, two teachers' test scores have high IR if they simply put the students in a similar order (from best to worst), even if the particular grades are very different. In contrast, researchers in the medical field tend to use the phrase interrater reliability when the goal is to measure how close raters' *actual values* are to one another (Gwet 2014). Thus, the term IA in one field is the same as the term IR in a different field.

In addition to this confusion, the study of IR (in medicine) and IA (in organizational research) has seen significant methodological development over the past 50 years. While IR/IA was originally calculated using nominal data scales with two raters and no missing scores, it has since expanded to include all types of data (nominal, ordinal, interval, and ratio levels of measurement), two or more raters, and even missing data (see, e.g., Janson and Olsson 2001). For the work below, we follow the terminology and meaning of authors like Gwet who use the phrase "interrater reliability" to mean the amount of agreement present in the *actual scores*, and adopt a modern stance that includes the use of multiple raters (three or more) and an interval level of measurement.

The goal of this article is two-fold. First, through the use of two motivating examples, one manufactured and one real, we revisit a paradox well-documented in the literature on interrater reliability, but with an eye toward the case of multiple raters. The artificial example seeks to pique the reader's interest from a theoretical perspective, showing that in a contrived setting, it is possible to experience counter-intuitive results. The real example, a dataset dealing with experts' assessments of Geometry problems, shows that this theoretical concern should be an actual concern, that is, the weaknesses inherent in standard IR indices can naturally be brought to the fore.

Spurred by these examples, the second goal is to explore the robustness (a term defined in Section 4) of many of the most common multirater (three or more) IR indices in more detail. While there has been some discussion of this topic in the research literature, it is limited to the simplest of setups: two raters, scales with only two rating categories, data at the nominal level, equal sensitivity and specificity levels (i.e., how often raters label category 1 choices as category 1, and category 2 as category

CONTACT David Quarfoot ✉ quarfoot@gmail.com 🖃 Center for Research in Mathematics and Science Education, San Diego State University, San Diego, CA 92182 and University of California, San Diego, CA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

Table 1. Scores of three hypothetical teams judging pilots' years of experience.

Team 1	P1	P2	P3	P4	P5	P6	P7
Rater 1	1	1	1	1	1	1	1
Rater 2	1	1	1	1	1	1	1
Rater 3	1	1	1	1	1	1	2
Team 2	P1	P2	P3	P4	P5	P6	P7
Rater 1	1	6	8	3	2	6	2
Rater 2	3	4	5	4	3	7	1
Rater 3	2	5	7	4	4	9	1
Team 3	P1	P2	P3	P4	P5	P6	P7
Rater 1	2	1	2	3	1	3	1
Rater 2	2	1	2	3	1	1	1
Rater 3	2	1	2	3	1	1	2

2), and related situational constraints. While these simplifications do allow for the derivation of closed-form expressions for the observed level of agreement and expected level of agreement due to chance (and hence, the IR value), they fail to explore the full richness of the IR landscape. It is now common to see studies with three or more raters, data at the ordinal and interval levels, and instruments with a large number of rating categories. As such, we use a Monte Carlo simulation to widen the exploration of the IR landscape.

2. A Theoretical Example

Consider first a theoretical case: Three teams of raters are assembled to estimate the number of years of experience they believe a set of airplane pilots have. Each team has three members and is located in a different city. For a given team, the three members will score a total of seven pilots (denoted P1 through P7), spending a day with each pilot. Table 1 shows the results from the three teams.

Intuition suggests that Team 1 should have the best IR, followed very closely by Team 3, with Team 2 somewhat worse. Looking at the data, it appears that while Team 1 was in a city with mostly rookie pilots, they agreed in all cases but one (P7). Team 3 had more diversity in its pilots and disagreed on only two pilots (P6 and P7). Meanwhile, Team 2 encountered quite a mix of experience levels and disagreed on the experience levels of all seven pilots.

It turns out that intuition is largely betrayed by the most commonly used IR indices. Table 2 presents the IR scores, to three decimal places, for these three teams using three different indices: Fleiss's kappa (a generalization of Scott's pi statistic; Fleiss 1971), Conger's kappa (Conger 1980), and Krippendorff's alpha (Krippendorff 2012).

These results were calculated in the statistical language R (R Core Development Team 2014) using implementations of these indices as outlined by Gwet (2014) and assume that the scores from Table 1 are viewed at an interval level of measurement. In general, an IR index returns 1 in the case of perfect agreement, 0 in the case of agreement equivalent to random chance, and negative values when the observed agreement is less than what is expect by chance—as in the case of intentional misrepresentation of ratings.

Table 2. IR values using three standard multirater indices for the teams in Table 1.

IR scores	Fleiss	Conger	Krippendorff
Team 1	− 0.050	0.000	0.000
Team 2	0.774	0.776	0.785
Team 3	0.593	0.598	0.612

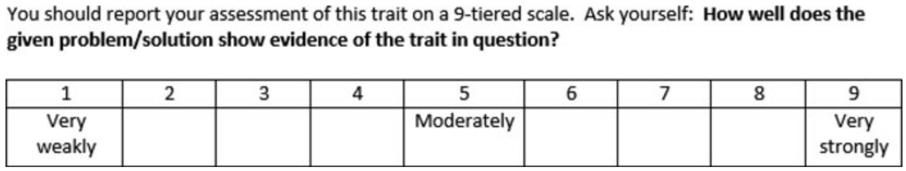
Two points are worth noting here. First, all three indices agree to within a few hundredths—so they agree with one another—and second, they order the three teams in precisely the *opposite* order that intuition would suggest. This type of IR paradox is well-documented in the literature, but only in the simple setups mentioned in Section 1 (see, e.g., Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990; Guggenmoos-Holzmann 1993; Nelson and Pepe 2000; Gwet 2002).

To understand this paradox, it is valuable to look at how IR values are determined. In general, IR indices calculate the observed level of agreement among the raters and then adjust this result by determining how much agreement could be expected from random chance. This is seen in the equation:

$$\text{IR} = \frac{(\text{observed agreement}) - (\text{chance agreement})}{1 - (\text{chance agreement})}. \tag{1}$$

Here, the numerator is the agreement beyond chance, and the denominator is the total possible agreement that could exist beyond chance. A quick example shows why this corrective action must be taken. Suppose two raters are asked to decide whether subjects are male or female simply by examining handwriting samples. At first blush, achieving a 75% agreement might appear impressive, but this finding must be tempered by the fact that the expected agreement due to random chance could be as high as 50% if the categories "male" and "female" appear with equal frequency in the test set. Thus, an effective IR index must look beyond the observed level of agreement and adjust for the fact that some agreement will appear simply due to chance. To do so, an index uses the histogram of ratings that were actually given (henceforth called the "frequency distribution" or FD) as the best estimate for the distribution of the units across the categories of the rating scale in the population. With this distribution in hand, it is possible, using probability theory, to figure out how much of the observed agreement is not really agreement after all, but instead, an artifact of throwing darts at a board with finitely many zones of certain sizes.

