Practical Assessment, Research, and Evaluation

Volume 9 *Volume 9, 2004* Article 4

2004

A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability

Steven E. Stemler

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 4, March, 2004 ISSN=1531-7714

A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability

Steven E. Stemler
Yale University PACE Center

This article argues that the general practice of describing interrater reliability as a single, unified concept is at best imprecise, and at worst potentially misleading. Rather than representing a single concept, different statistical methods for computing interrater reliability can be more accurately classified into one of three categories based upon the underlying goals of analysis. The three general categories introduced and described in this paper are: 1) consensus estimates, 2) consistency estimates, and 3) measurement estimates. The assumptions, interpretation, advantages, and disadvantages of estimates from each of these three categories are discussed, along with several popular methods of computing interrater reliability coefficients that fall under the umbrella of consensus, consistency, and measurement estimates. Researchers and practitioners should be aware that different approaches to estimating interrater reliability carry with them different implications for how ratings across multiple judges should be summarized, which may impact the validity of subsequent study results.

Many educational and psychological studies require the use of independent judges, or raters, in order to quantify some aspect of behavior. For example, judges may be used to score open-response items on a standardized test, to rate the performance of expert athletes at a sporting event, or to empirically test the viability of a new scoring rubric. Judges are most often used when behaviors of interest cannot be objectively scored in a simple right/wrong sense, but instead require some rating of the degree to which observed behaviors represent particular levels of a construct of interest (e.g., athletic excellence, history competence). The task of judging behavior invites some degree of subjectivity in that the rating given will depend upon the judge's interpretation of the construct. One strategy for reducing subjectivity is to develop scoring rubrics (Mertler, 2001; Moskal & Leydens, 2000; Tierney & Simon, 2004). The purpose of training judges how to interpret a scoring rubric and consistently apply the levels of a rating scale associated with the rubric is to impose some level of objectivity onto the rating scale.

Across all situations involving judges, it is important to estimate the degree of interrater reliability, as this value has important implications for the validity of the study results. If two judges cannot be shown to reliably rate individuals based on observed behaviors, then any subsequent analyses of the ratings given by those judges will yield spurious results. Furthermore, interrater reliability must be demonstrated anew for each new study, even if the study is using a scoring rubric or instrument that has been shown to have high interrater reliability in the past. Interrater reliability refers to the level of agreement between a particular set of judges on a particular instrument at a particular time. Thus, interrater reliability is a property of the testing situation, and not of the instrument itself.

A number of central textbooks in the field of educational and psychological measurement and statistics (e.g., Cohen, Cohen, West, & Aiken, 2003; Crocker & Algina, 1986; Hopkins, 1998; Winer, 1962) describe interrater reliability as if it were a single, unitary concept. In this article, I will argue that the widespread practice of describing interrater reliability as a single, universal concept is at best imprecise, and at worst potentially misleading. Instead, researchers and practitioners should begin to use more precise language to indicate the specific type of interrater reliability being discussed. Although there are numerous statistical methods for computing interrater reliability, I propose that those most commonly reported in the literature generally can be classified into one of three categories: 1) consensus estimates, 2) consistency estimates, or 3) measurement estimates. Reporting a single interrater reliability statistic without discussing the category of interrater reliability the statistic represents is problematic because the three different categories carry with them different implications for how data from multiple judges should be summarized most appropriately.

This article will describe the assumptions, interpretation, advantages, and disadvantages of consensus, consistency, and measurement estimates of interrater reliability. In addition, the article will also discuss some popular statistical methods for computing interrater reliability that fall under the umbrella of each of these three major categories.

Consensus Estimates

General Description

Consensus estimates of interrater reliability are based on the assumption that reasonable observers should be able to come to exact agreement about how to apply the various levels of a scoring rubric to the observed behaviors. If two judges come to exact agreement on how to use the rating scale to score behaviors, then the two judges may be said to share a common interpretation of the construct.

Consensus estimates tend to be the most useful when data are nominal in nature and different levels of the rating scale represent qualitatively different ideas. For example, Stemler and Bebell (1999) conducted a study aimed at detecting the various purposes of schooling articulated in school mission statements. Judges were given a scoring rubric that listed 10 possible thematic categories under which the main idea of each mission statement could be classified (e.g., social development, cognitive development, civic development). Judges then read a series of mission statements and attempted to classify each sampling unit according to the major purpose of schooling articulated. If both judges consistently rated the dominant theme of the mission statement as representing elements of citizenship, then they were said to have communicated with each other in a meaningful way because they had both classified the statement in exactly the same way. If one judge classified the major theme as social development, and the other judge classified the major theme as citizenship, then a breakdown in shared understanding occurred. In that case, the judges were not coming to a consensus on how to apply the levels of the scoring rubric.

