Communications

On "Different IQ's for the Same Individual Associated with Different Intelligence Tests"

In his small, clinically oriented study, Dreger (*1*) illustrates the point that even when a given subject is tested several times by the same examiner, using the intelligence scales with which this examiner feels most confident, the individual's IQ scores probably vary considerably. He concluded that "whereas group means on different tests of intelligence *may* not differ except by chance from one another, individual's IQ's may differ widely *and significantly* from one another on different tests" (p. 595, italics mine). Both portions of this statement may be misleading. By private correspondence, Dr. Dreger has assured me that he did not ignore the considerable body of literature which shows rather conclusively that the mean IQ score for a group of individuals usually varies markedly from test to test (for example, see *2–8*). Rather, he apparently meant that, even though the test means do not differ significantly for a certain group of testees, there may still be idiosyncrasies that result in greater-than-chance discrepancies between the highest and lowest obtained IQ scores of some individuals. This is equivalent to saying that testees have different "true" scores on the two tests which in the long run sum over individuals to approximately the same figures.

If only two scores on the same two tests were available for each testee, this true-score comparison would be tantamount to saying that the individuals by tests interaction is significant. But with a randomized block design, the $i \times t$ mean square serves as the denominator of the two F tests, and there is no "error" term available for testing its significance. This same situation prevails no matter how many individuals and/or tests are used. A design involving retesting or comparable forms of the same test is needed to provide an error estimate for the $i \times t$ interaction. I plan to discuss this problem elsewhere.

Furthermore, Dreger's "assumption that the tests which yielded the highest and lowest IQ's for each subject could have been by chance the two tests administered if only two were administered in any clinic" (p. 594) is an untenable basis for setting up a two-way classification (individuals by highest-lowest IQ) and running a t test of the difference between the correlated means, because it capitalizes heavily upon chance fluctuations to accentuate improperly the apparent significance of the difference. This is not the same as comparing individual means in the analysis of variance, with adjusted probabilities, after the overall F test for treatment means has been found significant.

Because of marked uncontrolled sources of heterogeneity in Dreger's study, coupled with the small number of subjects upon which it is based, the application of any type of statistics appears undesirable. Instead, I recommend that the 39 scores be viewed only as an illustration, certainly not as proof, of variability under particular clinical conditions.

Julian C. Stanley

Department of Education
University of Wisconsin

References and Notes

1. R. M. Dreger, *Science* **118**, 594 (1953).
2. J. C. Stanley, *Rev. Educ. Res.* **23**, 11 (1953).
3. S. W. Gellerman, *J. Consult. Psychol.* **16**, 127 (1952).
4. A. S. Elonen, *Psychol. Mongr.* **306**, 35 (1949).
5. M. E. Hamilton, *J. Consult. Psychol.* **13**, 44 (1499).
6. E. Z. Johnson, *J. Clin. Psychol.* **8**, 298 (1952).
7. G. Manolakes and W. D. Sheldon, *Educ. & Psychol. Meas.* **12**, 105 (1952).
8. W. G. Heil and A. M. Horn, "A Comparative Study of the Data for Five Different Intelligence Tests Administered to 284 Twelfth Grade Students at South Gate High School, Los Angeles." Los Angeles City School Districts, Curriculum Division, 1950 (mimeographed).

Received January 11, 1954.

A recent article by Dreger (*1*) ends with the sentence, "Therefore, it can be concluded that whereas group means on different tests of intelligence may not differ except by chance from one another, individual's IQ's may differ widely and significantly from one another on different tests."

Dreger's data do not establish this conclusion, because, as he seems to have suspected, he concocted an interesting way of misinterpreting results obtained from a perfectly respectable statistical method. Ten

Table 1. IQ's of 10 children, each tested four times on the Stanford-Binet (the form used is shown below the IQ).

Subject	Test 1		Test 2		Test 3		Test 4	
	CA	IQ	CA	IQ	CA	IQ	CA	IQ
1	37	111 L	50	106 L	64	106 M	76	118 L
2	73	110 L	85	112 L	97	109 M	109	116 L
3	30	123 L	42	129 M	54	113 L	60	117 M
4	35	109 L	45	107 L	55	129 M	62	123 L
5	37	97 L	51	106 L	53	104 M	57	100 M
6	35	143 L	43	142 M	54	131 L	61	125 M
7	31	126 L	44	127 L	55	125 M	62	115 L
8	56	141 L	57	151 M	60	143 L	92	146 M
9	30	110 L	36	106 L	48	121 L	50	122 M
10	84	155 L	95	179 L	107	157 M	119	173 L

children were given four (or, in one case, three) different tests, each of which yielded a score that could be converted into an IQ. Two columns were then set up, in one of which the child's highest IQ was entered; in the other, his lowest. Dreger then misused statistical methods to determine that the highest IQ's are higher than the lowest IQ's. Perhaps this is obvious to the reader; if not, our Tables 1 and 2 should clarify the matter.

