Generally, interrater reliability refers to the consistency with which two (or more) raters evaluate the same data using the same scoring criteria (Bailey, 1998) at a particular time (Stemler, 2004). Gwet (2001) states that “[v]irtually anything that is used to generate explicitly or implicitly a measure for classifying a subject into a predefined category can be considered a rater” (p. vii). Stemler (2004), however, argues that “the general practice of describing interrater reliability as a single, unified concept is at best imprecise, and at worst potentially misleading.”

So how is the consistency between two raters determined? Interrater reliability should be established “outside of the context of the measurement of your study” (Social Research Methods). It is best to do this through a pilot study prior to the full study. The two most common ways to measure interrater reliability are percent of agreement and correlation (Hayes and Hatch, 1999). Although some researchers prefer the percentage agreement measure over correlation measures because it is “conceptually simpler and easier to compute,” some would argue that percentage of agreement between raters is not the best measure (Grayson & Rust, 2001; Hayes & Hatch, 1999).

For the percentage of agreement, the agreement rate ($A$) is the observed agreement ($O$) divided by the possible agreement ($P$). Thus $A = O/P$ (Grayson & Rust, 2001). Cohen (1960) later proposed the kappa coefficient to improve upon the limitations found in the percentage of agreement. Hayes and Hatch (1999) posit that percentage of agreement does not take into account chance agreement to the total agreement score; whereas correlation measures do take into account the contribution of chance agreement in the calculation of reliability estimates. In addition, Uebersax (1988) provided the conceptual basis for making inferences concerning rating validity from agreement data. Regardless of the limitation cited for percentage of agreement, it may indeed be prudent to adopt this measure and to use correlation measures on a more limited scale.

According to Stemler (2004), the behavior of raters “invites some degree of subjectivity” and the ratings given will depend upon the rater’s interpretation of the criteria. An important strategy for reducing subjectivity, he suggests, is to develop sound rating scales (Mertler, 2001; Moskal, 2000; Moskal & Leydens, 2000; Tierney & Simon 2004). Training raters how to interpret the criteria and consistently apply the rating scale, Stemler contends, will “impose some level of objectivity” onto the process.

Stemler (2004) proposed the following three classifications of interrater reliability that synthesize the numerical statistical methods found in a thorough review of the literature: 1) consensus estimates, 2) consistency estimates, or 3) measurement estimates.

Consensus Estimates

Consensus estimates are based upon the idea that raters should be able to come to exact agreement about how to apply the rating scale. If the raters come to exact agreement, then it can be concluded that they share a common understanding of the criteria on the scale. If the raters can be trained on how to interpret the rating scale, then the scores given by the raters can be treated as equivalent.
Consensus estimates tend to be useful when the data are nominal and the different levels on the rating scale represent qualitatively different ideas. In addition, these estimates can be useful when the different levels of the rating scale represent a linear continuum, but are ordinal (e.g., a Likert scale). Bailey (1998) describes nominal (or categorical) data as a type of data that divides people or things into categories because the labels serve to name the classes of people or things under consideration. Ordinal data, Bailey continues, derive their name from the fact that the people or items in a series are rank ordered in terms of some quality or attribute.

The most popular method for calculating a consensus estimate is through the use of the percentage of agreement figure. As previously mentioned, the figure is calculated by adding up the number of cases that received the same rating by both raters and dividing that number by the total number of cases rated by the two raters. The advantage to this method is that it is intuitively appealing, easy to calculate, and easy to explain. The primary disadvantage is that it is possible to get artificially high percent-agreements because most of the values fall under one category of the rating scale (Hayes & Hatch, 1999). In addition, percentage of agreement is both time consuming and labor intensive. It is imperative that the raters are trained to the point of exact agreement.

A possible modification of the percentage of agreement method is to include adjacent scoring categories on the rating scale to include in the calculation of agreement. Raters would not have to come to exact agreement about the ratings they assign, so long as the ratings do not differ by more than one point above or below the other rater. The advantage to this method is that it relaxes the strict criterion that the raters agree exactly. The distinct disadvantage is that it can also lead to an inflated interrater reliability estimate if there are only a limited number of categories to choose from (e.g., a 1-4 scale).

A third consensus estimate is Cohen’s (1960) kappa coefficient. This statistic corrects the percentage of agreement estimate by taking into consideration the amount of agreement that could be expected by chance. Landis and Koch (1977) suggest that kappa values from 0.41-0.60 are moderate and that values above 0.60 are substantial. The major advantage of the kappa coefficient is that it is useful when one is concerned that the percentage of agreement statistic is artificially inflated. The major disadvantage is that the statistic can be somewhat difficult to interpret.

