Bases for Assessing Normative Samples

This Extension is an expansion and enrichment of pages 351-353 of your text that addresses the criteria by which to evaluate normative samples of published instruments. The first section of this Extension elaborates on the issue of the representativeness of the sample. The second section expands the text’s treatment of relevance of a reference group to one’s interpretive purpose.

Criteria for Judging Reference-Group Representativeness

The most common kind of reference-group data supplied for published instruments is national “norms” that enable an examinee’s performance to be compared with that of a representative national sample of other examinees of a given description. In this section we look at the issue of the representativeness of normative samples.

Educators who use published tests for making norm-referenced interpretations need to be able to judge the adequacy of the normative data for these instruments. Following is a description of things to look for in evaluating the adequacy of tests’ norming studies.

Samples can be categorized into two major groups, random and nonrandom. In random samples each person in a population (such as all fifth-graders in the country) has a known (and usually equal) probability of being selected; who gets picked is a matter of blind chance. In nonrandom samples, people are selected on some basis other than chance; who gets chosen in nonrandom studies is usually a matter of convenience and economy. Nonrandom samples may or may not be similar to the populations of interest, and there is no

---

1 We refer to studies that secure normative data from reference groups as norming studies. Most textbooks and test manuals call them standardization studies. Our reason is to avoid confusion between the processes of (a) gathering data for making norm-referenced interpretation of scores (i.e., norming a test) and (b) establishing standard procedures for administration and scoring a test.

2 In certain kinds of research (often qualitative in nature), purposeful sampling is a very reputable kind of nonrandom sampling. But that respectable method of nonrandom selection is not common among test publishers who use nonrandom sampling of normative samples.
way to know just how much they may misrepresent the populations. To make matters worse, publishers who sample Test users need to be astute enough to detect it from the procedures described in the test manual or to infer it from inadequate descriptions.

Reputable test publishers usually attempt—often diligently—to gather normative data from random samples. For example, the publisher of a third-grade math test should seek normative data that are based on a random sample of third-grade students. If, as is usually the case, the test is to be marketed nationally, the sampling is ordinarily national in scope.

Conceptually, this is simple. To illustrate, suppose a national sample of 5000 third graders is needed. Imagine that a list exists of all the third graders in the country. All their names are put into the proverbial hat, 5000 names are drawn out, and the test is given to each of these pupils. Unfortunately in practice, this simple random design is wholly impractical.

First, there is no such master list, and obtaining one would be prohibitively costly. Second, it would be unthinkable expensive and inconvenient to test 5000 pupils scattered in nearly that many schools all over the country.

To circumvent these problems, the publishers use cluster sampling by obtaining a list of all the school districts in the country and choosing only a sample of, say, 100 districts in such a way that the chance of selection is proportional to districts’ third-grade enrollment. Within each selected school district, participation might be sought for only one randomly selected school. (Let’s delay for the moment the matter of replacing those who decline.) Within each selected school 50 third graders might be randomly chosen to participate.³

A major concern with such a practical research design is the error that may result from the small number of sampling units (e.g., 100). To provide some protection against this source of error, stratification is often used. A stratified sample ensures proportional representation of each subgroup into which the population is divided. Within each stratum, selection is random.

³ For the technically inclined reader, we might say this is one method of random cluster sampling. Several other methods are also available.
To illustrate stratification, consider a well-known fact. Cognitive performance of students in some parts of the country is, on average, lower than performance of students in other regions (Hopkins, Stanley, & Hopkins, 1990). To protect against chance over- or under-representation of any region, publishers have long divided the country up into several regions, determined the fraction of the population in each, and then sampled that fraction of people from each region (Peterson, Kolen, & Hoover, 1989). This stratification of the random sampling procedures enhances the adequacy of the sample.

Another common stratification variable is population density (Peterson, Kolen, & Hoover, 1989). Rural students tend to perform better on cognitive tests than inner-city students, and suburban students perform best of all. Hence, most publishers stratify all the school districts of the country into several sizes and select the desired proportion from each size category. This can ensure the appropriate representation of students from each size of community. Another common basis for stratification is public vs. nonpublic schools. Ethnicity is yet another that receives considerable attention. Socioeconomic status (SES) is particularly useful for stratification. The correlation of school average SES and test performance is quite high.

If SES influence is statistically controlled, or held constant, the other stratification variables are much less important. For example, if school districts in the deep South are compared with those from other parts of the country having the same SES, little or no systematic difference in cognitive performance is likely. Similarly, if minority children are compared to nonminority children having the same SES, large differences in performance are unlikely. For these reasons, SES is the most important single stratification variable; therefore, test users should attend most closely to a research design’s treatment of SES.

