



Why Reliability and Validity Matter

(From *Benchmarks® Newsletter, Volume V: Issue 1*, ©Center for Creative Leadership)

Jill Wachholz interviews Cindy McCauley

It is clearly part of the mission and strategy of the Center for Creative Leadership to provide research-based, psychometrically sound instruments. Why do we care so much about these issues, and hope that HR professional and leaders do, too? To help simplify these complex concepts, I consulted with Cindy McCauley, Research Scientist and co-author of *Benchmarks*.

JW: Cindy, let's start off with the basics.

What is reliability?

CM: Reliability is about consistency of scores. There are different kinds of consistency that are important to 360-degree feedback instruments.

First, there should be consistency in the ratings managers receive on the various items that make up a particular scale on the instrument (Internal Reliability). For example, on the *Benchmarks* scale “Being a Quick Study,” there are four items. If each of these items measures the skill of being a quick study, then the ratings a manager receives on these four items should be fairly consistent – not perfectly consistent, but in a large group of managers, those who score high on one item should tend to score high on the others. If we don't find this consistency, we doubt whether it is meaningful to place these items together on a scale.

Secondly, the scores a manager receives on a feedback instrument should be consistent across short periods of time (Test-Retest Reliability). If you received a 4.5 rating on “Resourcefulness” one week and then received a 2.5 rating two weeks later, you would naturally question the reliability of those scores. We want to ensure that ratings are not influenced by extraneous factors that might fluctuate over short periods of time, such as the mood of the rater

or the content of his or her last interaction with the manager.

Finally, we expect to find some consistency across raters (Interrater Agreement). Not that all raters will have exactly the same perspective – that's why we divide the raters into categories of Superiors, Peers, and Direct Reports. But within a category, we expect there to be some consistency in how a single manager is rated rather than a random pattern of ratings. If we don't find some consistency, then we would be concerned that the ratings were entirely a function of something idiosyncratic about the raters rather than a reflection of the skill level of the manager.

As an analogy, think about how restaurant critics assess the reliability of a restaurant. There are various parts to a meal (like scales in an instrument). To rate the restaurant's appetizers, the critics examine consistency across various individual appetizers (Internal Reliability). A good critic, in my view, also goes to the restaurant more than once to be sure that the quality of the food is the same on different occasions (Test-Retest Reliability). Finally, critics often take others with them to see if there is a consistent view among different people (Interrater Agreement).

JW: What about validity? That's always tougher to nail down.

CM: You're right, it is, because validity is more of a subjective process, rather than 100% proof of anything. With validity, we continually ask: Does this instrument measure what it claims to measure? Does it fulfill the function it was created for? That's what Construct Validity is all about. We want to know that people who score highly on the *Benchmarks* scales actually are more effective and successful managers.

JW: How do you choose the best way to validate an instrument?

CM: It depends on the purpose of the instrument. Is it trying to measure a manager's communication style? Likability? Effectiveness? The most common way to build evidence that a claim is valid is to relate the scores to some outside criterion that we agree would be a reflection of what it claims to measure. This is called Criterion Validity, and in the case of Benchmarks, we looked at performance evaluations and predictions of promotability to gauge a manager's effectiveness and degree of success in the organization. To continue with the restaurant analogy, different reviewers have different purposes. Some may assess the gourmet quality of the food and relate that to the number of awards the chef has won, and others might measure the service by relating it to customer evaluations of the staff.

There's also Face or Content Validity, which looks at the questions and asks if these items relate well to the content of a scale such as "Leading Employees," and also if we've left something out. Usually we have experts look for that. When we developed Benchmarks, we asked practicing managers and clients to give us feedback about the content of the items, and then we refined the questions based upon their input.

JW: Why should HR professionals, managers and executives care about reliability and validity?

Why does it matter?

CM: It matters for two main reasons. First, because the scores the managers receive back mean a lot to them. They take them very seriously and are asked to make decisions and development plans based on those scores. So you want to be sure that you can rely on those scores, that they're consistent and reflect some kind of accuracy. Second, if you say these

areas are important to develop in order for you to be successful, then you ought to be sure that's true because people are going to put a lot of energy into supporting that.

In our work and personal lives, we usually do research on things we invest time and money into to see if there's evidence that this is a positive approach. So why wouldn't you investigate an instrument for accuracy and relation to important areas? Without reliability and validity measures, HR professionals and managers simply wouldn't know what they were getting.