Stemler proffers the following advantages and disadvantages of consensus estimates:

**Advantages**
- The calculations can be easily done by hand.
- The techniques are well suited to dealing with nominal variables whose levels on the rating scale represent qualitatively different categories.
- The estimates can be useful in diagnosing problems with raters’ interpretations of how to apply the rating scale.
- A high level of consensus implies that both raters are essentially providing the same information.
Disadvantages

- The interrater reliability statistics must be computed separately for each item and for each pair of raters.
- A considerable amount of time and energy is needed to train the raters to come to exact agreement.
- 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 (Linacre (2002)).
- The estimates can be overly conservative if two raters exhibit systematic differences in the way that they use the rating scale but simply cannot be trained to come to a consensus.

Consistency Estimates

Consistency estimates are based upon the idea that it is not really necessary for two raters to share a common meaning of the rating scale, as long as each rater is consistent in classifying the situation or event according to his or her own definition of the scale. These approaches are most useful when the data are continuous. Bailey (1998) describes continuous (or interval) data as a type of data that derives measurements from scales in which the intervals between one unit and the next on the scale are equal for the entire length of the scale. According to Barrett (2001), values greater than 0.70 are typically acceptable for consistency estimates of interrater reliability.

The most popular statistic for calculating the consistency between raters is the Pearson correlation coefficient. Like the percentage of agreement, this coefficient can be calculated for only one pair of judges at a time and for one item at a time. One limitation of this correlation is that it assumes that the data underlying the rating scale are normally distributed. A second consistency estimate is the Spearman’s rank coefficient. This coefficient, like Pearson’s, may be used where data are normally distributed. A major disadvantage is that it requires both raters to rate all cases. A third consistency estimate is to compute Cronbach’s alpha coefficient. This coefficient is useful for understanding the extent to which the ratings from a group of raters hold together to measure a common dimension. The major advantage of this coefficient comes from its capacity to yield a single consistency estimate of interrater reliability across multiple raters. The major disadvantage is that each rater must give a rating on every case, or else the alpha will only be computed on the basis of a subset of the data.

Stemler tenders the following advantages and disadvantages of consistency estimates:

Advantages

- The approach places less stringent demands upon the raters 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.
- Certain methods of this estimate (e.g., Cronbach’s alpha) allow for an overall estimate of consistency among multiple raters.
Disadvantages

- If the construct under investigation has some objective meaning, then it may not be desirable for the two raters to “agree to disagree.”
- Judges may not only differ systematically in the raw scores they apply, but also in the number of rating scale categories they use.
- The estimates are highly sensitive to the distribution of observed data.

Measurement Estimates

Measurement estimates are based upon the idea that one should use all of the information available from all raters (including discrepant ratings) when attempting to create a summary score for each respondent. In other words, each rater provides some unique information that is useful in generating a summary score for a person. It is not necessary for two judges to come to a consensus on how to apply the rating scale because rating differences can be estimated and accounted for in the creation of each participant’s final score. These estimates are best used when different levels of the rating scale are intended to represent different levels of an underlying unidimensional construct (e.g., competence in a subject).

One popular measurement estimate is computed using the factor analytic technique of principle components analysis as outlined by Harman (1967). If the shared variance is greater than 60%, then the raters are successfully rating a common construct. An advantage of this approach is that it assigns a summary score for each participant that is based on the strongest dimension underlying the data. The disadvantage is that it assumes that ratings are assigned without error by the raters. Another method has been through the use of generalizability theory described by Shavelson and Webb (1991). A thirds measurement approach is through the use of the many-facets Rasch model offered by Linacre (1994).

Stemler offers the following advantages and disadvantages of measurement estimates:

Advantages

- The estimates can take into account errors at the level of each rater or for groups of raters.
- The estimates effectively handle ratings from multiple raters by simultaneously computing estimates across all of the items that were rated.
- The estimates do not require all raters to rate all items in order to arrive at an estimate of interrater reliability.

Disadvantages

- The estimates are unwieldy to compute by hand.
- The estimates can only handle ordinal level data.
Conclusion

According to Gwet (2001), interrater reliability “can be improved at the planning stage, before data collection, in order to ensure the validity of the findings” (p. 234). Gwet purports that it is imperative to have “a well-thought theoretical framework that provides practical guidance for generalizing results” (p. 11). This framework allows for a statistical inference to be drawn from a sample of the larger population. The assessment of the level of agreement between raters, therefore, involves “a definition of the phenomenon under study and the choice of an estimation procedure: (p. 11). After a review of the three approaches to estimating interrater reliability, it is recommended that consensus estimates be employed by the Migrant Education Program because of the relative ease of calculating the percentage of agreement between raters and the relative ease of calculating the kappa statistic when warranted.

References