Publishers cannot (for practical and legal reasons) obtain SES data in individual students; therefore, to stratify for SES, they use mean

---


5 When individuals serve as the unit of analysis, cognitive performance is correlated only moderately with SES (or ethnicity, region of the country, etc.). However in norming studies, whole schools or school districts are the unit of sampling; hence, the correlation of school averages of SES and test performance is much higher.
or median SES of persons in each district. Because SES is a complex and subtle abstract attribute, publishers have to use indicants rather than direct measures of SES. Well-conducted norming studies usually estimated SES by use of two indicants that are available from U.S. census data for school districts. One is income. The other is education (e.g., median number of years of school completed by adults over 25 years of age).  

Such indicants enable publishers to:
- quantify all school districts in the country on the basis of estimated average SES,
- separate this list into several SES strata,
- determine the fraction of the population enrolled in schools within each stratum, and
- seek participation of a corresponding fraction of the sample from each stratum.

Most major publishers stratify samples on the bases of multiple variables, such as SES, geographic region, population density, public versus nonpublic schools, and ethnic group. Professionals who are considering the use of a cognitive test should examine its manual to verify that SES was a stratification variable.

Mere assurance that such variables were used in selecting the sample does not suffice; authors should report in the manual the success of their efforts. This can be achieved with a simple table for each stratification variable that shows the percentage of the population and the percentage of the sample in each stratum. Prospective users should be deeply suspicious of the adequacy of a test’s reference group data unless they are described by such data.

Sadly, there are manuals in which few if any aspects of the norming study are described in sufficient detail to enable their adequacy to be assessed. In general, one can assume that authors and publishers

---

6 An indicant that some publishers use is fraction of schools' enrollees who qualify for free or reduced lunch payments. Use of this sole indicant is not nearly as adequate as the use of dual indicants described. First, it has limited reliability (i.e., it suffers from great grouping error) because of having only two or three levels rather than continuity of SES data. Second, it is based on only one thing—income in relation to family size; it ignores the important issue of education level of children’s caregivers.
who have gone to the trouble and expense of conducting careful norming research will be willing to follow the dictates of *Standards of Educational and Psychological Testing* (AERA, 1999) and report their efforts. Those who resort to vague, glossed-over descriptions of their norming procedures usually have reason to avoid scrutiny of the details. Other things being equal, such tests should be avoided.

The issue remains of what publishers can best do about districts or schools that, when invited, elect not to participate in norming studies. This is a serious problem because of the likelihood that those who decline to participate may differ in systematic ways from those who are willing to take part. This lack of participation risks systematic bias in the reference group data. First, the problem should be minimized by making participation attractive. The test booklets and scoring service should, of course, be offered to participating schools free of charge. Additional incentives are common, such as discounts or credits on other purchases. Second, publishers can minimize the potential distortion resulting from nonparticipation by replacing schools that choose not to cooperate with others that are randomly selected *from the same strata* (Peterson, Kolen, & Hoover, 1989).

**Thinking CAP Exercise**

Following are two true stories about the norming of well-known tests. First, identify the fault in the procedure. Then speculate, if possible, on the likely direction, if any, of the bias that may have been introduced. That is, indicate whether a typical class’s status would be reported to be too high or too low as a result of the norming defects.

1 A major achievement battery was normed on a sample drawn by use of a random sampling design that was well stratified for SES and other appropriate variables. Selected schools were invited to participate for less charge than normal use of the test would entail.

2 A test was normed solely on white children in and around a particular city and was still being used thirty years later throughout the country.
Key to Thinking CAP Exercise

1 Although the sample selection procedures seem adequate, distortion would be introduced by requiring schools to pay to participate in the study. It is not only necessary to avoid nonrepresentativeness in sample selection, but also to minimize its introduction by schools declining to participate.

One might speculate that poorer schools would be less able to afford the bargain than more affluent ones. This would yield reference-group data that were too high. Therefore, a typical class's scores would appear to be lower (in comparison with this too-high reference group) than they should.

2 The extent to which that particular city is typical of the country is unknown. Restricting the norming to white children is—at best—puzzling. Use of very old normative data is another source of concern; populations undergo generational changes. Taken collectively, these problems seem damning.

(The appropriateness of the thirty-year-old content is another issue—a validity concern. Even when test content hasn’t changed—and in most aptitude and achievement areas, it has—curricular emphases may have.)

Assessing the Relevance of Normative Samples

We have been considering the representativeness of a sample to the population that it represents. Now we turn to the issue of the relevance of a normative-sample to one's interpretive purpose. This topic was treated on pages 351—353 of the text. Before reading on, we urge you to read that material in the text if you have not already done so.

OK, we have seen in the text that a user of a published test might have good reason to want to compare students' performance with a national reference group and/or with a local reference group and/or a gender-specific reference group and/or with a rural, inner-city, reservation, etc., reference group. In each case, the need is to compare an examinee with reference group(s) that is/are relevant to the purpose(s) of the interpretation—i.e., to the decision(s) the test is being used to help make.

We will now elaborate on topics treated only very briefly in the text.
Occupational Norms and Gender-Specific Norms. We have seen that there are, at present, gender differences in certain areas of achievement and aptitude. Even greater gender differences may be found in certain sections of interest inventories. Such scores can be interpreted by use of:

- same-gender reference-group data,
- same-gender and cross-gender data,
- total-group data.

