3 Psychometric Methods

JUDY A. SHEA
University of Pennsylvania

GREGORY S. FORTNA
American Board of Internal Medicine

SUMMARY

The purpose of this chapter is to provide an overview of several concepts and terms that were originally defined and investigated in the corner of education that housed psychometrics, but have migrated to the more general education literature. Definitions, explanations, and examples will be given for the commonly used terms including reliability, generalizability, and validity. Following the discussion of the common psychometric concepts and terms, the second part of the chapter provides an overview of how one might use these concepts in designing or choosing an instrument. The third part of the chapter will introduce some newer and more advanced topics that have received attention in recent years. The chapter will conclude with a brief review of practical suggestions for those engaged in educational research.

INTRODUCTION – THE PSYCHOMETRIC TRADITION

Psychometrics is one of the oldest research traditions in psychology. Its roots go back 100 years to two longstanding research programs in psychology. The first, *psychophysics*, addressed the issue of the relationship between the physical properties of an object, such as weight in kilograms or sound intensity in decibels, and the corresponding perception, such as heaviness and loudness. Methods such as Thurstone's Law of Comparative Judgment, which are occasionally applied to educational achievement, can trace their origins directly to issues in psychophysics. The second tradition has somewhat more nefarious roots. Since the late eighteenth century some psychologists have been preoccupied with measuring intelligence, and its relationship to other characteristics. Much of the early work in reliability and validity was rooted in research in intelligence.

International Handbook of Research in Medical Education ,97–126.
G.R. Norman, C.P.M. Van der Vleuten, D.I. Newble (eds.)

While these origins may appear as quaint historical side notes, it is important to recognize that many of the concepts of psychometrics, while they continue to stir debate among clinicians and educators, have been used in education and psychology for a very long time indeed. The concept of reliability, which we will soon encounter, is credited to Karl Spearman, a statistician working at the turn of the century (Spearman, 1904). The intraclass correlation coefficient, which has a fairly recent history in clinical measurement, was allocated a full chapter in the 1925 statistics book of Sir R. A. Fisher (Fisher, 1925). The Likert scale, the standard 7-point scale with Definitely Agree and Definitely Disagree at the two ends, was invented in a paper published in 1932 (Likert, 1932).

Moreover, there is a very good reason for medical educators to be conversant with these terms. A large proportion of research activity in medical education is related to issues of assessment, and the language of assessment is the language of psychometrics. It makes no more sense for a researcher in educational assessment to not understand these terms than for a physicist to be ignorant of Newton's Laws.

As a consequence, we will spend some time initially in an exposition of the terminology of psychometrics, if only because the terms are frequently misunderstood.

COMMON PSYCHOMETRIC TERMS

Reliability

One of the most common terms encountered in reading the medical education literature is reliability. There are two types of reliability: reproducibility [alternate forms, test-retest, inter-rater and intra-rater agreement] and internal consistency [also called homogeneity]. Both will be discussed in more detail. When used as an expression of reproducibility, reliability addresses the extent to which scores obtained on two occasions [test-retest] or with two equivalent forms [alternate form] or perhaps by two different raters or assessors [inter-rater] are similar. When used as an index of internal consistency, reliability refers to the degree to which the items are measuring a similar, unified concept or construct.

An understanding of reliability is enhanced by considering the technical meaning. A score on a scale (call it "X") is composed of two elements: true score or signal (T) and error or noise (E). Optimally, any measurement would contain only signal and no noise. The reliability of X is defined as the ratio of t/[t+e]. Reliability, then, is the ratio summarizing how much of the observed score is true score, or not due to error. Obviously, one cannot ever observe true score and thus it is necessary to estimate the statistic. Many sources provide the derivations of various reliability statistics, and computational formulas (Carmines & Zeller, 1979; Crocker & Algina, 1986).

Reproducibility

As mentioned earlier, one type of reliability is reproducibility. The question that is asked is "are scores obtained on one occasion [or with one rater, or with one scale] the same as those obtained on a second occasion?" The conditions of the occasions may be different times (test-retest reliability), different raters (inter-rater reliability), the same rater making repeated observations (intra-rater) or different forms of the test (alternate forms). The former two are frequently encountered; the latter are rare in education since few have the opportunity to have observers repeat observations under the same circumstances, and few, except national testing centers, have the resources or need to create multiple forms.

Most assessments of reliability are reported as some type of correlation, for example, an intra-class correlation, a Pearson product moment correlation, or a kappa. When a correlation is the statistic of choice, possible values range between 0 and 1.0 with higher values indicating more reliability. Some have argued that for scales that are used to make important decisions about individuals, coefficients should be 0.90 or higher. In educational research it is common to see values between 0.60 and 0.80.

A simple example will help to illustrate the meaning of repeatability. Assume that we had 10 medical students who were assessed by a tutor about their performance in the tutorial. (In reality we would almost always have a much larger sample of people.) Two weeks later, the tutor completed the scale a second time. Table 1 shows how the data might look.