Consensus estimates also can be useful when different levels of the rating scale are assumed to represent a linear continuum of the construct, but are ordinal in nature (e.g., a Likert scale). In that case, the judges must come to exact agreement about the quantitative levels of the construct under investigation, rather than attempting to evaluate qualitative differences in scoring categories, as in the previous example.

If judges can be trained to the point where they agree on how to interpret a rating scale, then scores given by the two judges may be treated as equivalent. Thus, the remaining work of rating subsequent items can be split between the judges without both judges having to rate all items. Furthermore, the summary scores may be calculated by simply taking the score from one of the judges or by averaging the scores given by all of the judges since high interrater reliability indicates that the judges agree about how to apply the rating scale. A typical guideline found in the literature for evaluating the quality of interrater reliability based upon consensus estimates is that they should be 70% or greater.

Popular Methods for Computing Consensus Estimates

Perhaps the most popular method for computing a consensus estimate of interrater reliability is through the use of the simple percent-agreement figure. Percent agreement is calculated by adding up the number of cases that received the same rating by both judges and dividing that number by the total number of cases rated by the two judges. The percent agreement statistic has several advantages. For example, it has a strong intuitive appeal, it is easy to calculate, and it is easy to explain.

The statistic also has some distinct disadvantages, however. For example, if the behavior of interest has a low incidence of occurrence in the population, then it is possible to get artificially inflated percent-agreement figures simply because most of the values fall under one category of the rating scale (Hayes & Hatch, 1999). Another disadvantage to using the simple percent-agreement figure is that it is often time consuming and labor intensive to train judges to the point of exact agreement.

Several other kinds of consensus estimates are also commonly reported in the literature. One popular modification of the percent-agreement figure found in the testing literature involves broadening the definition of agreement by including the adjacent scoring categories on the rating scale. For example, on a rating scale with levels ranging from 1–7, judges would not need to come to exact agreement about the ratings they assign. So long as the ratings did not differ by more than one point above or below the other judge, then the two judges would be said to have reached consensus. This approach is advantageous in that it relaxes the strict criterion that the judges agree exactly. On the other hand, percent agreement using adjacent categories can lead to inflated estimates of interrater reliability if there are only a limited number of categories to choose from (e.g., a 1–4 scale). If the rating scale has a limited number of points, then nearly all points will be adjacent, and it would be surprising to find agreement lower than 90%. The technique of using adjacent categories results in a situation where the percent agreement at the extreme ends of the rating scale is almost always lower than in the middle.

A third consensus estimate of interrater reliability is Cohen's kappa statistic (Cohen, 1960, 1968). Cohen's kappa was designed to estimate the degree of consensus between two judges after correcting the percent-agreement figure for the amount of agreement that could be expected by chance alone based upon the values of the marginal distributions (see Stemler, 2001 for a practical example with calculation). The interpretation of the kappa statistic is slightly different than the interpretation of the percent-agreement figure. A value of zero on kappa does not indicate that the two judges did not agree at all; rather, it indicates that the two judges did not agree with each other any more than would be predicted by chance alone. Consequently, it is possible to have negative values of kappa if judges agree less often than chance would predict. Landis and Koch (1977) suggest that kappa values from 0.41–0.60 are moderate, and that values above 0.60 are substantial. Kappa is a highly useful statistic when one is concerned that the percent-agreement statistic may be artificially inflated due to the fact that most observations fall into a single category.

The major disadvantage to the kappa coefficient is that it is can be somewhat difficult to interpret. Uebersax (1987) has

DOI: https://doi.org/10.7275/96jp-xz07

noted that one major problem is that values of kappa may differ depending upon the proportion of respondents falling into each category of a rating scale. Thus, kappa values for different items or from different studies cannot be meaningfully compared unless the base rates are identical. Consequently, although the statistic gives some indication as to whether the agreement is better than that predicted by chance alone, it is difficult to apply the kinds of rules of thumb suggested by Landis and Koch for interpreting kappa across different circumstances.

For a review of several other less frequently used methods for computing interrater reliability, some of which fall within the general category of consensus estimates of interrater reliability (e.g., Jaccard's J, The G-Index), see Barrett (2001).

