WYVERN
TOOLS FOR SOUND BUSINESS DECISIONS

The Validity and Reliability of Wyvern's Measurement Process

Validity of the Paired-Comparison Process

To the person working in the fields of measurement or psychometrics, the term "validity" has a specific meaning. Any rating or performance measurement project will be valid only to the extent that the criteria (survey items) are appropriate to the intended purpose of the project. That is why careful item development is very important. Items should be ***job relevant*** so that they apply to the issues or individuals being rated. They should be ***briefly defined*** so that they do not carry excess meaning or become all-encompassing. When measuring performance, the items should apply to ***observable behaviors*** *whenever possible* so that raters do not have to make inferences or resort to guesswork. The items should also be relatively ***independent*** of one another so that raters are not rating the same thing expressed in different ways.

When the items in a survey have been customized to meet the above criteria, the company and the participants can be confident that the survey results will be valid, i.e., they measure what they claim to measure.

Reliability of the Paired-Comparison Process

Reliability refers to the consistency of the ratings. Will the ratings be stable over a period of time if the survey is re-administered? The reliability of a measurement is defined by how closely the ratings will be identical when the measurement is retaken at different times. This is expressed as the degree of correlation between the first and second application and is expected to be .80 or higher out of a possible 1.0, i.e., the greater the number, the higher the similarity and the greater the evidence of reliability.

Reliability can be measured in several ways. One of these is inter-rater reliability, i.e., the extent to which the ratings of different raters agree with each other. For example, a group of 24 employees applying for promotion rated themselves by the Paired-Comparison Process, and they were rated by 13 supervisors, also using the Paired-Comparison Process. The correlation between the two groups of raters was .93. This is considered a very high reliability.

Another type of reliability is "test-retest" reliability. This refers to the stability of results when a procedure is repeated, usually within a short span of time. When the employee study mentioned above was repeated one week later and one month later, similar results were obtained, thereby supporting the test-retest reliability of the Paired-Comparison Process.

Reliability of the Paired-Comparison Process is achieved by two means: the use of **multiple raters** and the **multiple ratings of each behavior**.

The use of **multiple raters** reduces error in the rating process. For this reason, PSP asks that each ratee be rated independently by at least three raters. All that is required is that the raters have sufficient familiarity with the ratee. A single rater, even if trained in the rating process, typically will show less reliability than multiple raters because several raters produce a richer sample than does only one. That is why there are multiple judges in athletic events such as gymnastics and diving.

Making judgments about people is a difficult task, and various efforts have been made to achieve greater accuracy in the judgments. One popular attempt has been the adoption of Likert-type (typically five-point) scales for rating people on certain dimensions or competencies. In order to achieve any degree of accuracy from these Likert scaling procedures, considerable training of all raters is required, and there is the threat that the process will degrade as new – and untrained – raters become involved.

Training alone, however, does not solve the problems with the use of Likert-type scales. Other problems which could arise are as follows.

- Most raters don't use either the highest or the lowest ratings.
- Most raters put most people in the middle because it is the "safe" rating. This tendency is know as "average rater error."
- **Each behavior is typically rated only once; in the Paired-Comparison Process, the average rater rates a behavior multiple times.**

In an attempt to improve the accuracy of Likert-type rating scales, additional scale values are often added so that a five point scale becomes, for example, a ten point scale, the idea being that the more points on a scale, the more accurate it must be. However, the result is less accuracy because additional scale values make it even more difficult for raters to distinguish real differences. Adding more scale values simply introduces more variability (error) into the process.

The central problem with most rating systems is that they require judgments to be made in an absolute sense, i.e., they refer to ideals. But, **most real world judgments** are not made this way and – consciously or not – **usually involve a comparison of alternatives.** The Paired-Comparison Process makes use of this fact by asking raters to make comparisons rather than absolute judgments.

If a group of raters was asked to estimate the length of a piece of rope, there would be substantial disagreement among the judgments of the group members. If, on the other hand, the raters were shown two pieces of rope and asked which one is longer, a high degree of rater agreement would be obtained. Similarly, **the Paired-Comparison Process asks raters to compare a ratee's behavior with another behavior of that same ratee, thereby achieving a higher degree of rater agreement than other procedures that use non-comparative judgments.** The Paired-Comparison Process also requires more judgments or ratings before a final rating is reached. These additional judgments, made by comparing a single behavioral competency to other behavioral competencies, structure greater objectivity and accuracy in the final rating.

The Paired-Comparison Process also **prevents "gaming" the survey**. In a Likert-type rating scale, the ratings are transparent, i.e., the rater can readily "see" what rating he/she is giving the ratee. This enables the rater to adjust his/her ratings based on biases about the

individual. If the rater wishes to “send a message” to the ratee, either positive or negative, he/she can make their ratings artificially high or low.

The Paired-Comparison Process, however, asks the rater to judge whether the ratee is higher or lower on two behaviors, e.g., is John stronger in giving praise for a job well done or delegating tasks to the right person? To further strengthen the reliability of the rating, the rater is asked to judge the ratee’s performance in praising and delegating multiple times, each time against a different behavior. This makes it virtually impossible for a rater to “game” the ratings and further structures an objective end result.

Overall, **the Paired-Comparison Process structures a more objective and effective measurement system that ensures greater reliability for all concerned.**