The data in Table 1 were supplied to me by Stott (*2*) from records at the Merrill-Palmer School in Detroit. These cases were simply taken in order from among those who had been tested four times with the Stanford-Binet, the best known and the most reliable of the tests used by Dreger.

TABLE 2. Comparison of lowest and highest IQ's.

Lowest IQ	Highest IQ	Highest minus lowest
106	118	+12
109	116	+ 7
113	129	+16
107	129	+22
97	106	+ 9
125	143	+18
115	127	+12
141	151	+10
106	122	+16
155	179	+24

The highest and lowest IQ's and their differences are shown for each subject in Table 2. All differences are, of necessity, positive. Their mean is 14.6, their standard deviation is 5.3, the standard error of the mean is 1.8, and the t ratio of 8.1 is, as was Dreger's, significant beyond the 0.001 level. However, this does *not* prove that "IQ's may differ widely and significantly from one another on different tests." All that is proved is that in some or in all cases, a child may get at least two different IQ's if he is tested four times. This is true without regard to whether he is tested on four different tests or on the same test. It is also true of any kind of measurement. Thus, we would get exactly the same results if we simply weighed 10 children on four different days (using either the same or different scales).

Dreger's little study has, thus, contributed nothing to the problem he tackled, the constancy of the IQ. He has, however, well illustrated what every psychologist knows: A child's IQ will be similar, but not identical, on successive testings. This is true whether one or more tests are used.

ALBERT K. KURTZ

Department of Psychology
University of Florida

References

1. R. M. Dreger. *Science* **118**, 594 (1953).
2. L. H. Stott. Personal communication to the writer. Dec. 10, 1953.

Received January 11, 1954.

JULIAN C. STANLEY'S comment on my article helps to correct some misimpressions a brief article often gives. There are several points, however, that may need clarification in Dr. Stanley's letter.

The conclusion of my article, to which reference is made, was not intended to take in all the many experiments showing significant differences among IQ tests for the same individuals. These latter are mainly on group tests, of course, in contrast to the more carefully administered individual tests of my experiment. My conclusion was not based only on a comparison of highest and lowest IQ's for the same subject, but mostly on an analysis of variance and the relative lack of correlation among the tests. When I said that for the entire group IQ's from different tests may not differ except by chance, I merely intended to make a cautious introduction to the major clause based on the evidence immediately at hand, to the effect that individual IQ's may differ, *even though* group means may not, as they did not in this case.

I agree with Dr. Stanley's analysis of treatments by subjects design, what he calls $i \times t$ design. His discussion is relevant; though, from a strictly statistical standpoint, a mixed design of some sort might be better. From such a standpoint, a design unconfounding examiner and examinee variables would be best, but this type of design would circumvent the problem of my experiment.

In part, Dr. Stanley's discussion loses sight of the problem which was: Do IQ's obtained from the same individuals *in a clinical setting* on different tests differ and significantly? A careful reading of my article will reveal that the entire situation is couched in terms of clinics. I can, of course, think of several designs that could eliminate examiner variance from the error term for testing subject effects. Confounding was recognized and mentioned both in the report to the Florida Psychological Association and in the article. But to quote from the article: "The design was intended to duplicate the actual clinical situation where one examiner gives several tests to the same person." If the design of an experiment eliminates the conditions generating the problem it is intended to answer, the experiment may answer some question but not the experimental one.

In connection with the "untenable basis for setting up a two-way classification," a quotation from my original report to the Florida Psychological Association seems in order:

> The comparison here seems the most indefensible statistically, for the extremes are compared, rather than any two IQ's taken at random. Yet both from logical and statistical standpoints a rationale (which is not necessarily a rationalization) can be supplied. Statistically one may say that the test meets the demands for a significance level. Adopting Fisher's criterion of one in $\frac{1}{2}n(n-1) \cdot 20$ as the level of significance, the obtained level, beyond .001, is considerably below the required .008. Logically, the fact that in clinical practice the choice of any two of these tests could be made and could yield these extremes of IQ's justifies to an extent assigning each individual a High and a Low score for comparison.

I did point out in the *Science* article that the procedure was questionable, but limits of space prevented elaboration. Dr. Stanley is correct in stating that my procedure is not the same as a t test following an F test.