Advantages of Consensus Estimates

One particular advantage of the consensus approach to estimating interrater reliability is that the calculations are easily done by hand. A second advantage is that the techniques falling within this general category are well suited to dealing with nominal variables whose levels on the rating scale represent qualitatively different categories. A third advantage is that consensus estimates can be useful in diagnosing problems with judges' interpretations of how to apply the rating scale. For example, inspection of the information from a crosstab table may allow the researcher to realize that the judges may be unclear about the rules for when they are supposed to score an item as zero as opposed to when they are supposed to score the item as missing. A visual analysis of the output allows the researcher to go back to the data and clarify the discrepancy or retrain the judges.

Finally, when judges exhibit a high level of consensus, it implies that both judges are essentially providing the same information. One implication of a high consensus estimate of interrater reliability is that both judges need not score all remaining items. For example, if there were 100 tests to be scored after the interrater reliability study was finished, it would be most efficient to ask Judge A to rate exams 1–50 and Judge B to rate exams 51–100, because the two judges have empirically demonstrated that they share a similar meaning for the scoring rubric. In practice, however, it is usually a good idea to build in a 30% overlap between judges even after they have been trained in order to provide evidence that the judges are not drifting from their consensus as they read more items. Such an approach would suggest that Judge A rate items 1–35 by himself (35 items), Judge B should rate items 66–100 by herself (35 items), and both judges should rate students 36–65 (30 items).

Disadvantages of Consensus Estimates

One disadvantage of consensus estimates is that interrater reliability statistics must be computed separately for each item and for each pair of judges. Consequently, when reporting consensus-based interrater reliability estimates, one should report the minimum, maximum, and median estimates for all items and for all pairs of judges.

A second disadvantage is the amount of time and energy it takes to train judges to come to exact agreement is often substantial. In some instances it may not be important for the judges to come to exact agreement: for example, if the exact application of the levels of the scoring rubric is not important, but rather a means to the end of getting a summary score for each respondent.

Third, as Linacre (2002) has noted, training judges to a point of forced consensus may actually reduce the statistical independence of the ratings, and threaten the validity of the resulting scores.

Finally, consensus estimates can be overly conservative if two judges exhibit systematic differences in the way that they use the scoring rubric but simply cannot be trained to come to a consensus. As we will see in the next section, it is possible to have a low consensus estimate of interrater reliability while at the same time having a high consistency estimate, and vice versa. Consequently, sole reliance on consensus estimates of interrater reliability might lead researchers to conclude that "interrater reliability is low," when it may be more precisely stated that the *consensus estimate* of interrater reliability is low.

Consistency Estimates

General Description

Consistency estimates of interrater reliability are based upon the assumption that it is not really necessary for two judges to share a common meaning of the rating scale, so long as each judge is consistent in classifying the phenomenon according to his or her own definition of the scale. For example, if Judge A assigns a score of 3 to a certain group of essays, and Judge B assigns a score of 1 to that same group of essays, the two judges have not come to a consensus about how to apply the rating scale categories, but the difference in how they apply the rating scale categories is predictable.

Consistency approaches to estimating interrater reliability are most useful when the data are continuous in nature, although the technique can be applied to categorical data if the rating scale categories are thought to represent an underlying continuum along a unidimensional construct. Values greater than 0.70 are typically acceptable for consistency estimates of interrater reliability (Barrett, 2001).

It is important to recognize that although consistency estimates may be high, the means and medians of the different judges may be very different. Thus, if one judge consistently gives scores that are two points lower on the rating scale

than does a second judge, the scores will ultimately need to be corrected for this difference in judge severity if the final scores are to be summarized or subjected to further analyses.

Popular Methods for Computing Consistency Estimates

Perhaps the most popular statistic for calculating the degree of consistency between judges is the Pearson correlation coefficient. The Pearson correlation coefficient can be computed by hand (Glass & Hopkins, 1996), or can easily be computed using most statistical packages. One beneficial feature of the Pearson correlation coefficient is that the scores on the rating scale can be continuous in nature (e.g., they can take on partial values such as 1.5). Like the percent-agreement statistic, the Pearson correlation coefficients can be calculated only for one pair of judges at a time and for one item at a time.

A potential limitation of the Pearson correlation coefficient is that it assumes that the data underlying the rating scale are normally distributed. Consequently, if the data from the rating scale tend to be skewed toward one end of the distribution, this will attenuate the upper limit of the correlation coefficient that can be observed.

