ED 058 297 TM 001 012

AUTHOR	Ricks, James H., Jr.
TITLE	Local Norms--When and Why.
INSTITUTION	Psychological Corp., New York, N.Y.
REPORT NO	R-58
PUB DATE	Aug 71
NOTE	6p.
AVAILABLE FROM	The Psychological Corporation, Test Division, 304 East 45th Street, New York, New York 10017 (no charge)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Criterion Referenced Tests; Decision Making; Educational Counseling; *Group Norms; National Norms; Norm Referenced Tests; *Norms; *Predictive Measurement; Psychometrics; Standardized Tests; *Student Evaluation; *Test Results

ABSTRACT

A brief examination of norms, local and general, with emphasis on the former. The difficulties of using aptitude in counseling are discussed. Some data from the Differential Aptitude Tests is provided. (DLG)

THE PSYCHOLOGICAL CORPORATION INCORPORATED IN 1921 ®

Test Service Bulletin

THE PSYCHOLOGICAL CORPORATION
ROGER T. LENNON, *President*

August, 1971
No. 58

LOCAL NORMS—WHEN AND WHY

JAMES H. RICKS, JR.

TEST users often find themselves envying bankers and surveyors and people like that whose units are dollars or miles or other nice, definite and fixed ones. But how many 1940 dollars will your next car cost, and how do you measure what it takes to travel from New York to San Francisco, or from your home in the suburbs to your office in the city?

Personnel men, guidance counselors, psychologists and others who use tests may derive a small amount of wry comfort from the fact that others are finding *their* measuring units a bit slippery too. A dollar seems more like a percentile than before when we need a modifier (U.S. or Canadian or 1940) to give it specific meaning. Miles are less and less meaningful units of measure in a day of jet travel and monumental traffic jams. ("Agomins" or "Minutes of Agony" have been suggested by Russell Baker as the basic unit of sensible modern distance measure. He suggests that "Twelve and a half miles from downtown is a meaningless measure of the distance" compared with fifty-five minutes of agony in snarled traffic and exhaust-laden air.)

It's an old story to us, of course. People in the world of psychometrics have known so long that they tend to take for granted the fact that a percentile or a stanine or a mental age or I.Q. has no meaning apart from the observed collection of people's scores (norms) on which it is based. We even have become rather careful to specify what norms are being used when we think about a score. And this, in turn, raises the question as to what norms *ought* to be used for the purpose we have in mind at the moment.

Most readers of this BULLETIN realize that with few exceptions test scores have little or no meaning until we have an array or distribution of scores by some identified group with which to compare the score of the person in whom we are interested.[1] For the purposes of this discussion, however, there should be agreement as to what norms are. The dictionary says that "norm" is derived from the Latin *norma* meaning a carpenter's or mason's square and that in general it means a rule; an accepted standard; an authoritative model, pattern or type. One dictionary's definition of the statistical sense of "norm" is, "A quantitative standard determined by the average, median, or other measure of the central tendency among the varying individuals of a type of species."

That is the general definition. In this discussion, when we say "norms" we shall mean the percentile or standard score conversions derived from the distribution of scores earned by an identified group. These score conversions are used to help us make statements about the performance of an individual who may fairly be compared with the group. Or, in some cases, to make statements about

[1] No, Bligsby, it's not true that we haven't heard about criterion-referenced testing. Psych Corp published its first criterion-referenced tests in 1946 and 1947, and they're still in our catalog. (Ask us what they were if you can't tell.) For more on this point, see the last page of this article.

TEST SERVICE BULLETIN

Published from time to time by the Test Division of The Psychological Corporation *in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of interest. Address communications to*

304 East 45th Street
New York, N.Y. 10017

another group that may fairly be compared with the first.

"Local" and "general" as applied to norms are relative terms. In two of the most common settings for test usage, we may think of an order such as

IN EDUCATION	IN INDUSTRY
classroom	machine shop
building	plant
system	company
state or region	steel industry
nation	all industry

A unit occurring earlier in each list is, of course, "local" in comparison with any unit occurring later. In this discussion, "local" will usually refer to one of the first two or three levels, and our illustrations will be principally from the world of education and guidance rather than that of industry and selection.

The fact that we have the possibility of a proliferation of norms for any test is one of the blessings—some might say one of the curses—of modern computer technology. Preparing local norms for the tests we use or working up norms on a considerable variety of groups used to be a laborious undertaking. Only for a little more than a decade have computer-based scoring services made it possible to obtain local frequency distributions economically. Now that it *is* rather easy and inexpensive to obtain local norms, we have to deal more frequently and more seriously with the question of when it is worthwhile to prepare them and when it is wise to use them.