Which method is better? Three answers are appropriate. First, it depends to some extent upon the situation and the purpose of the interpretation. Second, experts don’t entirely agree which is best. Third, sometimes more than one of the above is best.

Let’s consider an example. Suppose eleventh-grader Sina has taken a mechanical aptitude test and is seeking advice from her counselor concerning her aptitude for (a) car mechanics and (b) engineering. Since males tend to receive higher mean scores than females on mechanical aptitude tests, test publishers often provide separate male and female norms.

Suppose Sina’s PR in the reference group of other eleventh-grade females is 99+. In an ideal situation (at Utopia High School), the counselor would also have data with which to compare her score to those received by successful car mechanics when they were in the eleventh grade. Of course most of them would be male, but the occupational group might best be taken as a whole. Suppose her PR in this reference group were 95. Suppose also that her PR in comparison with those received by successful engineers when they were high school juniors was 70. Her performance in comparison with these two groups would certainly be more helpful in judging how her mechanical aptitude compares with persons in those occupations than would her performance in comparison with the group to which she belongs—high school females.

Alas, the counselor probably wouldn’t have the desired occupational reference-group data. However, she or he would have data for high school males. The counselor might reason that because Sina is considering entry into what have traditionally been male-dominated fields, her prognosis would better be assessed by reference to male norms than by female reference-group data. Suppose Sina’s PR in comparison with eleventh-grade boys were 93. In the absence of the more desirable occupational reference-group data, the cross-gender data—along with the female data—would be useful.
Course-Based, Grade-Based, and Age-Based Reference-Group Data. We have seen in the text that published reference-group data may be based on people who are taking a particular course, are in a certain grade, or are a particular age. Which makes the best sense?

On the one hand, achievement in specialized secondary and college courses virtually has to be compared with others who are in the same course. On the other hand, aptitude and achievement in required courses can be referenced to persons of the same grade and/or to persons of the same age. In the United States achievement measures in the parts of the curriculum that are common to all students are most often interpreted with respect to others of the same grade. Aptitude tests, in contrast, are most often interpreted with respect to others of the same chronological age.

These conflicting traditions, of aptitude tests being referenced mainly to same-age peers and achievement tests being referenced mainly to same-grade peers, are unfortunate because use of different bases complicates comparison. For example, suppose Jake is a 12-year-old fourth-grader. If his national PR (NPR) on a reading test is 50 and his NPR on a verbal aptitude is also 50, many educators would conclude that his reading and his verbal aptitude are similar. Not necessarily! If, as is usual, the achievement test provides NPRs based on others of the same grade, then his reading NPR of 50 indicates that he performed about average for fourth graders. If, as is usual, the aptitude test provides NPRs based on others of the same age, then his aptitude NPR of 50 indicates that he performed about average for 12-year-olds. Thus his reading score is not as high as his aptitude score.

Comparison of aptitude and achievement scores—which is used to identify so-called underachievers and to quantify aptitude-achievement discrepancies for identification of students eligible for learning disability programs—is problematic for several reasons. The hazards include:

- the common referencing to different populations (i.e., age and grade discussed in this section),
- the modest reliability of difference scores (which is taken up later in Chapter 14),
- the blurred distinction between aptitude and achievement (which will be addressed in Chapter 16),
- the regression effect (about which an Extension was cited in Chapter 11).

Persons contemplating such test uses should be familiar with these and additional technical pitfalls that make comparisons of aptitude and achievement hazardous.
This difficulty is avoided if both the achievement and the aptitude test are referenced the same way. Whenever one wishes—in spite of the several hazards of doing so—to compare aptitude with achievement, it is important to ensure that the reference-group data are either both age based or are both grade based. This rule is especially important if the student is significantly underaged or overaged for his or her grade.

**Comparability of Reference Groups.** Another issue involved in comparing scores of different tests concerns the comparability of their normative samples. Suppose that Vincent obtained the $NPR$ of 75 on a reading test and of 50 on a math test. In judging whether he is stronger in reading than math, two issues must be considered. One concerns whether the difference is large enough to merit confidence that it was not a mere chance event (a topic addressed later in Chapter 14 and beyond). The other issue is whether the reference-group samples for the two tests were equally capable. If one test had been normed with a brighter sample than the other, then the comparison would be distorted.

Ideally, we would like for the reading and math tests to have been referenced to the same sample. Then comparability between the two $PR$s would be certain to be sound. As will be seen in Chapter 16, a major advantage to the use of an achievement test battery, which consists of several different subtests measuring various subjects, is that the whole battery is normed on the same sample. With batteries, comparability of normative data is assured.

If the two tests have been normed on different samples, then comparability merits intense consideration. Descriptions of the referencing studies should receive even greater scrutiny than usual in order to judge whether interpretation of both tests is referenced to samples that represent the same population.