Since I utilized statistics in planning and carrying out my study, I naturally disagree with Dr. Stanley's final paragraph. Various sources of heterogeneity were controlled experimentally, the training of examiners, the time lapse between testings, age distribution, representativeness of tests, and two other variables not mentioned in my article, socio-economic background to some extent, and "normality" of emotional behavior. From a methodological standpoint, I believe that all experiments are only illustrations. The only way I know further to show whether my experiment constitutes what might be called "proof" is to have it repeated, perhaps with a more efficient design, as long as the design in the interest of efficiency does not tackle a different problem from that set up here. *Whether or not* the means of groups differed, I should predict that the same general result would obtain, that individuals would differ by more than chance within themselves on different IQ tests.

Dr. Albert K. Kurtz' remarks are partially answered by the foregoing statements. His last paragraph and his Table 1 demonstrate that apparently I did not make the experimental problem clear: "Dreger's little study has, thus, contributed nothing to the problem he tackled: the constancy of the IQ." I was not concerned with that particular problem, which I did mention in passing in the first paragraph of my article. Instead, the question asked was: "What about the constancy of the same individual's IQ as reported on *different tests at approximately the same time?*" (Italics unfortunately are not in the original.)

Kurtz' Table 1 shows IQ's on the *same* test (assuming as we both do evidently that Binet L and M are equivalent) across periods ranging from 20 to 39 months. In the life of a child, the period from 3 to 6 years is a very long time, or even from 7 to 10 (Subject 10). As is apparent, Kurtz' problem and mine are different. An analysis of variance reveals, as in my case, that his Binet tests do not differ significantly from one another (as tested against interaction mean square). But aside from the fact that he is citing the same test used on different occasions, his procedure is not the same as mine. My experiment was set up so that a between-examiner variance would not inflate the differences among tests. I presume that Merrill-Palmer examiners would be different from time to time or randomized by happenstance among subjects and times. Kurtz' results are a tribute to the excellence of the Binet test but are not directly comparable to the results of an experiment employing a different procedure.

One comparison may be made between Kurtz' Table 1 and my Tables 1 and 2. A rank correlation of Kurtz' data, using Kendall's W, yields a coefficient of .81, which by a chi square test (*1*) is significant beyond the .001 point. Such a result might have occurred if only four tests had been used in my experiment, so that all ten subjects could have entered into the rank correlation. With six subjects, however, the correlation was not significant on three tests. With eight subjects on three tests, one different from the last, chi square is just below significance at the .05 point.

Rather than engage in this sometimes fruitless interchange on my experiment, I should rather repeat the experiment. Because of administrative changes, I am not at present in a position to do so, although I expect to be in such a position again. I hope somone will repeat it. If my results are not verified—on the same problem, not a different one—I shall be happy to acknowledge publicly that what I called a "limited answer to the question" is more limited than I am ready now to admit.

RALPH MASON DREGER

Department of Psychology
Florida State University

Reference

1. M. G. Kendall, *Rank Correlation Methods.* (Charles Griffin, London, 1948), p. 84.

Received February 15, 1954.

Chemical and Physical Characteristics of Delaware River Water from Trenton, N. J., to Marcus Hook, Pa., 1949–52

The Delaware River is the principal source of water for many industries and municipal water supplies in the reach of the river from Trenton, N. J., to Marcus Hook, Pa., and both industry and municipalities use it for disposal of their wastes.

Interest in the quality of the water in the Lower Delaware was manifested in the latter part of 1930 when the natural flow of the Delaware River was unusually low and the salinity of the river increased markedly. Officials of industries that were affected initiated salinity investigations of the stream. A daily sampling program by the U.S. Geological Survey was started in 1944, at Morrisville, Pa.

On the basis of measurements during the period between Aug. 1949 and Dec. 1952, we observed that the mineral content of the water increases from Trenton to Marcus Hook. During protracted periods of low flow (which occurs only during the late summer months) salt water moves up the river along the river bottom and is partially mixed with the river water as a result of currents from tidal action and other factors. This saline invasion causes chloride content to increase sharply at Eddystone and at Marcus Hook, and near its mouth the river water tends to approach the composition of sea water. During these periods, higher concentrations of dissolved solids are observed at the bottom of the river than near the surface.

During normal flow, there is more calcium than magnesium and more sulfate than chloride in the water. This relationship is reversed when the downstream flow is low and ocean water mixes with the river water. At such times, we observed dissolved solids concentrations as much as 4150 ppm at Mar-