Evaluating the COMPAS Risk/Needs Assessment Tool

Validity vs. Reliability vs. Norming

In looking at the COMPAS needs/risk assessment tool there are a few ways that we can evaluate how well the tool is working. These are split into the categories of validity, reliability and norming. These types of evaluation and how to look at them for the COMPAS assessment tool are briefly explained below.

Validity

Validity is the extent to which an assessment measures what it is supposed to. For our purposes in evaluating COMPAS, this means how well the scales (26) are measuring what they say they are measuring. In COMPAS there are two different types of scales, risk and needs. There are different methods to check the validity for these two types.

The risk scales are designed to predict recidivism. For the risk scales, we can look at their predictive validity, how well do they actually predict recidivism (both violent and general). In order to do this we can look at future outcome data to see how well the risk scales predict recidivism using the Area Under the Curve (AUC) test statistic. The AUC is a measure of predictive accuracy, telling us how well the risk scales predict recidivism.

The second type of scales, needs scales, are designed to measure needs, inform case plans and identify intervention targets. Essentially these scales are used to help understand and predict criminality. While these scales are not used to measure recidivism directly, they are used to identify relevant targets for intervention. To look at the validity of these scales we can look at the associations between each scale and the recidivism rates (through correlations and the AUC values). This demonstrates how well the needs scales are measuring factors that have clear and predictive relationships to recidivism, and are thus useful in identifying intervention targets.

Reliability

Reliability refers to the overall consistency of a measure. Something has high reliability if the same results are produced when similar conditions are in place; essentially if the results are consistent when tested multiple times.

There are two different ways that reliability can be looked at. The first is test-retest reliability. This is whether the same results are found when the test is repeated. Due to the needs scales being measured using a mix of self-reporting and interview methods, there is a chance that the results will not be consistent when retested. Measuring inter-rater reliability is one way that this type of reliability could be examined.
The second type of reliability is Internal Consistency Reliability. This is when the correlations between the different items (questions) which make up a scale are looked at. If the items which make up the scales are internally consistent, the correlations will be high, and the scales will be reliable. These correlations can be measured using Cronbach’s Alpha, a statistic which looks at how well items are measuring the same thing.

Norming

Norming refers to the process of creating a ‘norm’ group to use as a comparison when assessing a specific group. This is done by taking a large amount of cases of what is being studied (in this case scale scores) and creating a distribution of the scores. Ideally the number of cases that are being taken to create this distribution is large and from multiple areas in the population (the larger the number and the more variety the better). By accumulating this large group of scores, one creates a ‘normal’ distribution of the scale in the population. This tells us, in general, what the distribution of this scale is, and the general amount of people who fall at certain score levels. From this distribution, a person can create score categories (high, medium and low). This distribution is then used as a comparison for the specific group being assessed. The score of the person being assessed will be assigned a category of high, medium or low based upon where their score fell in comparison to the norm population.

Example 1:

![Test Scores Distribution](image)

An easy example to think about is a bell curve (shown above) which is often used in grading schemes. For this example, let’s say that the distribution above is the test scores on a final from every senior in a specific school. The line represents the distribution of scores. Based upon the bell curve distribution, the majority of the students which took the test would fall in the middle range, and only a small percent would be in the top range and in the bottom range. Categories can then be made (in the form of letter grades in this example) based upon this distribution. So a certain test score now represents a category (a grade). When a specific classroom takes the test next, their scores will be assigned a category based upon the larger grade distribution from all the other seniors in the school.

It is important to note however, that the distribution of the created from the larger ‘norm’ group does not have to look like the bell curve (low on either end and high in the middle). For example, if someone asked individuals in a population about the number of times that they have been married,
the distribution would be high (have lots of responses) at the low end and low (fewer responses) at the high end. This distribution would look like the one below:

**Example 2:**

![Graph showing a bell curve distribution](image)

Categories could then also be assigned to this distribution, leading to a person’s score reflecting a category which they are assigned based upon this comparison population. What becomes important then is not the actual number of marriages, but the category of High, Medium or Low, which has been assigned to the individual, based upon their number of marriages in comparison to everyone else.

By creating a ‘norm’ group, individuals who receive a certain value on a scale can be compared to the larger ‘norm’ group, allowing them to be placed in the appropriate category. This allows for individuals to be placed into a category not based upon their absolute score on a scale, but based upon their where they fall on the scale in comparison to others in the population being looked at. This is the process of ‘norming’.

When using the process of norming it is important be sure that the groups which are being compared are similar both in composition and score distributions. If the ‘norm’ group (the group being used as a comparison) is too different from the group being studied then the results may not be accurate. For example if we look at a distribution of age in two groups, one, a town, which has an average age of death at 30, and one, the rest of the USA at an average age of death of 85, these two distributions will be different (shown below).
If we use the first group (average age of death at 85) as our norm group or our comparison group the categories don’t make sense for our specific group being compared (town with average age of death at 30). The categories of young, middle and old are created based upon the norm group distribution. So someone at an age younger than 20 is termed young. However, these categories do not make sense for the specific society being looked at in this case (the town with an average age of death at 30). If someone is 29 in this town they will be placed into the ‘middle’ category because they are middle aged compared to everyone in all the larger group. This does not work for the specific town however considering they are almost at the average age of death, and thus should be considered in the older category. Thus what would be considered ‘old’ in the first town, where the average age of death is 30, would be very different than what would be considered ‘old’ in the second group. Thus it would not make sense to use the norm group as a comparison for this specific town.

Part of evaluating the COMPAS program involves looking at their process of norming, and whether it works for the WI data. The COMPAS Reentry Norms Report for Men and Women evaluates whether or not using their ‘norm’ group as a comparison for the WIDOC data makes sense. The report looks at the different distributions of each of the 26 scales to see how well the distributions match.