Something not readily seen in Equation (1) is that each IR index uses a different approach when calculating the (chance agreement) term. Indeed, the interested reader can find an overview of the common approaches and technical details in Gwet (2014). For some indices, the correction for chance agreement can be particularly draconian, especially when the distribution of the raters' scores is unbalanced. Gwet has shown this tendency in both computer simulations (2002) and mathematical analyses (2008), writing that the behaviors of pi and kappa indices are "very erratic . . . as soon as trait prevalence goes to the extremes" (2008, p. 35). Indeed, under a fixed level of agreement between two raters, it is possible to get *almost any IR value* between the observed percentage of agreement and zero simply by adjusting the frequency with which certain ratings appear (Guggenmoos-Holzmann 1993; Gwet 2002). Said differently, the (chance agreement) term in Equation (1) appears to be heavily influenced by the types of units the raters are asked to score (the FD), and this can result in unexpected behavior for the IR value. Sadly, this is precisely the *opposite* of what one should desire from a functioning IR index: the hope is that an index will show



You should report your assessment of this trait on a 9-tiered scale. Ask yourself: **How well does the given problem/solution show evidence of the trait in question?**

1	2	3	4	5	6	7	8	9
Very weakly				Moderately				Very strongly

Figure 1. Example of the rating scale used for the Geometry problems study.

robustness—that is, that it will look past the frequency distribution it is given and get an accurate read on the agreement level of the raters.

This discussion explains why traditional IR indices fail to match common sense in the pilot examples above: Team 1 from Table 1 has an extremely skewed FD (20 1's and one 2), while Team 3 has a strongly skewed distribution (10 1's, 7 2's, 4 3's). Hence, traditional chance-agreement-correction indices are so heavy-handed that they grossly underestimate what intuition suggests are very high levels of agreement when FDs are skewed. In contrast, Team 2 has a more balanced distribution across the rating levels and receives a higher IR. Thus, it appears that many common indices are powerfully impacted by frequency distribution, *even in the case of multiple raters*.

3. A Practical Example

Now, one might argue that the above airplane pilot example is contrived. As such, we supplement it with an actual research setting and dataset in which the issues discussed above were critically important. In this project, 8 expert raters—mathematicians, teachers, mathematics educators, and problem solvers—were sent a collection of 12 high school Geometry problems and asked to work and then rate them using a set of 14 different instruments (referred to as "metrics" below). These metrics were designed to gauge how much a trait like Difficulty or Novelty (metric names are capitalized) was present in each of the problems. All ratings were assigned on a 1–9 Likert scale similar to Figure 1.

The goal of the study was to determine the level of agreement among experts when assessing features of mathematical problems. The need for such a skill is critical, for many national standards documents and articles in the mathematics education literature suggest assigning problems with specific features—for example, cognitive demand, multiple solutions, high authenticity, etc. (National Council of Teachers of Mathematics 1991, 2000; Gutiérrez 2007, 2013; National Governors Association Center for Best Practices, Council of Chief State School Officers 2012). After collecting the raters' scores, a separate IR analysis was conducted for each of the 14 instruments (see column names in Table 3) for five different IR indices (rows in Table 3). These indices were Fleiss's kappa, Conger's kappa, Krippendorff's alpha, the Brennan-Prediger coefficient, and Gwet's AC2 index (Fleiss 1971; Conger 1980; Brennan and Prediger 1981; Krippendorff 2012; Gwet 2014). While the first three indices are well-cited, to our knowledge, the Brennan–Prediger coefficient is not well-known, and Gwet's AC2 index has been applied only a limited number of times in the literature (see, e.g., Baethge, Franklin, and Mertens 2013; Wongpakaran et al. 2013; Lang et al. 2014). It is important to note that each of these indices uses a different approach to correct for chance agreement, and as such, responds differently to the frequency distribution paradox discussed above. We collect Fleiss's kappa, Conger's kappa, and Krippendorff's alpha under the label "Group 1" indices, and the Brennan–Prediger coefficient and Gwet's AC2 under the label "Group 2" indices.

We note upfront that the small number of raters ($n = 8$) and problems ($u = 12$) in this example preclude robust statistical inference; indeed, the 95% confidence interval for each IR value in Table 3 is quite wide, with standard errors often around 0.2 (see Gwet 2014, chap. 5 for details on inference in IR calculations). Nevertheless, we try to sketch some general themes that are useful from these IR point estimates. One immediate observation from Table 3 is just how different the IR results are for different indices. For example, using any of the Group 1 indices, the Productive Dispositions metric—an assessment of how likely a problem is to instill a positive, rich view of mathematics—appears to have very weak reliability (0.17–0.20). Under the Group 2 indices, the IR has skyrocketed to 0.61 or 0.71! Similar leaps can be seen in Authenticity (0.11–0.74), Affective Engagement (0.30–0.76), and several of the other metrics. Each of these differences serves as an important example of how a researcher's choice of IR can have a profound impact on the conclusions he or she draws from an IR analysis. With a Group 1 IR choice, an instrument might appear hopelessly flawed or the raters using it might seem incapable of agreement; under a Group 2 IR choice, the same instrument could be deemed perfectly reliable and the raters in agreement.

Why do these IR indices, which purport to measure the level of agreement in a set of ratings, return such different results for a fixed metric? As with the airplane pilot example, the answer lies in the frequency distribution paradox. If, for example, we

Table 3. IR values for each of the metrics using five different indices. These values are found by choosing a metric, collecting the scores of the 8 raters across the 12 problems for that metric, and then calculating the IR using one of the possible indices.

IR index	Difficulty	Elegance	Novelty	Cognitive sophistication	Internal resource collaboration	Number of steps	Creativity	Representational media	Resource creation	Presence of misconceptions	Number of solutions	Productive Dispositions	Affective Engagement	Authenticity
Fleiss	0.62	0.42	0.47	0.44	0.31	0.48	0.42	0.37	0.23	0.42	0.29	0.17	0.30	0.11
Conger	0.62	0.43	0.48	0.45	0.32	0.49	0.43	0.38	0.24	0.43	0.30	0.20	0.31	0.13
Kripp	0.62	0.42	0.48	0.45	0.32	0.49	0.43	0.37	0.24	0.43	0.30	0.18	0.30	0.12
B-P	0.63	0.57	0.50	0.70	0.61	0.86	0.56	0.62	0.58	0.74	0.55	0.61	0.66	0.59
Gwet	0.67	0.64	0.54	0.75	0.69	0.91	0.63	0.72	0.70	0.81	0.78	0.71	0.76	0.74

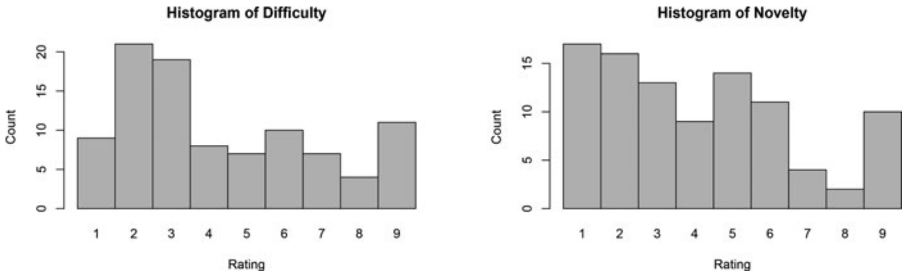


Figure 2. Histograms for the two most balanced metrics, Difficulty and Novelty.

compare the distribution of ratings on the Difficulty and Novelty metrics (for which the IR values are quite close) with those given on the Productive Dispositions and Authenticity metrics (for which the IR values vary greatly), we see much different FDs.