We shall begin by presenting a couple of propositions. The reader will recognize at once that the propositions are debatable and that some exceptions must be taken to them. Debates on this topic all too often have taken place in the absence of real, concrete information and data. Therefore we shall set forth some actual data and illustrations in the hope of justifying the propositions or, at least, illuminating the debate.

The first proposition will be that **local norms often make sense when we are looking back at what a group or an individual has done but are less likely to make sense when we are looking ahead to what they may be expected to do in the future.** Roughly this corresponds to endorsing the use of local norms for achievement tests but opposing the use of local norms for aptitude measures. Of course, we all know that it is possible to use a so-called achievement test as a predictor of future performance or an indicator of readiness, and to use a so-called aptitude test to measure accomplishment or proficiency that has already been gained. This proposition has to do with the function of looking backward or looking ahead rather than with the label or title that the author may have chosen for the test.[2]

The second proposition will be that **local norms are frequently useful for administrative purposes, but are less often valuable for counseling purposes.** The administrative purposes may be forward-looking as in the instance of sectioning a class for the coming year. Similarly some counseling uses of tests may be backward-looking—it often is appropriate to review with a counselee the use he has made of his educational opportunities and experiences to date. It probably will surprise no one that these propositions are not altogether simple and clear-cut. With balanced consideration, however, they can serve as

[2] And here we have to begin acknowledging the exceptions. If all one wants to do is predict Johnny's chances of passing Spanish 1 in Elmwood Junior High School next year, local norms may indeed be useful. The two propositions are intended, however, in the more general guidance and counseling context.

a sound guide in the general run of situations where special considerations are not involved.

It is not often that test users have a real opportunity to see how much difference local norms can make. It may be worthwhile to look at some actual data of the kind that one finds at the back of a publisher's file when cleaning out the accumulations of a decade or two. Specifically, consider an unplanned, opportunistic collection of local norms for the *Differential Aptitude Tests*, prepared by or for a variety of groups.

The groups include several college freshman classes, a highly selective New England prep school, a number of other private preparatory schools both religious and secular, several vocational high school and technical high school groups, and a number of public high schools. The entire body of data would be quite unwieldy and indigestible, so we shall approach it by pulling out of the body a few bones to show the shape of the skeleton and then patching on a little flesh here and there to illustrate what happens when local norms are actually used.

To begin with, there were 38 major groups in this accumulation, that is, 38 areas or institutions for which normative data were conveniently accessible. Treating each sex and each grade separately, and in a few instances treating separately the data from two or more different years of testing, we have 126 subgroups in these 38 major groups.

By way of getting down to the bare bones, we shall ignore all of the information about each group except its average (mean or median) scores on the eight tests of the *DAT* battery—or the seven tests or the six tests and so on in those instances where something less than the entire battery was given. The percentile equivalent on national norms of each of these mean or median scores is presented in Table 1, in the form of a frequency distribution.

Under "Frequency," the first number in the body of the table indicates that there were two instances of groups with average scores on one test (or one group on two tests) that would convert to the 99th percentile on national norms for the appropriate grade and sex. Reading down the column, there were three more instances in which average scores were equivalent to the 95th percentile on national norms, ten at the 90th percentile, nineteen averages at the 85th percentile, and so on down to one instance of an average score at the 5th percentile.

TABLE 1

Distribution of National Percentile Equivalents of Average Scores of Special Norms Groups on the Differential Aptitude Tests[a]

Based on 126 Groups (by Grade and Sex) from 38 Areas or Institutions

Percentile	Frequency (No. of Groups)	Percent	Cumulative Frequency	Cumulative Percent
99	2	—	956	100
97	—	—	954	99.8
95	3	—	954	99.8
90	10	1	951	99.5
85	19	2	941	98
80	16	2	922	96
75	28	3	906	95
70	40	4	878	92
65	66	7	838	88
60	98	10	772	81
55	103	11	674	70
50	148	15	571	60
45	107	11	423	44
40	113	12	316	33
35	77	8	203	21
30	46	5	126	13
25	30	3	80	8
20	19	2	50	5
15	23	2	31	3
10	7	1	8	1
5	1	—	1	—
3	—	—	—	—
1	—	—	—	—
N	956[b]			

[a] Data from testings during 1947–1960, not included in Manual. See Appendix for additional data.

[b] The product of the number of groups times the number of tests; 956 instead of 1008 because a few groups did not take all eight tests.