Another popular consistency estimate of interrater reliability is Spearmans's rank coefficient. The Spearman rank coefficient provides an approximation of the Pearson correlation coefficient, but may be used in circumstances where the data under investigation are not normally distributed. For example, rather than using a continuous rating scale, each judge may rank order the essays that he or she has scored from best to worst. In this case, then, since both ratings being correlated are in the form of rankings, a correlation coefficient can be computed that is governed by the number of pairs of ratings (Glass & Hopkins, 1996). The major disadvantage to Spearman's rank coefficient is that it requires both judges to rate all cases.

In situations where multiple judges are used, another approach to computing a consistency estimate of interrater reliability would be to compute Cronbach's alpha coefficient (Crocker & Algina, 1986). Cronbach's alpha coefficient is a measure of internal consistency reliability and is useful for understanding the extent to which the ratings from a group of judges hold together to measure a common dimension. If the Cronbach's alpha estimate among the judges is low, then this implies that the majority of the variance in the total composite score is really due to error variance, and not true score variance (Crocker & Algina, 1986).

The major advantage of using Cronbach's alpha comes from its capacity to yield a single consistency estimate of interrater reliability across multiple judges. The major disadvantage of the method is that each judge must give a rating on every case, or else the alpha will only be computed on the bases of a subset of the data. In other words, if just one rater fails to score a particular individual, that individual will be left out of the analysis. In addition, as Barrett (2001) has noted, "…because of this 'averaging' of ratings, we reduce the variability of the judges ratings such that when we average all judges ratings, we effectively remove all the error variance for judges." (p. 7).

For further information regarding various consistency estimates of interrater reliability, see Bock, Brennan, and Muraki (2002), Burke and Dunlap (2002), LeBreton, Burgess, Kaiser, Atchley, and James (2003), and Uebersax (2002).

Advantages of Consistency Estimates

There are three major advantages to using consistency estimates of interrater reliability. First, the approach places less stringent demands upon the judges in that they need not be trained to come to exact agreement with one another so long as each judge is consistent within his or her own definition of the rating scale (i.e., exhibits high intrarater reliability). It is sometimes the case that the exact application of the levels of the scoring rubric is not important in itself. Instead, the scoring rubric is a means to the end of creating scores for each participant that can be summarized in a meaningful way. If summarization is the goal, then what is most important is that each judge apply the rating scale consistently within his or her own definition of the rating scale, regardless of whether the two judges exhibit exact agreement. Consistency estimates allow for the detection of systematic differences between judges, which may then be adjusted statistically. For example, if Judge A consistently gives scores that are 2 points lower than Judge B does, then adding 2 extra points to the exams of all students who were scored by Judge A would provide an equitable adjustment to the raw scores.

A second advantage of consistency estimates is that certain methods within this category (e.g., Cronbach's alpha) allow for an overall estimate of consistency among multiple judges. In addition, consistency estimates readily handle continuous data.

Disadvantage of Consistency Estimates

One disadvantage of consistency estimates is that if the construct under investigation has some objective meaning, then it may not be desirable for the two judges to "agree to disagree." Instead, it may be important for the judges to come to an exact agreement of the scores that they are generating.

A second disadvantage of consistency estimates is that judges may not only differ systematically in the raw scores they apply, but also in the number of rating scale categories they use. In that case, a mean adjustment for a severe judge may provide a partial solution, but the two judges may also differ on the variability in scores they give. Thus, a mean

adjustment alone will not effectively correct for this difference.

A third disadvantage of consistency estimates is that they are highly sensitive to the distribution of the observed data. In other words, if most of the ratings fall into one or two categories, the correlation coefficient will necessarily be deflated due to restricted variability. Consequently, a reliance on the consistency estimate alone may lead the researcher to falsely conclude that interrater reliability was poor without specifying more precisely that the *consistency estimate* of interrater reliability was poor and providing an appropriate rationale.

Measurement Estimates

General Description

Measurement estimates are based upon the assumption that one should use all of the information available from all judges (including discrepant ratings) when attempting to create a summary score for each respondent. In other words, each judge is seen as providing some unique information that is useful in generating a summary score for a person. As Linacre (2002) has noted, "It is the accumulation of information, not the ratings themselves, that is decisive" (p. 858). Consequently, under the measurement approach, it is not necessary for two judges to come to a consensus on how to apply a scoring rubric because differences in judge severity can be estimated and accounted for in the creation of each participant's final score.

Measurement estimates are best used when different levels of the rating scale are intended to represent different levels of an underlying unidimensional construct (e.g., mathematical competence). They are also useful in circumstances where multiple judges are involved in administering ratings, and it is impossible for all judges to rate all items.

Measurement estimates allow for the creation of a summary score for each participant that represents that participant's score on the underlying factor of interest, taking into account the extent to which each judge influences the score.