First note that the disparity in Figure 2 and 3 suggests that highly skewed rating distributions do occur in the wild, and thus, the frequency distribution paradox is not simply an esoteric theoretical construct, which never occurs in practice. Furthermore, it appears that when the units to be rated are well-distributed across the rating instrument, IR indices agree with one another and appear to report the level of agreement among raters. As the FD becomes more skewed, IR indices begin to diverge and it is uncertain what they are truly measuring.

Perhaps most importantly, this example clearly shows that if a study involves the use of a collection of metrics (instruments) scoring a *common* set of units, then researchers will almost surely want to be careful in their choice of IR index. The reason is that while it is possible to design a *single* metric so that the units are well-distributed across that metric, this is difficult to do when *many* metrics need to be used on a common set of units. Consequently, the units might be well-distributed on some instruments and skewed on others, exposing the study to the frequency distribution paradox. In the Geometry study, it was simply impractical to ask experts to work and rate hundreds of problems. Thus, the same set of problems had to be used for each metric, and as a result, the full scale range was present in some metrics (e.g., Difficulty and Novelty), but not in others (e.g., Productive Dispositions and Authenticity). Indeed, when the problems were originally selected for inclusion in the study, they were chosen precisely because they appeared to represent a wide swath of Difficulty levels. Meanwhile, no attention was paid to how the problems were distributed across the other 13 metrics. In this setting of multiple metrics and common units, skewed FDs might occur, and researchers will need a robust IR index to avoid drawing false conclusions.

4. The Simulation

A “robust” IR index is one that gives roughly the same result for a fixed level of agreement irrespective of the frequency distribution it must use when correcting for chance agreement. In this section, we run a Monte Carlo simulation to explore the robustness of the five IR indices mentioned above. To do so, we begin by constructing six different FDs. One way to think about an FD is as the type of units the raters are exposed to as filtered through the rating instrument in question. For example, an intelligence-rating scale that reads (Lowest Third, Middle Third, Highest Third) will produce much different results than (Lowest 5%, Next 5%, Top 90%), even when scoring the same individuals. Figure 4 shows the FDs used in this study.

When creating FDs, we wanted to mimic a Likert scale with L levels, and so it was necessary to use a discrete distribution with finite support. All six of the pictured FDs are special cases of the beta-binomial distribution. We used the `dbetabinom.ab` function in T. W. Yee’s VGAM library in R to create each distribution. In brief, the beta-binomial distribution is similar to a standard binomial distribution, but the probability of success changes, following a beta distribution, as described by two shape parameters. To create the six plots in Figure 4, the following pairs of shape parameters were passed to the beta-binomial distribution: FD 1 (0.25, 0.25), FD 2 (1, 1), FD 3 (2, 2), FD 4 (50, 50), FD 5 (25, 50), and FD 6 (5, 50). For a shape pair (a, b), the histogram is symmetric if $a = b$, and the overall graph approximates a binomial distribution for large values of $a = b$.

After creating the FDs, we designed a system to simulate the agreement among raters. To do this, we set up six different agreement distributions (ADs), as seen in Figure 5. Note that ADs are always centered around Rater 1’s score, and show the probability that a new rater will use a certain score given that Rater 1’s score is known. Details about these ADs are provided later.

Together, a particular FD and particular AD can be used to create a table of ratings in the following fashion: First, Rater 1’s

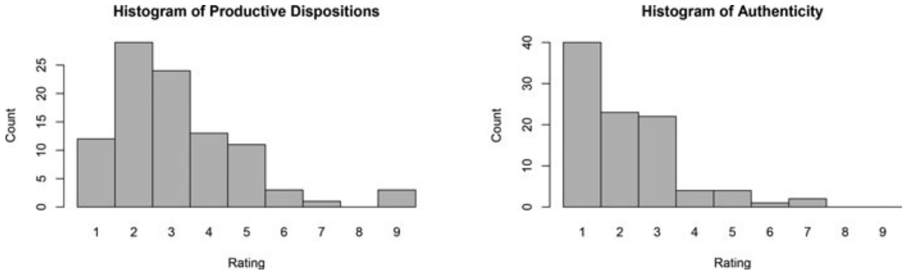
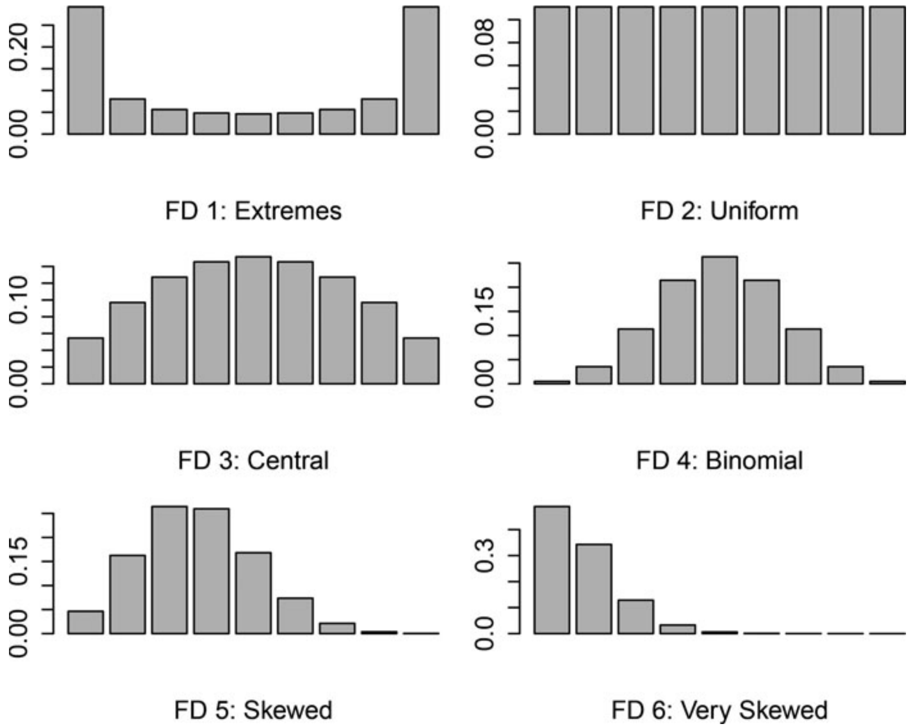


Figure 3. Histograms for the two most skewed metrics, Productive Dispositions and Authenticity.