Remember: Unlike most percentile tables this one presents not scores of individuals, but averages of groups. The data demonstrate impressively that groups do differ; it is clear that local frequency distributions often will provide a basis of comparison very different indeed from that provided by national norms.[3]

[3] Some readers will want to know to what extent special kinds of schools may have contributed to the rather surprising range of averages shown. In the Appendix on page 6 are presented the data for 40 public high school groups on all eight tests, and then data on two of the tests for three kinds of special groups: 19 private or independent schools, 8 vocational and technical high schools, and freshmen in 7 colleges or junior colleges.

There can be little doubt that such local distributions can be both interesting and useful to researchers and to school administrators (and to the latter in some of their forward-looking, planning functions as well as in their appraisal of the past). Illustrative instances and a discussion of such uses of local distributions are to be found in TEST SERVICE 'BULLETIN NO. 41, "Human Resources and the Aptitude Inventory."

Now let us look at what happens when we shift norms in the case of a real student or two. William Swan and James Wallace are two ninth-grade boys whose cases are reported in the *DAT* casebook, *Counseling from Profiles.*

Bill represents one of those happy instances in which it was the counselor's pleasurable duty to point out to him and to his parents that his goals were rather low in the light of his measured talents. Jim, on the other hand, is the kind of case that the unhappy counselor probably feels he encounters much more frequently—the problem of preparing a student or parents with apparently unrealistically high ambitions for the kinds of learning and occupations from which he can really profit and in which he can make a self-realizing contribution.

The profiles of these two boys on each of four different sets of norms appear below and on the next page. In each case, No. 1 is the profile on national norms as presented in the casebook. The next profile, No. 2, represents the same scores plotted according to California norms as of some years ago. There are some minor differences in the patterns on the state norms as compared with the national norms, but not anything that would lead the counselor to offer very different advice or the boy or his parents to make very different plans or decisions.

But now look at the third and fourth charts. No. 3 shows how each set of scores would look when plotted on the norms for boys in segregated rural Negro schools in a southern state something over ten years ago. No. 4 shows how each boy would look if compared at about that time with the students at a New England preparatory school which maintains a highly selective admissions policy, exacts high performance standards from its students, and sends nearly every one of them on to a "good" college.

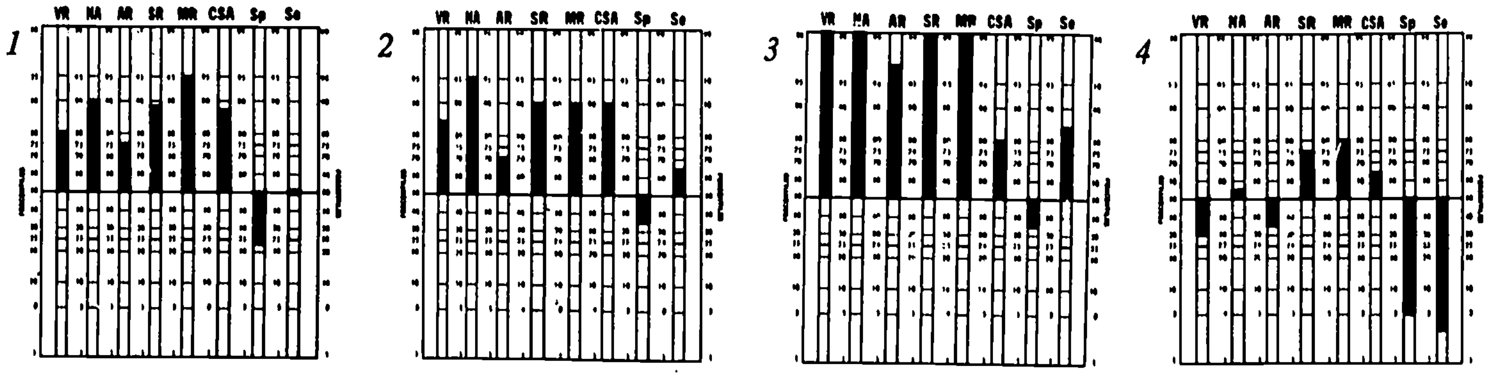
If we look at Bill's profile No. 4, our very able boy appears from mediocre to very low on the Verbal Reasoning and Language Usage tests that are usually the most effective predictors of scholastic success. If we look at Jim's profile No. 3 without considering the norm group, we might send him away from the counseling interview with a quite unrealistically optimistic picture of how well he is likely to do and what sort of program he is likely to profit from in the years of school ahead of him.

So it seems that if we compare each boy with state (one state, at least) rather than national norms it doesn't make much difference—we might just as well present the national norms profile and let it go at that. And when we find a norms table that *does* make a real difference, the effect of its use by the careless or unsophisticated may be to mislead rather than to provide a wise guide—at least so far as general educational and vocational counseling for his future are concerned.

