Test Scores Can Be Meaningless: Just Because It's a Number Doesn't Mean It Has Value

Let's start by contrasting "time measurement" and "skill measurement."

"Time" is an important concept. Understanding it is vital in managing our personal and work lives, so we measure it with consistent standards (seconds, minutes, days, etc.). We even toss in an extra day once in a while to "make up" for the accumulated sloppiness in our more common time measures. February 29, 2024 will "reset" time measurement for a while.

"Skill" is an important concept too. (As are "Skill's" cousins - "Knowledge", "Ability", and "Personal Characteristic.") "Skill" is at the root of workforce performance. Understanding it is vital to achieving desirable business outcomes, so we measure it too.

That's where the "time" and "skill" analogy falls apart. Consistency in skill measurement is hard to achieve and not guaranteed. Unlike time measures, test scores can be consistent or so sloppy as to render them useless.

Consider a test score like "75." "75" can be interpreted in these ways.

- *Higher Score = More Skill* – Almost everyone interprets scores this way. A score of "75" reflects a higher skill level than does a score of "70." A score of "80" reflects a higher level of skill than "75", etc. (1:30pm is later than 12:00pm…consistent measurement.)
- *A Meaningless Number* – Rarely do people interpret scores this way. Scores of "70", "75", and "80" have no relationship to an underlying skill. "70" is no better than an arbitrary label like "meat". Rank-ordering people with these scores has no value. Skeptical? Stop it. Some tests are just not consistent measures of anything…particularly ones that well-intentioned employees write from scratch. (As if 1:30pm and 12:00pm are arbitrary labels like "hood ornament" and "peanut".)

The two interpretations above represent polar opposites, of course. In reality, "test consistency" will fall somewhere in the middle.

Just because a test score is a number doesn't mean it has value.

Measurement Consistency = Reliability

How do you tell whether a test score is consistent? A tool's consistency can be measured on a 0.00 (inconsistent) to 1.00 (consistent) scale. This consistency scale is known as "reliability." Higher reliability results in more consistent scores ("70" < "75" < "80").

Here are some common ways to gauge the reliability of an assessment tool:

- **Internal Consistency Reliability**
  - *Split Half* – Scores based on ½ of the test's items are compared to scores based on the other half of the test's items.

    - *All Possible Split Halves* – Very commonly used. It estimates reliability based on all possible split halves correlated with their corresponding alternative.
- **Alternate Forms Reliability** – Essentially the same as "Split Half" except the "halves" represent different versions (commonly referred to as Form A and Form B).
- **Test-Retest Reliability** – A test is administered to a group of people twice, each administration separated by time. Time 1 scores are compared to Time 2 scores.

ChatGPT captures these and less-common alternatives to measuring the reliability of assessment tools. Click here (free account required).

Important Considerations Related to Test Reliability

Here are some other facts that you might find useful as you consider leveraging skill measurement in your business:

- *The longer the better*. Adding similar items will improve reliability – but there are diminishing returns (not to mention practical considerations related to the time required of candidates).
- *Ditto for 360^0 and engagement surveys.* While the content above is focused on skill measurement reliability, it also directly applies to 360^0 development and engagement surveys. While these two survey options tend to be rooted in observable behaviors, the behaviors and resulting scale values should always be scrutinized from a reliability standpoint.
- *What Value Between 0.00 and 1.00 Represents "Good" Reliability?* - As you can imagine, deciding when to deem a test/survey "reliable" requires professional judgement. While "0.00" is decidedly unreliable and "1.00" is decidedly reliable, those values are exceedingly rare and anything in between takes training and experience to interpret. Just call us for advice here…no charge!
- "*Reliability" and "Validity" are different.* Reliability simply means that items work together in a consistent fashion. What they work together to measure is an entirely different matter known as validity. As you can imagine, only reliable assessments/surveys have a chance to measure something valuable (a.k.a., valid). Reliability is a necessary precondition for validity. Stay tuned for more on validity.

So, the next time you see a test score, keep in mind that it just might be meaningless.

This is the 2nd in a series on improving business performance via skills assessment. Our first is here – Legal Standards and Regulatory Compliance.

Also, check out our post 6 Reasons Employees Leave (And 24 Others that Don't Matter That Much).

Catch Rick's presentation to the Lake Washington Human Resources Association on December 13, 2023. Topic and registration – Is Your Company Culture a Competitive Advantage?.