FD 1: Extremes

FD 2: Uniform

FD 3: Central

FD 4: Binomial

FD 5: Skewed

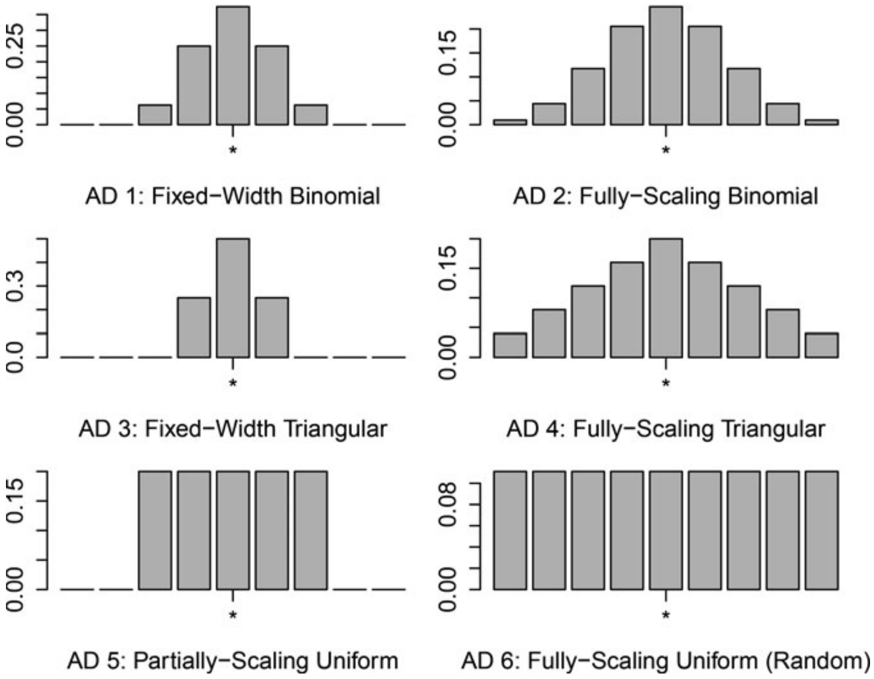
FD 6: Very Skewed

Figure 4. The six frequency distributions (FDs) used in the simulation, assuming a nine-level instrument.

scores for all the units are selected by sampling from the FD; this fills the top row of the ratings table in Figure 6. Next, for a given unit, we center an AD around Rater 1's score (R1) and then sample from this to fill the scores for the other raters; this populates a given column in Figure 6.

There are a few technical challenges in this process. For example, if Rater 1's score is near either end of the scale, then centering the AD about this value will result in scores that exceed the boundaries of the Likert scale. In this case, values outside the span 1 to L were dropped, and the probabilities for values in the 1 to L span were rescaled to have a sum of 1. Figure 7 shows what AD 2 looks like for each of the possible values of R1 assuming a nine-level Likert scale.

One feature of this simulation that has yet to be discussed is that it is designed to work with rating instruments of varying sizes. While the narrative so far has focused on a nine-level Likert scale, we ultimately also wish to understand how the number of levels, L, in a scale affects IR calculations. Changing the



AD 1: Fixed−Width Binomial

AD 2: Fully−Scaling Binomial

AD 3: Fixed−Width Triangular

AD 4: Fully−Scaling Triangular

AD 5: Partially−Scaling Uniform

AD 6: Fully−Scaling Uniform (Random)

Figure 5. The six agreement distributions (ADs) used in the simulation where Rater 1's score is denoted by ⋆. The case of $L = 9$ is shown assuming Rater 1's score is a 5.

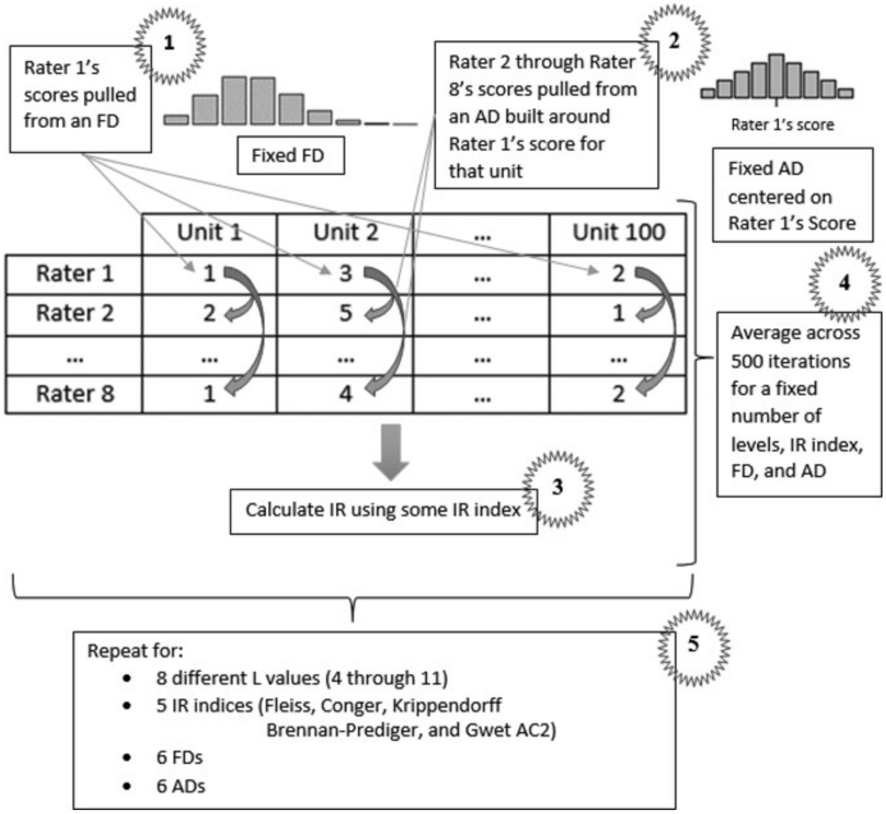

1
Rater 1's scores pulled from an FD
Fixed FD
2
Rater 2 through Rater 8's scores pulled from an AD built around Rater 1's score for that unit
Rater 1's score
Fixed AD centered on Rater 1's Score
Unit 1
Unit 2
...
Unit 100
Rater 1
Rater 2
...
Rater 8
1
3
2
2
5
1
1
4
2
4
Average across 500 iterations for a fixed number of levels, IR index, FD, and AD
Calculate IR using some IR index
3
5
Repeat for:
8 different L values (4 through 11)
5 IR indices (Fleiss, Conger, Krippendorff Brennan-Prediger, and Gwet AC2)
6 FDs
6 ADs

Figure 6. Overview of the simulation.

number of levels in the scale offers no challenge when setting up the FDs: the beta-binomial distribution requires the size of the rating instrument when it is created, and hence, scales based on L. ADs, in contrast, might or might not scale based on the size of the rating instrument. Indeed, it is unclear how a rater's agreement changes based on the size and granularity of the rating scale. To account for the different cases that could occur in practice, we include ADs that maintain a fixed width (ADs 1 and 3), an AD that scales to fit part of the range of scores (AD 5), and three ADs that fully scale to the range 1 through L (ADs 2, 4, and 6). We also use a variety of possible shapes for the ADs to simulate the different types of agreement that might appear in practice.

