Norm Referenced Assessments

W. Blaine Dowler

August 5, 2010

1 Norm Referenced Results

One of the two main ways to report student achievement is through norm referenced results. With this style of reporting, students are compared to age or grade peers, and performance is measured relative to this group. Because of this, interpreting norm referenced results correctly can be difficult. These interpretations are easier to make if one understands how norm referenced results are produced in the first place, and work onward from there.

2 Norming Groups

In order to produce a norm referenced assessment, one must first find a group of typical students who may be assessed with the tool. The group must be large; to do an accurate statistical analysis, you need a lot of data.[1] Furthermore, you need a spread of data.

Imagine you are creating a pencil and paper assessment tool used to measure performance of grade five students. The industry standard way to label school grade levels is with numbers that have one decimal place. The number before the decimal is the current student grade, while the number after the decimal is the number of months the student has completed in that grade. So, if the new school year begins on September 1, then a grade five student would be in grade 5.0 on September 15, moving up to grade 5.1 on October 1, and so

[1] As a rule of thumb, one wants 30 students to form a bell or normal curve. This is enough to rate students in a single post secondary course, but when the goal is to build a standardized tool for any peer group, groups need to number in the hundreds or thousands to work effectively.

forth, reaching grade 5.9 on June 1.[2] For accurate testing, you would want to know how typical students perform on the assessment at different points in their academic careers, so you would want to administer this grade 5 test to typical students every month of the year to build data.

There is a drawback to this, of course. Logistically, one typically needs to pay a school to use that school's students to build data. Furthermore, administering the same test to the same students every single month skews results, as the students start choosing answers they remember being confident in on the last assessment, or the answers their friends told them were right while discussing it after school the first time, and so forth. Therefore, it is more common to use "seasonal norms" by administering the assessment once per season, and then using mathematical interpolation[3] techniques to "predict" the values in between administrations.

Now, in any group, one will run across students who perform significantly above or below average. There are two ways to deal with this when building a reference. It is common to administer the assessment to students outside the intended grade to get the actual data to model; a test intended for grade five is typically given to students in grades four, five and six to generate the normative data. However, drifting too far from the intended grade level skews results. Administering a grade five assessment to a grade one class is virtually worthless, as students will be emotionally frustrated in trying to complete it, and the responses given would be effectively random. Administering the grade five test to a grade twelve class would be worthless on the other end; students would feel their time was being wasted by an assessment that easy. Typically, norm referenced results that are more than a year apart from the intended grade level are compared to data mathematically extrapolated[4] from the original set. As such, accurate results are highly dependent on the quality of the model used to fit the data. The different models will be discussed in more detail in lesson seven.

A final obstacle in this type of assessment is the fact that one needs a variety of possible scores to build an accurate model, which means longer assessments in many cases. With more assessment items, one is more likely to assess multiple skills at once. As such, students who excel in one category but struggle with

[2] The grade level assigned in July and August for the traditional ten month school schedule is ambiguous. Most bodies increment the grade on July 1, so this student would be in grade 5.9 on June 30, and then grade 6.0 from July 1 to September 30 inclusive. Some bodies leave the student grade as 5.9 through July and August, and increment on September 1.

[3] Mathematically speaking, interpolation is like playing "connect the dots" when you have some idea about what the picture should look like, but when some dots are missing. Given the expected picture, which would be the theoretical model of student performance in this case, one tries to deduce where the missing dots are and reports the values as such.

[4] Mathematically speaking, extrapolation is similar to interpolation, but is used to continue to "connect the dots" beyond the limits of the original picture. Because the reference dots are only on one side, it is harder to get accurate data in this fashion.

another may not fit the model very effectively. For example, if a student who excels at computation but struggles with reading is given a typical math assessment, the student may perform well above average on the computational items, but struggle with word problems due to the difficulties with reading itself.[5] Assessments that measure multiple skills are called *multidimensional assessments*, while assessments which measure only a single skill are called *unidimensional assessments*.