Remember, the first proposition was that while local norms or distributions may be quite useful in looking back at past performance they will be less so in looking ahead and making predictions.[4] When a student has

[4] An exception occurs, as noted earlier, when we know that a particular future course involves competition only or mainly with the group of which the counselee already is a member. (Expectancy tables are "norms" too.)

Profiles of Bill S.



come through the ninth grade in a particular school, we know that he has been a member of that group and may be compared with them on past performance. But it would be unreasonably limiting to give him advice about his future that requires the assumption that he will always 'emain a member of this group. And this leads to the second proposition: Local norms and distributions are more likely to be useful for administrative and research purposes than when they are used for counseling purposes.

As noted when they were first presented, these two propositions do *not* relieve us of the burden of thinking about what kind of norms to seek or which table of available norms to use. We cannot call the propositions *rules*—it is far too easy to point out cases and situations to which their application would be foolish. They will have served their purpose if they shake up just a bit the kind of automatic thinking that seems all too often manifest, the tendency to accept unquestioningly the table of norms that is either most fashionable or most conveniently available. (See also TEST SERVICE BULLETIN No. 39, "Norms Must Be Relevant." May, 1950.)

Before leaving the topic, we must not fail to recognize the increased attention currently being directed to mastery tests—"criterion-referenced tests" in the now-fashionable term, or "the normless wonders" as one city school superintendent called them. They have an important role to play in the improvement of instruction and learning—even in learning without instruction.

Their time is here. It is one thing for a psychometric in-group to be discussing the uses of criterion-referenced tests as compared with norm-referenced tests. It is quite another for the head of the measurement division of the American Educational Research Association to be telling a Congressional committee that

> "... a new approach to measurement must be used which is capable of determining what a learner can do, regardless of whether he can do it better than another learner."[5]

But. This new emphasis on mastery tests will not relieve us of concern with norms. For too many of the essential purposes and uses of tests, the kind of norm-based unit so familiar to all remains the only meaningful way of recording or reporting test outcomes. In addition, it seems assured that norms will be acquired for many, perhaps most, of the tests built as criterion-referenced measures, for the sake of the additional useful information that data on the performance of others can furnish.

In an illuminating paper on norms presented in 1963, Roger Lennon pointed out that

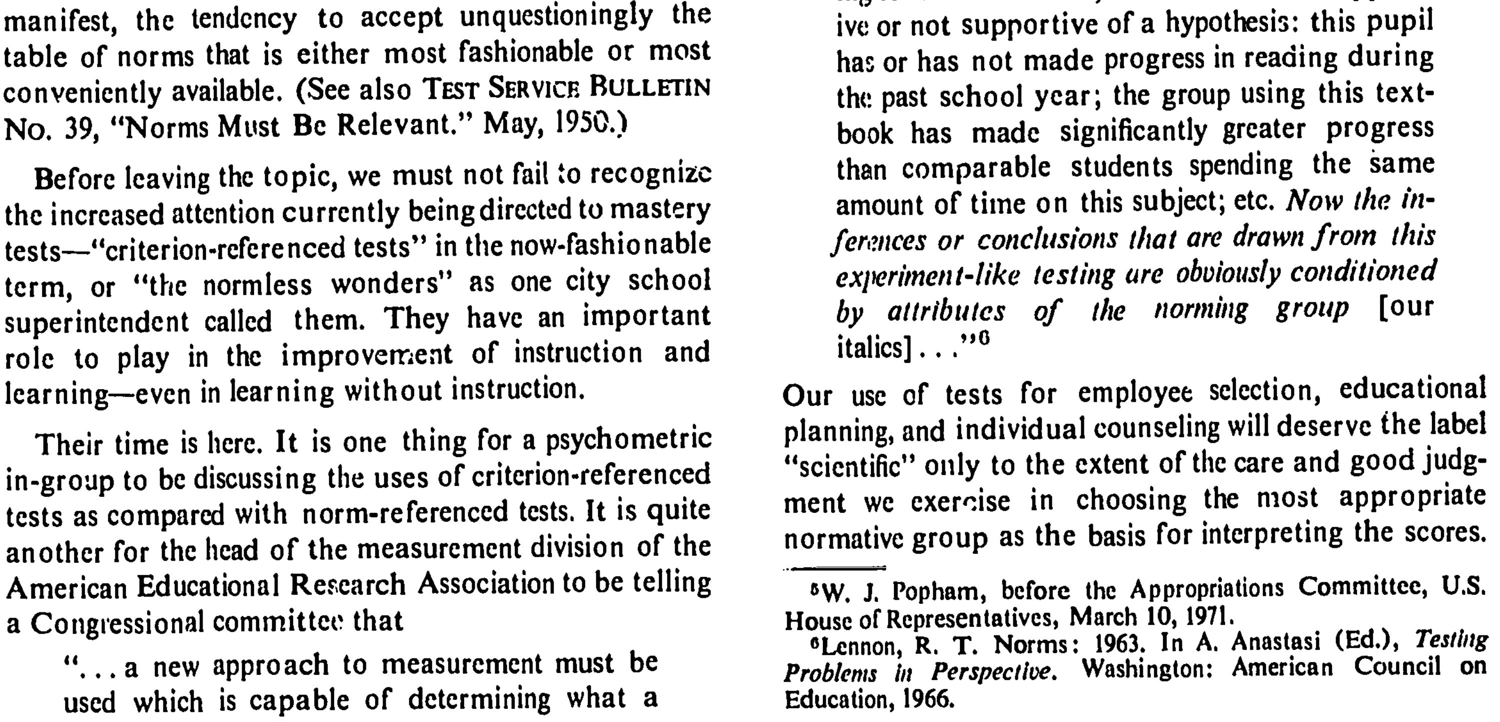
> "... the administration of a test to an individual or a group can, in most instances, be thought of as akin to the conduct of a scientific experiment. Performance on a test, when interpreted according to suitable norms, serves as evidence supportive or not supportive of a hypothesis: this pupil has or has not made progress in reading during the past school year; the group using this textbook has made significantly greater progress than comparable students spending the same amount of time on this subject; etc. *Now the inferences or conclusions that are drawn from this experiment-like testing are obviously conditioned by attributes of the norming group* [our italics]..."[6]