More specifically, AD 1 is set up as a binomial distribution with $n = 4$ and $p = \frac{1}{2}$. This creates a distribution with five points of support (0 to 4, inclusive), which can be centered at Rater 1's score of choice. AD 2 is also a binomial distribution, but it scales based on the size of the rating instrument. Thus, it uses $n = L$ and $p = \frac{1}{2}$. One difficulty that arises in this case is that if L is odd, then the finite support of $\{0, 1, \ldots, L\}$ has an even number of points, and thus, the graph does not have a unique maximum (which we want to center at R1). In this

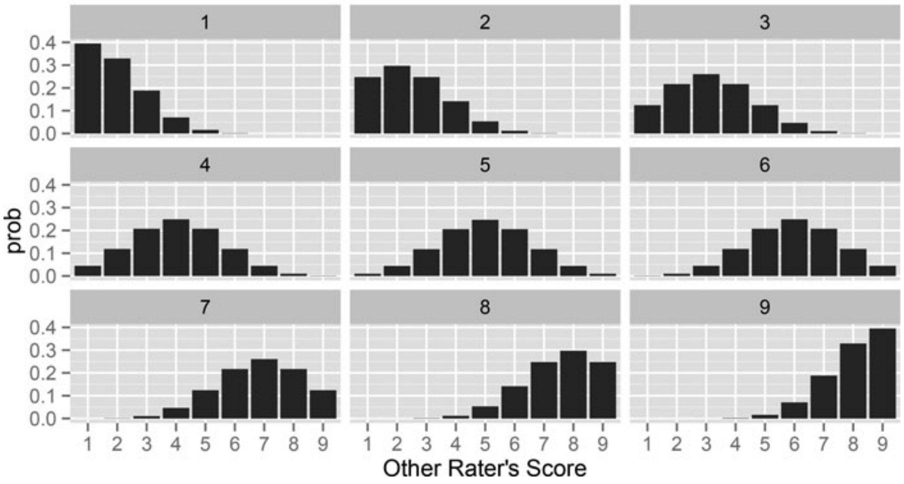


Figure 7. The actual nature of AD 2 depending on Rater 1's score (the number shown in the strip atop each graph). As Rater 1's score moves along the Likert scale, the AD must move as well, adjusting for the finite nature of the rating scale.

situation (and others that were similar), the support size of AD was increased from L to $L + 1$, so that a unique maximum would occur and could be centered at R1.

AD 3 is similar to AD 1 in that it does not scale as L changes. It is designed as a discrete triangular distribution on $\{-2, -1, 0, 1, 2\}$, where the probabilities at $x = -2$ and $x = 2$ are 0. AD 4 is a scalable version of the triangular distribution with support on $\{-\lfloor\frac{L+2}{2}\rfloor, \ldots, \lfloor\frac{L+2}{2}\rfloor\}$ (again, the probability is set to be 0 at the endpoints). As always, whatever zone of support an AD is initially assigned, it is eventually shifted so that it is centered about Rater 1's score. Finally, ADs 5 and 6 are discrete uniform distributions. AD 5 has support on $\{-\lfloor\frac{L+2}{4}\rfloor, \ldots, \lfloor\frac{L+2}{4}\rfloor\}$, and AD 6 is always the uniform distribution on $\{1, 2, \ldots, L\}$, making it equivalent to random chance: new raters' scores have no relation to Rater 1's scores (again, see Figure 5).

Returning to the simulation, Figure 6 depicts a ratings table with 8 raters scoring 100 hypothetical units. The choice of 8 raters, 100 units, and (as discussed later) 500 iterations ensures that (1) Rater 1's scores are a reasonable approximation of the FD from which they are sampled, (2) Raters 2–8's scores are a reasonable approximation of the AD from which they are sampled, and (3) the overall simulated IR values have a 95% confidence interval within a few hundredths of the calculated values.

After creating a single ratings table for a fixed FD and AD as outlined above, we found the IR using one of the five indices and then repeated the whole process a total of 500 iterations. These 500 IR values were averaged to get a final IR value for a particular IR index, fixed FD, fixed AD, and fixed value of L. This procedure was repeated for all combinations of five IR indices, six FDs, six ADs, and eight different values for L (4 through 11). Note that we used the implementations of the five IR indices outlined in Gwet (2014), and code for these can be found at *www.agreestat.com/r_functions.html*. Given that we interpreted our Likert data at the interval level of measurement—a reasonable assumption per Carifio and Perla (2007)—the standard quadratic penalty function was employed in the IR calculations. That is, a rating difference of 3 (say 2 vs. 5) was penalized at a level of $3^2 = 9$, and a near agreement (say 2 vs. 3) was penalized at a level of $1^2 = 1$.

5. Simulation Results

Two important ideas are studied from the data created in this simulation. First, for a fixed number of levels, we explore how robust the five IR indices are to changes in the frequency distribution (Figure 8). Second, for a fixed FD and AD, we examine if changing the number of levels in the rating scale has an effect on the IRs for each of the five indices (Figure 9 and 10). Given that some ADs adjust to the number of levels, while others do not, we expect the IR values for some FD/AD pairings to be unaffected by changes in L, while others will show changes.

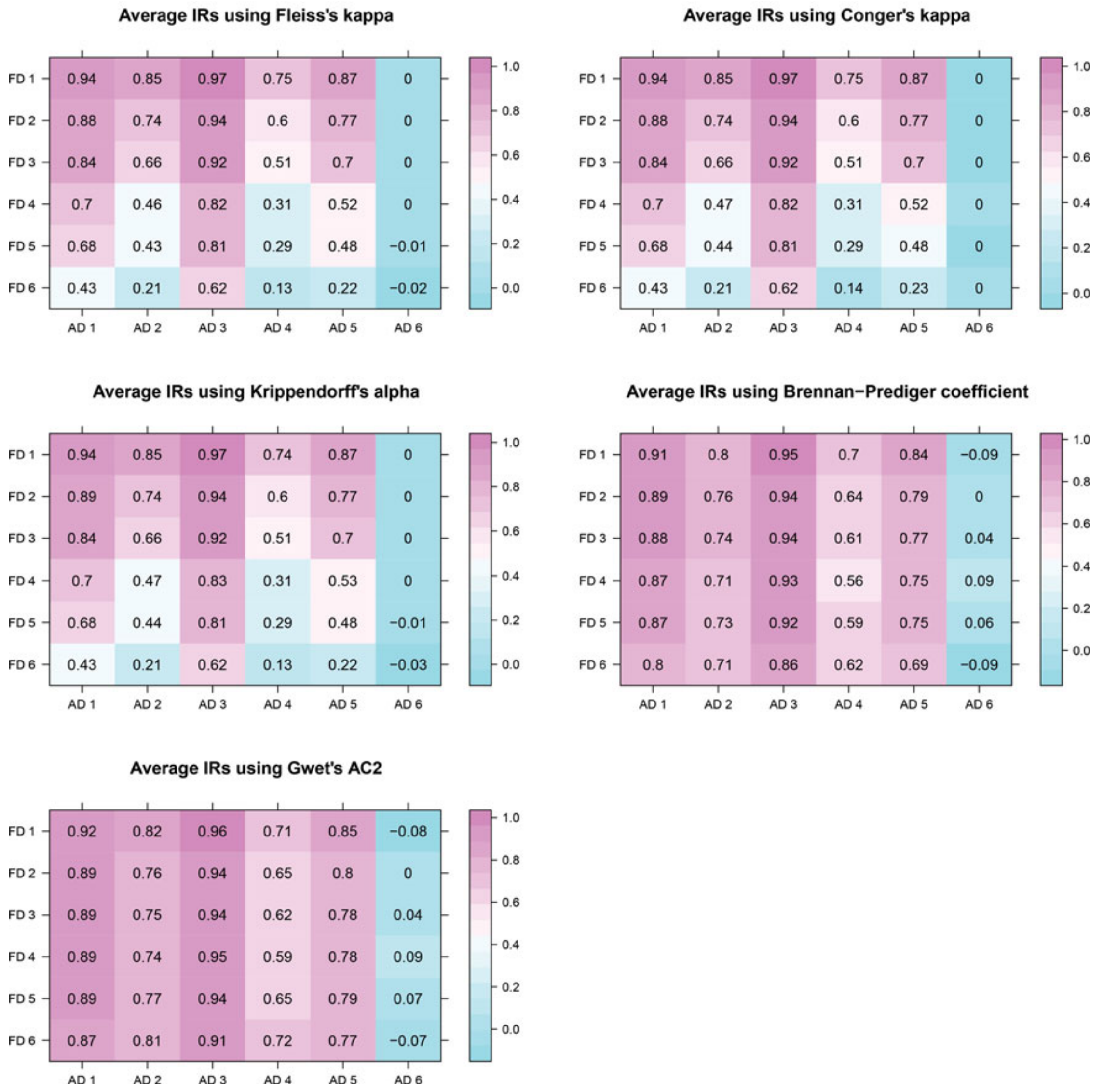
Turning to Figure 8, which is based on a 9-tiered Likert scale ($L = 9$), we can visually see the robustness of a given IR index to various frequency distributions by looking at the colors or values in the columns of the five 6×6 level plots. Here, each column represents a given (fixed) AD, and if an IR index is robust to the FDs it encounters, then the colors going down a column should be nearly identical. The level plots suggest four primary findings:

1. Looking at the overall coloring of the level plots, it appears that Fleiss's kappa, Conger's kappa, and Krippendorff's alpha perform almost identically, and that the Brennan–Prediger coefficient and Gwet's AC2 are also nearly identical. Thus, these five IR indices fall into two distinct groups (hence the grouping terminology introduced earlier).
2. The Group 1 indices each vary greatly for a fixed agreement level (i.e., the scores in a particular column). For example, AD 1 is a strong agreement index, and surveying the average IR results in this column for Krippendorff's alpha reveals values between 0.43 (FD 6, very skewed) and 0.94 (FD 1, extremes). In contrast, the Group 2 indices are far more robust to different distributions: the columns of each have consistent coloring. These findings generalize similar results that have been observed in the case of two raters and two nominal categories (Gwet 2008).
3. Overall, all five indices report that AD 6 has the lowest IR. Given that this AD represents chance agreement—Raters 2–8 do not even consider Rater 1's scores when recording their scores—we would expect these IR values to be zero. Interestingly, the Group 1 indices do a more consistent job of reporting this randomness as zero than those in Group 2. In fact, the Brennan–Prediger coefficient and Gwet's AC2 appear to be less robust to distribution changes when agreement nears chance. For example, the Brennan–Prediger coefficient reports IR values between 0.09 (FD 4, binomial) and −0.09 (FD 1, extremes).
4. The results from Group 1 indices are particularly difficult to trust. For example, in just the six FDs used in this study, Fleiss's kappa routinely gave IR values that were 0.5 apart for a fixed AD. Indeed, the IR spreads for the first five ADs in the Fleiss table are 0.51 (0.94 − 0.43), 0.64, 0.35, 0.62, and 0.65. This finding makes the interpretation of a Fleiss-generated IR value especially difficult. Surprisingly, the only time Fleiss's kappa consistently reports the level of agreement is when there is no agreement to report (AD 6). Furthermore, notice that the range of IR values for ADs 1–5 all overlap. Thus, if we take any value in this overlap, say 0.7, then we see that by simply varying the FD, any of our five ADs can report the value 0.7 *despite the fact* that ADs 1–5 are hard-coded to contain different levels of agreement!

   This problem is not present in the Group 2 indices. As a comparison, the spreads for the first five ADs in the Gwet AC2 table are 0.05 (0.92 − 0.87), 0.08, 0.05, 0.13, and 0.08. So, while variation in the frequency distribution does create variation in the IR, Gwet's index is capable of dramatically lessening this effect.

We now turn to Figure 9 and 10. These figures show the effect of changing the number of levels in the Likert scale that raters use under each of the possible FD/AD combinations and for each IR index. We offer a few observations:

1. As before, the five indices cluster into the two groups previously delineated. If we look at a fixed AD, say AD 2,



Average IRs using Fleiss's kappa

Average IRs using Conger's kappa

Average IRs using Krippendorff's alpha

Average IRs using Brennan−Prediger coefficient

Average IRs using Gwet's AC2

Figure 8. Level plots for all five IR indices for each of the six FDs and ADs assuming a nine-level rating scale of interval data.

and step through the graphs of rows FD 1 through FD 6, we see the two index groupings pull apart from one another; this echoes the results from Figure 8 and suggests the trends seen in that figure occur at all the studied levels (4 through 11), not simply $L = 9$.

2. The graphs for ADs 1–3 (across all FDs) show that increasing the number of levels, while holding all other factors constant, actually results in an increased IR! This finding is quite surprising and immediately suggests that researchers must think about more than the frequency distribution paradox when designing a measurement instrument. Indeed, it appears that the AD, FD, and the number of levels all influence the IR for a given experiment. The reason for this finding is the following: As the number of levels increases (linearly), one must ask how the AD will expand to fit this scaling. If the standard deviation of the AD does not also grow linearly, then the growth of the rating instrument will not match the expansion of the AD. For example, AD 2 is a fully expanding binomial distribution; specifically, it is a binomial distribution with $p = \frac{1}{2}$ and support on $\{0, 1, 2, \ldots, L\}$ (or $L + 1$ if L is odd). The variance of this distribution is $np(1 - p) = \frac{L}{4}$, and thus, the standard deviation is $\frac{\sqrt{L}}{2}$, which is less than linear growth in L. This implies that as L grows larger, AD 2 actually becomes a stronger agreement metric, and hence, the IR should increase.

   There are other factors also at play in the graphs from ADs 1–3 (and other ADs that might occur in practice). Note that an AD does not always look like its idealized shape as seen in Figure 5. Indeed, as Rater 1's scores move to the extreme ends of the scale, the AD (which determines the scores of the other raters) begins to deform, as seen in Figure 7. These deformed versions of a given AD can offer a weaker level of agreement, and hence the IR is, in part, influenced by how often an AD is pulled into its deformed versions. This is certainly a function of the FD, but also of L. If we consider FD 2 (uniform on 1 through L), we see that adding more levels to the scale tends to decrease the likelihood that R1 will take an

