Some Thoughts on Concordances

Michael E. Walker, PhD, NCME President

I have noticed a recent proliferation in the number of concordances on the internet, especially among various tests; but not as much guidance on how to use them. We encounter lookup tables, crosswalks, or concordances, quite often, possibly every day. These include translations of physical measurements from metric (meters, kilograms, liters, degrees Centigrade) to United States customary units (feet, pounds, quarts, degrees Fahrenheit); of one manufacturer's product numbers to another's (say, for watch batteries); or from one classification of educational courses to another. We also see them quite often with educational tests for college entrance, language fluency, medical licensure, and much more. I will share some thoughts that might help people to evaluate the quality and the limitations of such concordances.

What are concordances? Concordance tables have been around for more than 100 years. Warnings about possible misuse have continued unabated during that entire span (e.g., see Kelley, 1923, pp. 109-122; Eignor 2007 NCME Presidential Address; Dorans, 2020). For the most part, the warnings go unheeded. A big problem may be that people have faith that measurement is all high quality: as reflected in the attitude that "numbers don't lie," that sort of thing. People also believe that concordances, too, are pretty much just what they claim to be. That is not always the case.

With physical attributes, there is a neat one-to-one correspondence between two measurement scales. In this sense, the concorded scores and the original scores are interchangeable. For example, whether we measure length in meters, or whether we measure length in feet and convert to meters, we should get the same answer. With other kinds of measurement, there may be subtle differences between the two sides of the concordance that we need to know about. I found out first-hand, for example, that just because two button batteries have the same size and voltage, and a lookup table says they are equivalent, that does not mean I can use them interchangeably in my equipment.

With educational tests, we need to pay even closer attention to what is being concorded. A bad decision involves more than just a possible equipment malfunction; people's futures and livelihoods may be at stake. That is why a user of any concordance between educational tests should ask certain questions, which the responsible concordance creator should answer. I will discuss what I consider the most important issues here.

Are both tests valid, reliable and fair? Before we can reasonably concord two tests, there must be evidence to show that each test consistently (reliability) measures what it claims to measure (validity), and that it does so equally well for everyone (fairness). Whenever an organization produces a concordance for public use, that organization should include information on the quality of both tests.

Are the two tests similar in purpose and content? People construct concordances for tests that are used for similar purposes. For a concordance to make sense, the tests need to measure similar subject matter and have the same level of difficulty. Just because two tests both measure math, that doesn't mean they can be successfully concorded. One test may measure mostly algebra, while the other measures geometry. Or one may be a test of basic skills, while the other is a test of advanced skills. If a testing organization is being responsible, before it concords two tests, it will show the correspondence

between the two tests in terms of content. The organization will also show the relationship, or correlation, between the two sets of scores. Correlation measures the degree to which people tend to score about the same (either high or low) on both tests. A perfect correlation is 1.0, indicating that both tests would place test takers in exactly the same rank order. That won't happen, even if people are taking two versions of the same test. But the goal is to get as close to possible to the ideal. Dorans and Walker (2007) argue that a correlation of at least 0.87 generally indicates a level of correspondence between tests that should result in a good concordance. The lower the correlation, the less we can trust the concordance.

Are the tests designed for the same population? Continuing the previous example, we would expect a math proficiency test for engineers and a math proficiency test for Humanities majors to be different, even if they are both covering math. And even if the corresponding tests were each built to have a moderate level of difficulty in their respective populations, the difficulties would look quite different if we gave the tests to the opposite populations. Consequently, it is important for an organization issuing a concordance to mention the intended population. A concordance will be most defensible when the two tests are designed to be used by the same population of test takers. If not, make sure the organization issuing the concordance explains the population each test is designed for, as well how and why a concordance makes sense in the first place.

