**2022 November Psymposium**

**Exploring Assessment in Professional Psychology**

**Beyond Test Scores – Why Numbers Are Not Enough**

From the early days of psychology’s development as a field, psychometric testing has been an important part of psychological assessment. A client’s responses to questions, ratings on test items, and performance on tasks can generate scores that allow us to compare them to other people. That’s helpful. But what do the numbers tell us?

When I started in this field, I had to calculate item scores by hand, tabulate results, and look up norm-reference tables to summarize findings. We were closer to the data, so it was easier to appreciate how the test items contributed to the scaled scores. Over time, computer-based test administration and scoring has made the process easier by allowing clients to complete rating scales at their leisure, or examiners to simply sit the client at the computer and initiate automated testing. Computers now provide us with hyper-detailed results, comparison scores, and even offer interpretive descriptions. The convenience is appealing, and the breadth of detail is alluring. But we can easily be lulled into complacency and place too much trust in the test scores or a computer algorithm’s cold interpretations. In this article, I will review some of the common pitfalls of numbers-based decision-making and offer suggestions to alleviate them.

**Who is the Examinee?**

It’s handy when the test responder knows themself well and understands the test questions. Language, culture and interpretation, current emotional state, fatigue, attitude and motivation can all play a role in the way individuals respond to test items. If the person is completing a rating scale, where are they doing it? In a quiet space with time to reflect, or a public place with many distractions? On their coffee break at work, or after dinner when they are tired? Who might be looking over their shoulder to see their answers, or even influence their responses? And not everyone completes all the items properly. Which items did they miss or skip? What about responders who try to provide a rating in the space between “sometimes” and “often” rather than selecting one of the options. Then there was the client who simply scribbled overtop of his form: *“This is a government conspiracy.”* That told me a lot more than any test item did!

**Computerized Interpretations**

Computer algorithms can be used to generate hypotheses about the individual. But that’s often as far as they can go. A hypothesis is not a sure bet. It’s a conjecture. It’s an opinion from the professional who developed the test or its interpretive system. From that perspective, it can be helpful. But the writer never met your client, the client who just completed the form. The client whose responses contributed to the test scores, which in turn contributed to the descriptive interpretation. It’s important to be suspicious about the accuracy of any such interpretations. Consider them, but don’t embrace them with an uncritical eye. And even if you like what they say, try not to use their wording verbatim. After all, it needs to be your formulation, not theirs.

 - 500 Words -

**Who Generated the Data?**

Observer-raters are simply that. They may have the benefit of knowing the individual better than you; however, they rarely have the experience or expertise to judge items with a balanced eye. They don’t always understand what the test developer was asking, so be careful not to put too much trust in test scores that emerge from their ratings. We’ve all seen what happens when we give the same test to three different observers (e.g., mom, dad, and the teacher). That’s why I prefer to complete the Vineland Adaptive Behaviour Scales as an interview rather than trusting parents, teachers, or others to complete a rating scale accurately (e.g., the ABAS). When making decisions about intellectual disabilities, care, caution, and attention to detail is crucial. It’s risky to trust too heavily in the judgement of an untrained rater.

**Factors are Just Correlated Items**

Factors tell us which items hang out together, then the test developer assigns a name to the factor. Almost every factor constructed in this way will contain one or two items that will have you scratching your head wondering why it is there. That’s why qualitative item analysis matters. It does take time, but it will help to ensure that your interpretation of results is more accurate, more valid, and ultimately more useful. Rather than reading all of the test items in the order they were completed on the form, I prefer to review the items in their factor or scale groupings. Even if a scale is coherent, an item analysis will help you to understand how the individual got the score they did. For example, on a Likert scale from 0 to 4, was a raw score of 8 generated by two 4-point answers or eight 1-point responses? Also, you will then see how many items contributed to the scale score. I’ve seen some scales with only one or two items in them! This type of item-by-scale analysis will give you insight into the individual. Short on time to review the entire test? Then at least review the scale scores that were elevated (or low).

**Bimodal Scales**

With rare or unusual symptoms, a few endorsed items can generate an elevated scale score. Autism is a fine example. Psychosis is another. I always recommend reviewing test items to see what contributed to a high score. You might find a few surprises, like discovering that the endorsed item can be explained by something other than what the scale’s name suggests. For example, anxiety can create rigidity and social avoidance just as autism can.

**When Statistics Mislead**

Some tests calculate a lot of comparison scores (e.g., the Delis-Kaplan Executive Function System, continuous performance tests, and measures that compare cognitive test scores to academic test findings). Each calculation is typically based on a 95% confidence level, which means it’s accurate 19 times of 20 (within a margin or error). This also means that if you conduct 20 test-to-test comparisons, you can bet that one of them will be misleadingly high or low. The problem is you won’t know which one… that’s why I suggest limiting your contrast calculations to just the ones that you need. Also, it can be helpful to plan for overlap in test domains (e.g. two measures of auditory working memory). And if a score looks unexpectedly low, administer a second one (e.g., word reading accuracy) and pay close attention while they complete it.

**Investigate Academic Test Errors**

If you find a low reading comprehension score, you probably haven’t discovered anything that the teacher didn’t already know. And if an examinee performs poorly on a measure of math, a subtest that amalgamates multiple domains into a single test score won’t tell you where the individual’s weaknesses are. That’s when a little item analysis can help. As they complete the testing, keep track of their errors. Are their mistakes in fractions? Long division? Decimals? Carrying over? Basic math facts? When reading comprehension is a problem, after you finish testing, ask participants to read a troublesome passage out loud to you so you can see where their errors are. Then try some trial teaching. Can you get them to engage with the reading material to better understand the content? Can you help them to identify their procedural error in math? Or can you teach them a specific math skill to see if they are able to learn it? This can be translated into intervention guidance, which can be extremely helpful to teachers, parents, and the students themselves.

**Social-Emotional Tests aren’t Diagnostic**

Measures of psychosocial functioning can be extremely helpful in formulating hypotheses. Sometimes they even offer a score on a risk factor scale identifying the likelihood they actually have a disorder. However, no test score will be diagnostic in itself. For lots of reasons, very few tests actually include the exact items from diagnostic criterion sets. And even if items are closely aligned to diagnostic criterion sets, do they survey the time frame required for the diagnosis (e.g., two weeks? six months?). Ultimately, a high score on a measure of depression doesn’t necessarily mean the individual has a Major Depressive Disorder. And a high score on an anxiety scale doesn’t tell you what type of anxiety the person might have. Review the test items. You might even want to ask some follow-up questions based on what you find.

Psychometric testing can be a valuable component of psychological assessment. I encourage you to create a habit of looking beyond the test scores so that you can better serve your clients, and your profession.

Dr. Michael Lee Zwiers, R. Psych.