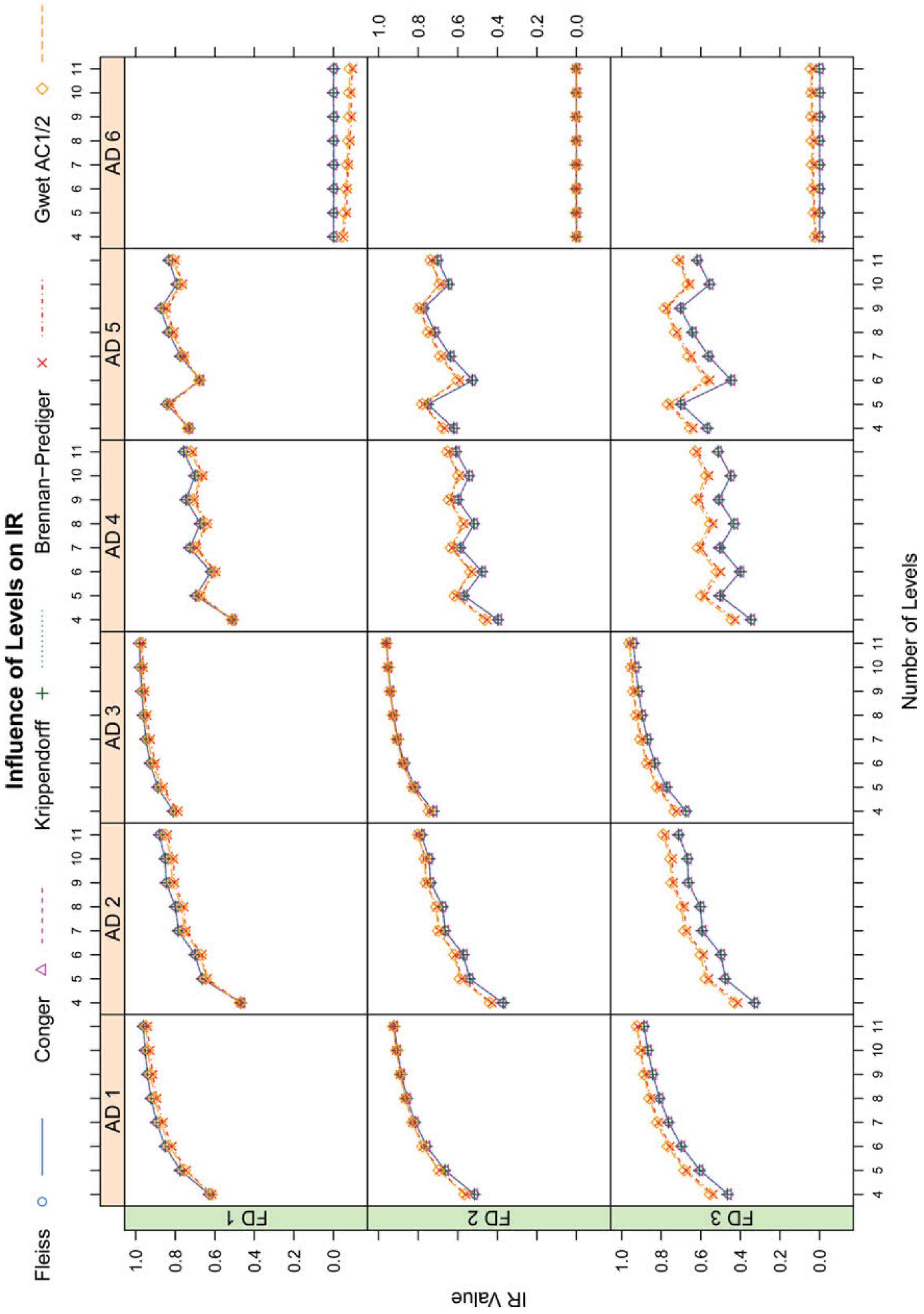


Figure 9. Plots comparing IR values to number of levels in FDs 1–3 with ADs 1–6.

extreme value, and hence that any AD will need to invoke its deformed versions.

3. In contrast with ADs 1–3, ADs 4 and 5 both have standard deviations that grow (roughly) linearly in L. This can be proven directly from the functional definition of variance or intuitively, from the shapes of the ADs themselves. Because the growth of the scale and the growth of standard deviation of the AD match, the graphs for these ADs are roughly horizontal. The observed kinks in the graphs occur because the growth in the

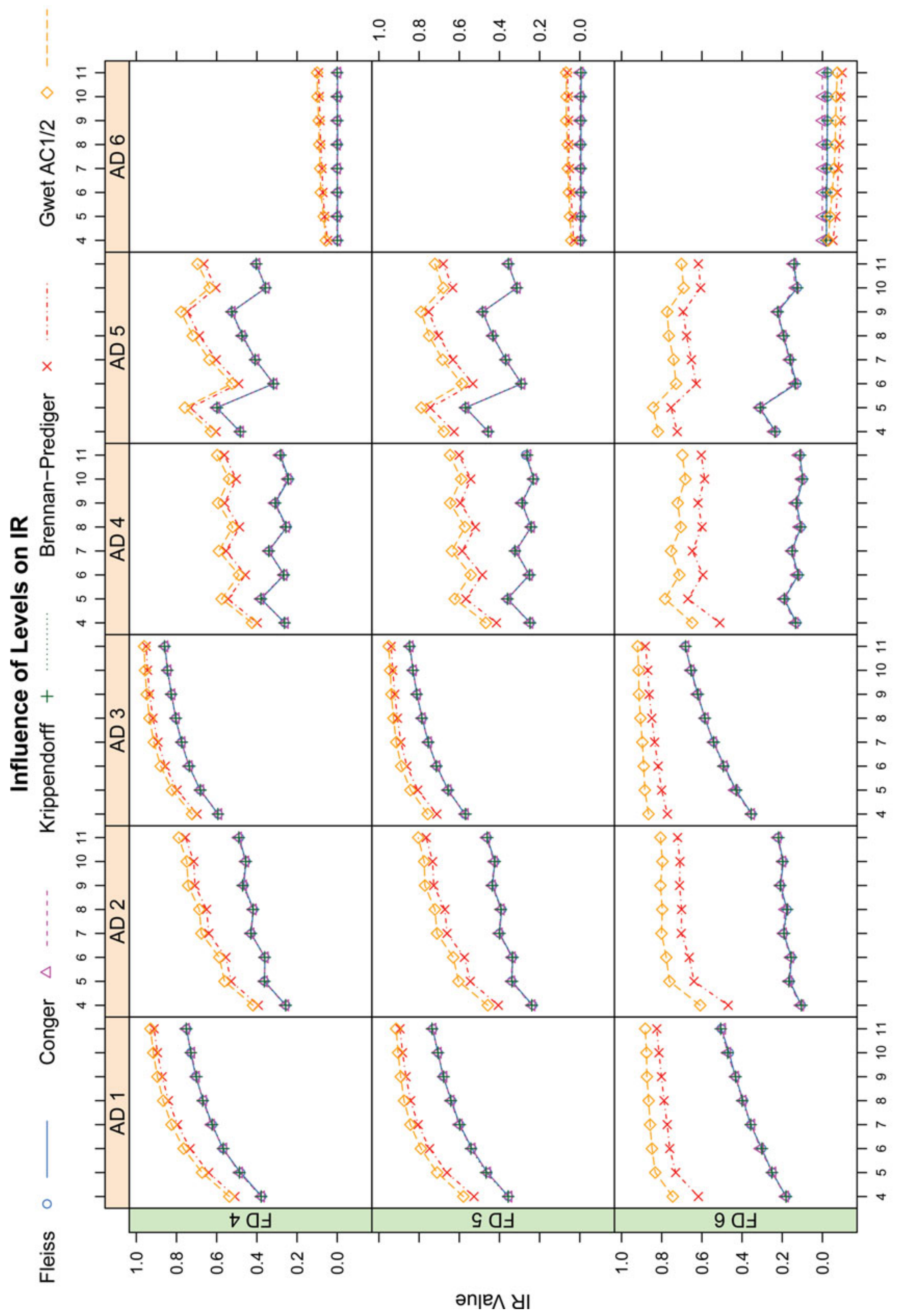


Figure 10. Plots comparing IR values to number of levels in FDs 4–6 with ADs 1–6.

standard deviation is not precisely linear (more accurately, it is a step function). For example, AD 4 is a fully scaling triangular distribution designed to have support on $\{-\lfloor\frac{L+2}{2}\rfloor, \ldots, \lfloor\frac{L+2}{2}\rfloor\}$. As L increases, the AD rescales *every other integer*; the floor function causes this and was included because of the discrete nature of the support. Thus, in cell (FD 4, AD 4) of Figure 10, we see a pattern in the graph that repeats every two units. Note that the graph goes up from $L = 4$ to $L = 5$ (and 6 to 7, 8 to 9, etc.) because both of these scales are using the same AD, but the increased scale size leads to less appearance of the scale-end deformation effect mentioned above, and hence, a higher IR. As L becomes larger, the scale-end effect weakens, and the zig-zag pattern reduces in intensity.