Popular Methods of Computing Measurement Estimates

One popular measurement estimate of interrater reliability is computed using the factor analytic technique of principal components analysis (Harman, 1967). Using this method, multiple judges may rate a set of participants. The judge's scores are then subjected to a principal components analysis to determine the amount of shared variance in the ratings that could be accounted for by the first principal component. The percentage of variance that is explainable by the first principal component gives some indication of the extent to which the multiple judges are reaching agreement. If the shared variance is high (e.g., greater than 60%), then this gives some indication that the judges are rating a common construct.

Once interrater reliability has been established in this way, each participant may then receive a single summary score corresponding to his or her loading on the first principal component underlying the set of ratings. This score can be computed automatically by most statistical packages. The advantage of this approach is that it assigns a summary score for each participant that is based only on the relevance of the strongest dimension underlying the data. The disadvantage to the approach is that it assumes that ratings are assigned without error by the judges.

Another popular method for computing a measurement estimate of interrater reliability has been through the use of generalizability theory (Shavelson & Webb, 1991). Linacre (1994) has noted that the goal of a generalizability study is to determine,

> *…the error variance associated with each judge's ratings, so that correction can be made to ratings awarded by a judge when he is the only one to rate an examinee. For this to be useful, examinees must be regarded as randomly sampled from some population of examinees which means that there is no way to correct an individual examinee's score for judge behavior, in a way which would be helpful to an examining board. This approach, however, was developed for use in contexts in which only estimates of population parameters are of interest to researchers (p. 29).*

A third measurement approaches to estimating interrater reliability is through the use of the many-facets Rasch model (Linacre, 1994). Recent advances in the field of measurement have led to an extension of the standard Rasch measurement model (Rasch, 1960/1980; Wright & Stone, 1979). This new, extended model, known as the many-facets Rasch model, allows judge severity to be derived using the same scale (i.e., the logit scale) as person ability and item difficulty. In other words, rather than simply assuming that a score of 3 from Judge A is equally difficult for a participant to achieve as a score of 3 from Judge B, the equivalence of the ratings between judges can be empirically determined. Thus, it could be the case that a score of 3 from Judge A is really closer to a score of 5 from Judge B (i.e., Judge A is a more severe rater). Using a many-facets analysis, each essay item or behavior that was rated can be directly compared.

In addition, the difficulty of each item, as well as the severity of all judges who rated the items, can also be directly compared. For example, if a history exam included five essay questions and each of the essay questions was rated by three judges (two unique judges per item, and one judge who scored all items), the facets approach would allow the researcher to directly compare the severity of a judge who rated only Item 1 with the severity of a judge who rated only

Item 4. Each of the 11 judges (2 unique judges per item + 1 judge who rated all items = 5*2 + 1 = 11) could be directly compared. The mathematical representation of the many-facets Rasch model is fully described in Linacre (1994).

Finally, in addition to providing information that allows for the evaluation of the severity of each judge in relation to all other judges, the facets approach also allows one to evaluate the extent to which each of the individual judges is using the scoring rubric in a manner that is internally consistent (i.e., an estimate of intrarater reliability). In other words, even if judges differ in their own definition of how they use the scale, the fit statistics will indicate the extent to which a given judge is faithful to his or her own definition of the scale categories across items and people.

Advantages of Measurement Estimates

There are several advantages to estimating interrater reliability using the measurement approach. First, measurement estimates can take into account errors at the level of each judge or for groups of judges. Consequently, the summary scores generated from measurement estimates of interrater reliability tend to more accurately represent the underlying construct of interest than do the simple raw score ratings from the judges.

Second, measurement estimates effectively handle ratings from multiple judges by simultaneously computing estimates across all of the items that were rated, as opposed to calculating estimates separately for each item and each pair of judges.

Third, measurement estimates have the distinct advantage of not requiring all judges to rate all items in order to arrive at an estimate of interrater reliability. Rather, judges may rate a particular subset of items, and as long as there is sufficient connectedness (Linacre, 1994; Linacre, Englehard, Tatem, & Myford, 1994) across the judges and ratings, it will be possible to directly compare judges.

Disadvantages of Measurement Estimates

The major disadvantage of measurement estimates is that they are unwieldy to compute by hand. Unlike the percent-agreement figure or correlation coefficient, measurement approaches typically require the use of specialized software to compute.

A second disadvantage is that certain methods for computing measurement estimates (e.g., facets) can handle only ordinal level data. Furthermore, the file structure required to use facets is somewhat counterintuitive.