Does the sample used to create the concordance match the test-taking population? Concordances are built using real data from test takers. The resulting concordance depends on exactly who is included in the data and who is left out. Because two tests to be concorded are not built to be identical, their relationship to each other will be different for different populations. As mentioned earlier, ideally both tests would rank order the population of test takers in the same way. If they did (because the correlation was near 1.0), then the tests would also order test takers in different subgroups of the population in the same way. As tests become more different, we can expect different results in different subgroups. Continuing our example from before, we can imagine that Humanities majors would show a range of scores on the Math for Humanities Test. We might expect scores in this group to be mostly low on the Math for Engineers test. By contrast, Engineers might show mostly high scores on the Math for Humanities test but a wide range of scores on the Math for Engineers test. If we tried to build concordances in the two groups of test takers, these two different patterns of results would lead to two different concordances. Even if each concordance were useful for the group in which it was built, it might give misleading results if used in the other group. Make sure the organization issuing the concordance clearly describes the sample used to make the concordance, as well as the population for which the concordance should be used. Be wary if the two don't match. Take care applying the concordance to any population that does not look like the sample used to create it.

Did the sample used to create the concordance take both tests under authentic testing conditions? Whenever an organization produces a concordance, it needs to find test takers who took both tests, during actual test administrations, under actual test-taking conditions. Otherwise, the motivation of test takers (and therefore their performance) might be different from actual testing conditions. Ideally, the two tests would be taken close in time to each other (and with different people taking the tests in different orders), so that the test takers would have the same level of knowledge for both tests. Be wary of concordances based on tests that were not taken under actual test-taking conditions (e.g., in a "special study"), or tests that were taken far apart in time. Good data on both tests can be hard to get. The process becomes much easier when the creators of both tests cooperate to make the concordance.

Are chains of concordances involved? If A is related to B, and B is related to C, that doesn't necessarily mean that A is related to C. Anyone who has played the Telephone Game can appreciate this. It is generally inappropriate to concord Test A to Test B; to concord Test B to Test C; and then to derive a concordance of Test A to Test C by sticking the first two concordances together. Although there may be rare instances where this practice actually works, for example when the concordances are produced from a sample that took all three tests at about the same time, there are usually too many unknown factors to rely on such a concordance.

Are concorded scores interchangeable? The answer is always "No." Unlike meters and feet, two different test scores will not measure the same thing. In this sense, we cannot consider the score on one test to be an exact substitute for the score on another. Be wary of any test concordance that includes language suggesting exact interchangeability (e.g., "this is the score you would have gotten if you took the other test"). Rather, a concorded score is an estimate of what a person would have gotten had the person taken a different test. That estimate has a margin of error. As long as a score change within the margin of error does not change the conclusion about the person's status, it is generally okay to use the concorded score instead of the actual score.

Are the limitations of the concordance clearly stated? As I mentioned earlier, no concordance is perfect. There is always error involved. A concordance cannot apply to every situation, and no concordance is appropriate for every population. Before using a concordance, look for any documentation from the issuing organization explaining when and how (and, more importantly, when *not* and how *not*) to use it, as well as on the margin of error around concorded scores. One example of such a document is for the ACT/SAT concordance, developed jointly by ACT and College Board, in collaboration with the NCAA Technical Advisory Board. The document lists some uses of the concordance, followed by key limitations and technical information. A somewhat different approach was used for the COMLEX-USA/USMLE concordance. In a blog, the National Board of Osteopathic Medical Examiners gives guidance on when and how to use concorded scores. Linked documents give technical details. Both are easy to read and meet the majority of the criteria I have listed here.

May the concordance user beware. The bottom line is that, like anything we find on the internet, we need to be savvy consumers of concordances. Don't just take at face value that a published correspondence between two tests is appropriate for a given use, especially any use that requires interchangeability of the scores. Check the details for answers to the questions above. If the answer is missing, assume the answer is "no." The answers to these questions help me know how to use a concordance responsibly. Many "no" answers make me more wary of using the concordance at all. Given how improper use can affect a person's life, I think this is the right approach.