Our use of tests for employee selection, educational planning, and individual counseling will deserve the label "scientific" only to the extent of the care and good judgment we exercise in choosing the most appropriate normative group as the basis for interpreting the scores.

[5] W. J. Popham, before the Appropriations Committee, U.S. House of Representatives, March 10, 1971.

[6] Lennon, R. T. Norms: 1963. In A. Anastasi (Ed.), *Testing Problems in Perspective*. Washington: American Council on Education, 1966.

Profiles of Jim W.



APPENDIX

Distribution of National Percentile Equivalents of Average Scores of Special Norms Groups on the Differential Aptitude Tests, by Type of School and Subtest

	State-Representative Puhlic High School Groups								Private High Schools		Vocational and Technical High Schools		Colleges and Junior Colleges	
Percentile	VR	NA	AR	SR	MR	CSA	Sp	Se	VR	SR	VR	SR	VR	SR
99														
97														
95													1	
90									3				—	
85						1			—				1	1
80						2			1				—	—
75			1			1			2	1			—	—
70		3	1	3	1	1		1	1	1			—	—
65	3	1	3	2	4	1	1	3	5	8			—	—
60	3	5	4	4	4	8	1	2	1	4		3	3	—
55	3	1	12	12	12	4	5	—	2	—		—	—	1
50	5	5	12	12	12	9	2	2	3	1		1	1	1
45	10	13	1	1	2	8	9	8	—	—		2	1	—
40	10	8	2	2	1	4	14	10	1		1	1		1
35	2	—	—	—	—		6	6			2	1		2
30	—	—	—	—	—			2			3			1
25	—	—	1	1	—			1			2			
20	—	—	1	1	2			1						
15	1	4	2	2	2			2						
10	3													
5														
3														
1														
N	40	40	40	40	40	39	38	38	19	15	8	8	7	7

Note.—There are four sections in the table, running from left to right. The first section presents eight columns of data—all eight tests. The next three sections present two columns each—two tests.

Of the 92 groups representing public high schools, 40 were groups that were reasonably representative of a state or a substantial part of a state. For these 40 groups, the first section of the table presents a frequency distribution of the percentile equivalents of the mean or median scores. Under VR, the first number in the body of the table indicates that there were three groups with mean scores on the Verbal Reasoning test that would convert to the 65th percentile on national norms for the appropriate grade and sex. In the second column, headed NA, we see that there were three groups whose average score on the Numerical Ability test was equivalent to the 70th percentile on national norms, one group at the 65th percentile, five groups that averaged at the 60th percentile, and so on down to four groups with average scores at the 15th percentile.

Rather than burden the reader with similar data for each kind of subgroup, the table next presents these distributions for only two tests (the Verbal Reasoning and the Space Relations, chosen because they are about as different as any two tests in the *DAT* battery) for three kinds of groups: the private school group, the vocational high school and technical high school group, and the group of freshmen in colleges, engineering schools, and junior colleges.