Comparing Consensus, Consistency, and Measurement Estimates

Table 1 presents a comparison of the key features of consensus, consistency, and measurement approaches to estimating interrater reliability.

Table 1: Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability

	Consensus	Consistency	Measurement
Purpose	Demonstrate exact agreement among independent judges	Functional purpose of getting judges to consistently apply a scoring rubric Consistent application is a means to the end of creating a summary score for each participant	Goal is the preserve as much information as possible from each judge and to incorporate that information into the model
Nature of the data	Nominal, Ordinal, or Interval (if there are few categories)	Ordinal, Interval, Ratio	Ordinal, Interval, Ratio
Advantages	Easy to compute by hand	Can be used when rating scales are continuous.	Provides one statistic that allows for direct comparison

	Strong intuitive appeal Effectively deals with nominal data Easy to diagnose rater discrepancies	Makes less stringent demands about training judges to exact agreement Possible to summarize rating from multiple judges	of the severity of all judges on all items, even if they did not rate the same items Differences in rater severity are taken into account at the level of the individual person (facets) or group (generalizability theory) Provides an empirical estimate of the extent to which judges consistently apply the rating scale across participants
Disadvantages	Must be calculated separately for each pair of judges and for each unique item Often requires substantial time and resources to train judges to exact agreement Rater independence can be trained away, thereby threatening the validity of the summary scores	Magnitude of correlation coefficients is affected by distribution of observed ratings. Can lead to artificially deflated estimates. Can be burdensome to compute the necessary adjustments to rater severity	Difficult to calculate by hand — requires specialized software Structure of the raw data file may be counterintuitive to set up (facets) Does not handle nominal level data
Example statistics	Percent agreement Cohen's kappa Jaccard's J	Pearson's r Spearman's rho Cronbach's alpha	Principal components analysis Generalizability theory Facet rater severity indices and fit statistics
End Goal / Assumption	Scores from multiple judges can be averaged once high consensus has been demonstrated	Requires some statistical adjustment for rater differences before scores from multiple judges can be summarized. Summarizing scores from multiple judges without adjusting for differences in rater severity could threaten the validity of study conclusions.	Rater severity is independently estimated and put onto a linear measure of underlying factor Differences in rater severity can be adjusted for on an individual and item level in the algorithm that calculates participant ability estimates (facets) Judges with high infit statistics may be drifting in their application of the rating scale and may need to be retrained

Intuitively, one may think that if two judges reach consensus, they will also be consistent. Although that is true in some circumstances, it is not true in others. For example, it is entirely possible for two judges to have an extremely high consensus estimate of interrater reliability and at the same time have a very low consistency estimate of interrater reliability. Table 2 provides a practical example of data that give a high consensus estimate and a low consistency

estimate of interrater reliability. The data indicate that the two judges have rated 80 participants on a single item. The rating scale for that item ranges from 0 to 1.

Table 2: Frequencies of ratings for two judges demonstrating high consensus and low consistency

		Judge 2		
		0	1	Total
Judge 1	0	76	1	77
	1	2	1	3
	Total	78	2	80
Percent Agreement	96%			
Pearson's r	0.39			

The reason for the large discrepancy between the consensus and consistency estimates of interrater reliability is because the Pearson r statistic is heavily influenced by the distribution of the data. Specifically, the Pearson correlation coefficient is based upon the assumption that the data are normally distributed. The example presented in Table 2 contains ratings that are highly skewed across both judges in that the majority of judge ratings fall into only one of the rating categories. This leads to an instance of an exceptionally high consensus estimates of interrater reliability (96% agreement) and at the same time, an exceptionally low consistency estimate (r = 0.36). Thus, it is important to recognize that some types of interrater reliability may be low simply as a function of the distribution of the data as opposed to being deflated due to a problem with the scoring rubric or with the judge training.

Conversely, the fact that two judges have a high consistency estimate of interrater reliability does not automatically imply that they will also have a high consensus estimate. Table 3 provides a practical example of data that give a high consistency estimate and a low consensus estimate of interrater reliability. The data indicate that the two judges have rated 60 participants on a single item. The rating scale for that item ranges from 0 to 4.