The case of AD 5 makes sense as well. This distribution is set up to have support on $\{-\lfloor \frac{L+2}{4} \rfloor, \ldots, \lfloor \frac{L+2}{4} \rfloor\}$. Because of the floor function, the AD only expands every four values of L, and hence, we see jumps in the graphs of the AD 5 column every four steps. Because of the precise way the support is set up, the L values 4 through 11 show the end of a four-cycle, a complete four-cycle, and the start of another four-cycle. Note that when the AD does not scale (e.g., in the middle of a four-cycle), the IR grows because the scale is expanding and the ends are less likely to be used.

6. Discussion

What is to be made of the above results? To start, our findings strongly suggest that Group 1 IR indices are difficult to trust. Because of how they correct for chance agreement, Fleiss's kappa, Conger's kappa, and Krippendorff's alpha all suffer from the same problem: they are profoundly influenced by the frequency distribution of the units being scored. Group 2 indices (the Brennan–Prediger coefficient and Gwet's AC2) represent a hopeful step forward in this regard. As seen in Figure 8, they perform quite consistently across various FDs for a fixed AD. This issue is important both in theory and in practice. As the Geometry problems example shows, the need for a robust IR index is especially high when a set of instruments are applied to a common set of units. In such a case, the ratings distributions for each instrument are unlikely to all be well-balanced, and Group 1 indices underreport IR values in skewed frequency distributions. Indeed, Group 1 indices are so sensitive to distributional variation that one can largely predict the frequency distribution for a given set of units/instrument simply by examining how differently the instrument performs on Group 1 and Group 2 indices, as we saw in the discussion of Figure 2 and 3.

This does not, however, suggest that Group 2 indices are the IR panacea. At an important, core level, it appears that many factors affect the final, single number that is produced by an IR index: the approach the index uses to correct for chance agreement, the AD of the raters, the FD of the units, and the number of levels in the rating instrument. Many of these factors are intertwined, and not even Group 2 indices are capable of disentangling these threads. As seen in Figure 9 and 10, the relationship of L, an FD, and an AD is particularly thorny. If, as in the case of ADs 1–3, increasing L linearly results in less-than-linear increase in the standard deviation of the AD, then the IR value will increase as the instrument grows in size. What makes the structuring of IR studies particularly frustrating is that both the "true" FD of the scored units and the AD of the raters are invisible to the researcher (unless he or she back-derives them at the end of the rating process). Hence, it is impossible to know at the start of an experiment if changing L will alter the IR. In some cases (e.g., cell AD 1, FD 4 of Figure 10), the choice of L dramatically influences the IR; in other cases (e.g., cell AD 4, FD 6 of Figure 10), the effects are minimal. This issue remains a topic for future research.

Supplementary Material

The supplementary materials contains the R code for the simulation described in the article.

References

Baethge, C., Franklin, J., and Mertens, S. (2013), "Substantial Agreement of Referee Recommendations at a General Medical Journal—A Peer Review Evaluation at Deutsches Ärzteblatt International," *PloS One*, 8, 1–7. [375]

Brennan, R. L., and Prediger, D. J. (1981), "Coefficient Kappa: Some Uses, Misuses, and Alternatives," *Educational and Psychological Measurement*, 41, 687–699. [375]

Carifio, J., and Perla, R. J. (2007), "Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends About Likert Scales and Likert Response Formats and Their Antidotes," *Journal of Social Sciences*, 3, 106–116. [379]

Cicchetti, D. V., and Feinstein, A. R. (1990), "High Agreement but Low Kappa: II. Resolving the Paradoxes," *Journal of Clinical Epidemiology*, 43, 551–558. [374]

Conger, A. J. (1980), "Integration and Generalization of Kappas for Multiple Raters," *Psychological Bulletin*, 88, 322–328. [374,375]

Feinstein, A. R., and Cicchetti, D. V. (1990), "High Agreement but Low Kappa: I. The Problems of Two Paradoxes," *Journal of Clinical Epidemiology*, 43, 543–549. [374]

Fleiss, J. L. (1971), "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, 76, 378–382. [374,375]

Guggenmoos-Holzmann, I. (1993), "How Reliable are Chance-Corrected Measures of Agreement?," *Statistics in Medicine*, 12, 2191–2205. [374]

Gutiérrez, R. (2007), "(Re)Defining Equity: The Importance of a Critical Perspective," in *Improving Access to Mathematics: Diversity and Equity in the Classroom*, eds. N. S. Nasir and P. Cobb, New York, NY: Teachers College Press, pp. 37–50. [375]

——— (2013), "The Sociopolitical Turn in Mathematics Education," *Journal for Research in Mathematics Education*, 44, 37–68. [375]

Gwet, K. L. (2002), "Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity," *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2, 1–9. [374]

——— (2008), "Computing Inter-Rater Reliability and its Variance in the Presence of High Agreement," *British Journal of Mathematical and Statistical Psychology*, 61, 29–48. [374,379]

——— (2014), *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, Gaithersburg, MD: Advanced Analytics, LLC. [373,374,375,379]

Janson, H., and Olsson, U. (2001), "A Measure of Agreement for Interval or Nominal Multivariate Observations," *Educational and Psychological Measurement*, 61, 277–289. [373]

Krippendorff, K. (2012), *Content Analysis: An Introduction to its Methodology*, Thousand Oaks, CA: Sage Publications. [374,375]

Lang, A. T., Grooms, L. P., Sturm, M., Walsh, M., Koch, T., and O'Brien, S. H. (2014), "The Accuracy of a Parent-Administered Bleeding Assessment Tool in a Pediatric Hematology Clinic," *Haemophilia*, 20, 807–813. [375]

LeBreton, J. M, and Senter, J. L. (2008), "Answers to 20 Questions About Interrater Reliability and Interrater Agreement," *Organizational Research Methods*, 11, 815–852. [373]

National Council of Teachers of Mathematics (1991), *Professional Standards for Teaching Mathematics*, Reston, VA: Author. [375]

——— (2000), *Principles and Standards for School Mathematics*, Reston, VA: Author. [375]

National Governors Association Center for Best Practices, Council of Chief State School Officers (2012), *Common Core State Standards for Mathematics*, Washington, DC: Author. [375]

Nelson, J. C., and Pepe, M. S. (2000), "Statistical Description of Interrater Variability in Ordinal Ratings," *Statistical Methods in Medical Research*, 9, 475–496. [374]

R Core Development Team (2014), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org/*. [374]

Wongpakaran, N., Wongpakaran, T., Wedding, D., and Gwet, K. L. (2013), "A Comparison of Cohen's Kappa and Gwet's AC1 When Calculating Inter-Rater Reliability Coefficients: A Study Conducted with Personality Disorder Samples," *BMC Medical Research Methodology*, 13, 1–7. [375]