Table 1. Sample data to illustrate reproducibility

Person ID	Day 1	Day 14
1	13	11
2	12	14
3	10	11
4	10	7
5	8	9
6	6	11
7	6	3
8	5	7
9	3	6
10	2	1

By looking at the data we can see that, in general, the people who scored the highest on Day 1 also scored the highest on Day 14, so that students who did well or poorly the first time also did so the second time. Note, however, that scores are not exactly the same on the two occasions, presumably due to random error. If we wanted to quantify the reliability, a simple procedure (though generally not the best statistical choice) would be to compute a Pearson product moment correlation, which is 0.76 for this example. However, a simple correlation is usually not the best

way to assess repeatability, since it is a measure of association – a linear relationship between the two scores.

A better statistical alternative is to compute an intraclass correlation coefficient which is derived from an analysis of variance framework, and directly estimates the relevant variances as well as dealing with multiple observations. For example, consider a simple study where three psychiatrists rate the "degree of sadness" in 10 patients, using an 11 point scale. Table 2 shows how the data might look.

It remains the case that the "signal" is captured in the systematic differences among patients, which is the averages on the right column. Systematic differences among raters are reflected in the column means of 4.0, 6.0 and 8.0. Finally, the "noise" is the random variation of the individual observations around the values predicted by the marginals. Table 3 shows how the ANOVA table would look[1]:

Table 2. Sample data to illustrate intraclass correlation

	Rater			
Patient	**1**	**2**	**3**	**Mean**
1	5	7	9	7.0
2	3	5	7	5.0
3	1	2	3	2.0
4	2	4	6	4.0
5	4	4	7	5.0
6	7	9	11	9.0
7	4	7	10	7.0
8	5	7	9	7.0
9	3	6	9	6.0
10	6	9	9	8.0
Mean	**4.0**	**6.0**	**8.0**	**6.0**

Table 3. Sample data to illustrate ANOVA summaries

Source	**SS**	**d.f.**	**MS**	**F**	**p**	**Variance component**
Patients	114.00	9	12.67	22.80	0.0001	4.04
Raters	80.00	2	40.00	72.00	0.0001	3.94
Error	10.00	18	0.56			0.56
Total	204.00	29				

Note that in the last column variance components are shown. These are not routinely provided in most ANOVA summaries, but they are straightforward to calculate (formulas are given in Streiner & Norman, 1995; Brennan, 1992). Essentially, variance components are summaries of "how much" of the observed variability in the data set is due to the various sources.

[1] Readers who are unfamiliar with analysis of variance may wish to consult *PDQ Statistics*, by Norman, G. R. and Streiner, D. L. Hamilton, ON: Decker, 1997.

If the scale is working appropriately, we would expect to see that some patients are consistently rated higher or lower than others regardless of who is doing the rating; that is, we would expect to see that most of the variance in the scores is due to differences among patients. The definition of reliability as a ratio of true score/(true score + error) or t/(t + e), captures this expectation. In this case the reliability coefficient would be 4.04/(4.04 + 0.56) or 0.88.

But what about the differences in raters that are apparent from looking at the table? Traditionally, this source of variance would be ignored (and it would be appropriate to do so if these were the only three raters that would ever be used – a "fixed" effect in ANOVA parlance). Probably it is more accurate to include the variance component for raters in the denominator as a source of error, assuming that the three particular raters in the study were a sample of all possible raters that could have been employed: 4.04/(4.04 + 0.56 + 3.94) = 0.47. The reliability is considerably lower, because we have treated systematic differences among raters as a source of error.

This general formulation can be applied to many reliability studies. Although we have chosen to use "rater" as a repeated measure, and therefore have computed an "inter-rater reliability", we could equally well have created a reliability study where students complete parallel tests on two different occasions, which we would call "test-retest reliability". We could consider an examination where the student sees multiple cases or problems (as in the Objective Structured Clinical Examination), in which case the coefficient might be labeled an "inter-case" reliability study.

Finally, one of the standard forms of reliability is called "intra-rater" reliability, in which the raters must see exactly the same stimulus (a video of the encounter, a chest film) on two occasions. This particular design is very difficult in practice. It is very expensive and time-consuming to get raters to repeat the task of rating examinees, essays or videos. Moreover, those who agree to do so may not be representative of the original sample. Completing an instrument on a second occasion raises many threats to the integrity of the scores (boredom, acquiescence, recall) (Campbell & Stanley, 1963; Streiner & Norman, 1995). In any case, it can be argued that, since different raters are likely to yield lower reliability than the same rater on two occasions, if the inter-rater reliability is adequate, the intra-rater can only be higher, so it is unnecessary.

In summary, reliability, and the reliability coefficient, reflects the ability of a test to differentiate among individuals; that is, to identify systematic differences between people (usually students). Many designs that occur in educational research, such as those that use multiple raters, occasions, etc., fit well with the ANOVA framework from which intraclass correlation coefficients are produced. More will be said about the ANOVA framework in the brief presentation of generalizability theory, which extends reliability to consider multiple sources of error together.

http://www.springer.com/978-1-4020-0466-7

International Handbook of Research in Medical Education

Norman, G.R.; van der Vleuten, C.P.M.; Newble, D.I. (Eds.)

2002, XIII, 1106 p. In 2 volumes, not available separately., Hardcover
ISBN: 978-1-4020-0466-7