Table 3: Frequencies of ratings for two judges demonstrating high consistency and low consensus

		Judge 2					
		0	1	2	3	4	Total
Judge 1	0	9	0	0	0	0	9
	1	7	5	3	0	0	15
	2	2	5	9	9	0	25
	3	0	1	0	4	5	10
	4	0	0	0	0	1	1
	Total	18	11	12	13	6	60
Percent Agreement	47%						
Pearson's r	0.80						

Table 3 shows that the Pearson correlation coefficient between the judges is 0.80, while their percent-agreement figure is only 47%. If it is important for the study that the two judges reach exact agreement about how to use the various points on the rating scale, then the judges will need to be re-trained and/or the scoring rubric will need to be revised. If, however, the ratings are simply a means to the end of getting a summary score for each participant, then it may be possible to summarize the scores after using some adjustment factor. It is important to note that if a researcher were to report a high consistency and low consensus estimate of interrater reliability, but then take no steps to appropriately correct for the discrepancy when summarizing the data, systematic bias could be introduced into the results, which would then invalidate inferences from subsequent analyses.

One method for computing the differences in rater severity and incorporating them into the final summary score for each participant is through the use of the many-facets Rasch model. Table 4 presents a rater-measurement report generated by the FACETS computer program (Linacre, 1988). The table presents an indication of the severity level of each individual judge, along with various fit statistics to help in diagnosing the extent to which each judge was consistent within his or her own use of the scoring rubric. The benefit of this table is that the relative severity of all judges can be compared simultaneously.

The first column in Table 4 provides the initials for 12 judges who scored the open-response items on the essay portion of this particular exam. The second column indicates the number of items that were scored by each judge. The third column, "Measure", indicates the severity level of each judge along a linear continuum (using a logit scale), where

higher values indicate more severe judges. The fourth column, "SE", presents the standard error associated with each of the rater-severity measures. In general, the more essays a particular judge scored, the smaller the standard error of the rater severity measure (Myford & Cline, 2002). Finally, columns 5–8 present the fit statistics associated with each judge.

Table 4: Interrater reliability output: Measurement approach

Rater	N of ratings	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
CL	258	0.89	0.04	3.4	9	3.3	9
SK	1896	0.38	0.01	1.4	9	1.6	9
AM	1491	0.14	0.02	2.2	9	2.1	9
JJ	756	0.11	0.02	1.3	9	1.4	9
LM	758	0.09	0.02	1.2	5	1.3	7
CK	956	0.03	0.02	1.0	1	1.0	-1
SG	815	0.02	0.02	1.9	9	2.0	9
KV	2685	0.01	0.01	1.6	9	1.6	9
ER	3369	-0.02	0.01	1.0	2	1.1	4
JW	940	-0.17	0.02	2.4	9	2.5	9
KA	1662	-0.57	0.02	1.1	4	1.0	0
AP	869	-0.91	0.03	1.0	0	0.9	-2

The judge-severity indices (i.e., "Measure") are useful for estimating the extent to which systematic differences exist between judges with regard to their level of severity. For example, Judge CL was the most severe rater with an estimated severity measure of +0.89 logits. Consequently, students whose test items were scored by CL would be more likely to receive lower raw scores than students who had the same test item scored by any of the other judges. At the other extreme, Judge AP was the most lenient rater with a rater-severity measure of -0.91 logits. Students who had their exams scored by AP would be more likely to receive substantially higher raw scores than if the same item were rated by any of the other judges. The results presented in Table 3 show that there is about a 2-logit spread in systematic differences in rater severity (from -0.91 to +0.89). This spread indicates that simply assuming that all judges are defining the rating scales they are using in the same way is not a tenable assumption, and that differences in rater severity must be taken into account in order to come up with precise estimates of student ability. The many-facets Rasch model allows for these differences to be incorporated into the final participant ability estimate (Linacre, 1994).

In addition, Table 4 provides a series of fit statistics that are useful for interpreting the degree of intrarater reliability. Rater infit statistics are interpreted much the same way as item or person infit statistics are interpreted (Bond & Fox, 2001; Wright & Stone, 1979). Infit mean squares greater than 1.3 indicate that there is more unpredictable variation in the judges' responses than we would expect based on the model. Infit mean-square values less than 0.7 indicate that there is less variation in the judges' responses than we would predict based on the model. Myford and Cline (2002) have noted that high infit values may suggest that ratings are noisy as a result of the judges' overuse of the extreme scale categories (i.e., the lowest and highest values on the rating scale), whereas low infit mean-square indices may be a consequence of overuse of the middle-scale categories (e.g., moderate response bias).