To overcome these issues, many modern norm-referenced assessments are being administered with technology, and will adapt their level to the student's functioning level. Thus, the assessment isn't tied to a particular grade, and can provide accurately normed results for students at all levels of ability, and with fewer assessment items. They tend to model each question individually, using advanced techniques discussed in lesson seven.

3 Typical Score Reports

The scores on norm-referenced assessments are typically published in one (or more) of four ways:

1. **Percentiles**: These are distinct from percentages. This compares the student to a percentage of the population that did as well as the student, or worse than the student. A student who is average among his or her peers will score at the 50th percentile on a properly normed assessment. A percentile score of 10 doesn't mean the student only got 10% of the questions correct; it means that 90% of the population scored higher than that student. In a class of 100 students, if the lowest mark on a test is 25%, then a raw score of 25% means scoring in the 1st percentile. This gets limiting when students get the same score: if 10 out of 30 students tie for the highest mark on a test, it can be mathematically ambiguous when it comes to determining which percentile to assign. Technically, they are all in the 100th percentile by definition, but the next lowest attainable percentile would be the 67th. That's a rather large range of ambiguity, which is why the average difficulty of most norm-referenced paper based tests is higher than the average difficulty of the typical classroom test.

2. **Grade Equivalents**: These are the scores used most often. A grade equivalent of 4.2 GE means that the testing student performed as well *on that particular testing tool* as you would expect from a typical student who had completed 2 full months of grade 4. Note that this is *NOT* a grade level. It's easier to see that from a top performer: an honours

[5] For example, the student may misinterpret a word problem requiring division, and then correctly multiply the numbers provided instead.

grade 8 math student could score, say, 11.4 GE on a grade 8 math test, because most students would be in the eleventh grade before they do that well on a grade 8 test. That does *not* mean that same grade 8 student could skip ahead to grade 11 math and expect to succeed. There are too many intermediate skills that he or she has never seen. It is entirely possible for a strong grade 3 or 4 student to score a grade 5 equivalent on a grade 2 test, and score a grade 2 equivalent on a grade 5 test. It also has the disadvantage that the actual ability represented by divisions in the numbers is variable. The average grade 12 student learns more in a month than the average grade 2 student, so the difference between grade equivalents 12.2 and 12.3 represents more skills and more learning than the difference between grade equivalents 2.2 and 2.3. It is not an "equal ability" scale. Grade equivalent results are particularly sensitive to the limitations of extrapolation. If a score is more than a year away from the intended grade level for the assessment tool, then the score really only means "too easy" or "too hard."

3. **Normal Curve Equivalent (NCE)**: This avoids some of the difficulty with percentile score. Instead of grouping the student population into 100 categories, it groups the actual scores into 100 categories. So, if the lowest score in a class on a test out of 50 is 17, and the highest score is 42, then there is a 25 point range in test scores. Each mark is worth $100 \div 25 = 4$ points on the NCE scale, so a score of 17 out of 50 has NCE 4, 18 out of 50 has NCE 8, and so forth. Assigning a student a score with the NCE score means comparing them to average performance relative to the test instead of performance relative to their peers. Unlike grade equivalents, this is an equal ability scale.

4. **Standard or Scale Score (SS)**: This is often reported, but the meaning is hardest to interpret. Part of the mathematical process is norming a test is setting up a completely arbitrary reference scale to measure scores against. This is the standard score, also referred to as a scale score. This is typically an equal ability scale, but it's generally meaningless outside the context of that particular assessment. Instead, it is most often used internally to represent a student's ability, and then converted into one of the three types of scores above.

Note that all of these reporting methods share a common limitation: the student reading the results has no idea which areas need improvement and how much improvement is needed. For example, if a student scores in the 50th percentile, he or she is at class average. That student doesn't know if the class average was 50% or 80%, and so on. Some computerized norm referenced assessments are starting to provide norms in different areas, but even then, norm referenced results do not provide the information needed to determine what needs to be improved.

4 Upcoming Lessons

In lesson five, we will discuss criterion referenced assessments and the manner in which the results are reported. In lessons six through nine, we will discuss different means of analyzing assessment items for a variety of purposes.