Summary and Conclusions

Interrater reliability is one of the most important concepts in educational and psychological measurement. Without demonstrating that two independent judges can be reliably trained to rate a particular behavior, our hope for achieving objective measurement of behavioral phenomena is diminished. The concept of interrater reliability, however, has received far less theoretical attention than it warrants. Many key authors in the area of educational and psychological measurement and statistics have described interrater reliability as if it were a single construct with a single meaning. In this paper, I have attempted to demonstrate that the broad range of meanings implied by the term interrater reliability suggest that the field may benefit from increasing the precision of the use of the term interrater reliability. Toward this end, this article describes one potential framework for categorizing the various statistical methods for computing interrater reliability into a meaningful system based upon shared assumptions of the aim behind each technique.

When evaluating the results of an interrater reliability study, there are two important components to note. The first is the specific index of interrater reliability the authors have used. Some statistics include percent-agreement, Cohen's kappa, Pearson's r, Spearman's rho, percentage of variance accounted for by the first principal component, rater severity estimates, or fit statistics. Each of these statistics will provide a statistical estimate of the extent to which two or more judges are applying their ratings in a manner that is predictable and replicable.

The second important thing to note is whether the data are summarized in a manner that is consistent with the assumptions of the approach to estimating interrater reliability that was used. For example, if a researcher study reports a high consistency estimate of interrater reliability, this does not imply that the researcher is then free to simply average the scores across the judges. In fact, simply averaging the scores could lead to spurious results in subsequent analyses.

At times, it will make sense to report indicators of interrater reliability that represent more than one approach (e.g., both consensus and consistency estimates) because the distributional properties of the data will artificially deflate one or the other. The appropriate approach to estimating interrater reliability will always depend upon the purpose at hand.

Author's Note

Preparation of this article was supported by a Grant Award # 31-1992-701 from the United States Department of Education, Institute for Educational Sciences, as administered by the Temple University Laboratory for Student Success. Grantees undertaking such projects are encouraged to express freely their professional judgment. This article, therefore, does not necessarily represent the position or policies of the U.S. Department of Education, and no official endorsement should be inferred.

I would like to thank my colleagues at the Yale University PACE Center, especially Robert J. Sternberg, Elena Grigorenko, Robyn Rissman, Cynthia Matthew, Jonna Kwiatkowski, Damian Birney, and Carolyn Parish, for their helpful comments and suggestions during the preparation of this article.

References

Barrett, P. (2001, March). *Assessing the reliability of rating data*. Retrieved June 16, 2003, from http://www.liv.ac.uk/~pbarrett/rater.pdf

Bock, R., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*(4), 364-375.

Bond, T., & Fox, C. (2001). *Applying the Rasch model*. Mahwaw, NJ: Lawrence Erlbaum Associates.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*(2), 159-172.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (Third ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Glass, G. V., & Hopkins, K. H. (1996). *Statistical methods in education and psychology*. Boston: Allyn and Bacon.

Harman, H. H. (1967). *Modern factor analysis*. Chicago: University of Chicago Press.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*(3), 354-367.

Hopkins, K. H. (1998). *Educational and psychological measurement and evaluation* (Eighth ed.). Boston: Allyn and Bacon.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

LeBreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80-128.

Linacre, J. M. (1988). FACETS: a computer program for many-facet Rasch measurement (Version 3.3.0). Chicago: MESA Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2002). Judge ratings with forced agreement. *Rasch Measurement Transactions, 16*(1), 857-858.

Linacre, J. M., Englehard, G., Tatem, D. S., & Myford, C. M. (1994). Measurement with judges: many-faceted conjoint measurement. *International Journal of Educational Research, 21*(4), 569-577.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation, 7*(25).

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation, 7*(10).

Myford, C. M., & Cline, F. (2002, April). *Looking for patterns in disagreements: A Facets analysis of human raters'*

and e-raters' scores on essays written for the Graduate Management Admission Test (GMAT). Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LA.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Stemler, S. E. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation, 7*(17), Available online: http://PAREonline.net/getvn.asp?v=7&n=17.

Stemler, S. E., & Bebell, D. (1999, April). *An empirical approach to understanding and analyzing the mission statements of selected educational institutions.* Paper presented at the New England Educational Research Organization (NEERO), Portsmouth, NH.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2), Retreived February 16, 2004 from http://PAREonline.net/getvn.asp?v=9&n=2.

Uebersax, J. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101*(1), 140-146.

Uebersax, J. (2002). *Statistical methods for rater agreement*. Retrieved August 9, 2002, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

Winer, B. J. (1962). *Statistical principals in experimental design*. New York: McGraw-Hill.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Descriptors: Interrater Reliability; Rating Scales; Scoring; Scoring Rubrics; Error of Measurement; Evaluation Methods; Evaluators; Examiners

Citation: Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Available online: http://PAREonline.net/getvn.asp?v=9&n=4.