REPORT RESUMES

ED 017 810 AC 002 021

SAMPLE-FREE TEST CALIBRATION AND PERSON MEASUREMENT. PAPER PRESENTED AT THE NATIONAL SEMINAR ON ADULT EDUCATION RESEARCH (CHICAGO, FEBRUARY 11-13, 1968).

BY- WRIGHT, BENJAMIN D.

PUB DATE 28 OCT 67

EDRS PRICE MF-$0.25 HC-$0.96 22P.

DESCRIPTORS- *MEASUREMENT TECHNIQUES, *TESTING, *INTELLIGENCE, *TEST RELIABILITY, MODELS, ITEM ANALYSIS, TEST RESULTS, GOERG RASCH MEASUREMENT MODEL,

OBJECTIVITY IN MENTAL TESTING REQUIRES THAT TEST CALIBRATION BE INDEPENDENT OF WHICH PERSONS ARE USED FOR THE CALIBRATION AND THAT PERSON MEASUREMENT BE INDEPENDENT OF WHICH ITEMS ARE USED FOR THE MEASUREMENT. PRESENT PRACTICE IS NOT OBJECTIVE, BUT COULD BE SO, AS SHOWN BY THE EXAMPLE HERE PRESENTED. DATA COME FROM THE RESPONSES OF 976 LAW STUDENTS TO 48 READING COMPREHENSION ITEMS ON THE LAW SCHOOL ADMISSIONS TEST. THE POSSIBILITY OF PERSON FREE TEST CALIBRATION IS DEMONSTRATED BY SHOWING THAT A CALIBRATION BASED ON THE RESPONSES OF A DUMB GROUP OF STUDENTS CAN BE NEARLY IDENTICAL WITH ONE BASED ON A SMART GROUP. THE POSSIBILITY OF ITEM FREE PERSON MEASUREMENT IS DEMONSTRATED BY SHOWING THAT ABILITY ESTIMATES MADE FROM SCORES ON AN EASY TEST CAN BE STATISTICALLY EQUIVALENT TO THOSE MADE FROM A HARD TEST. THE MEASUREMENT MODEL WHICH MAKES THIS OBJECTIVITY POSSIBLE WAS DEVELOPED BY GEORG RASCH. IN THIS MODEL THE ODDS OF SUCCESS ON A TEST ITEM ARE HYPOTHESIZED TO BE GIVEN BY THE PRODUCT OF THE PERSON'S ABILITY AND THE ITEM'S EASINESS. IN ORDER TO FIT THIS MODEL ITEMS MUST BE CHOSEN OR CONSTRUCTED TO HAVE SIMILAR DISCRIMINATION. THE RESULTING MEASURES OF PERSON ABILITY AND ITEM EASINESS ARE ON A RATIO SCALE WITH A NATURAL ZERO AND A DEFINABLE UNIT. THIS PAPER WAS PRESENTED AT THE NATIONAL SEMINAR ON ADULT EDUCATION RESEARCH, CHICAGO, FEBRUARY 11-13, 1968. (AUTHOR/RT)

October [illegible], 1967

SAMPLE-FREE TEST CALIBRATION AND PERSON MEASUREMENT

Benjamin D. Wright
Professor of Education
University of Chicago

Ever since I was old enough to argue with my pals over who had the best I.Q., I say "best" because some thought 100 was perfect and 60 was passing, I have been puzzled by mental measurement. Even that noble achievement, 100 per cent, is ambiguous. One hundred may signify the welcome news that we are smart. Or it may just mean the test is easy. Some students pray for easier tests to make them smarter.

We all know one way a test score can more or less be used. If you are willing to accept as a whole the set of items making up a standardized test, you can get a relative measure of ability. If your performance puts you at the eightieth percentile among college men, you'll know where you stand. Or will you? The same score will also put you at the eighty-fifth percentile among college women, at the ninetieth percentile among high school seniors and above the ninety-ninth percentile among high school juniors. Your ability will depend not only on which items you take but on who you are and the company you keep.

The truth is that a scientific study of changes in ability, of mental development, is far beyond our feeble capacities to make measurements. How can we possibly obtain quantitative answers to questions like: How much does reading comprehension increase in the first three years of school? What proportion of ability is native and what learned? or, What proportion of mature ability is achieved by each year of childhood?

I hope I am reminding you of some problems which afflict present practice in mental measurement. The scales on which ability is measured are elusive and slippery. They have no logical zero point and no regular unit. Their meaning and estimated quality depend upon the specific set of items actually standardized and the particular ability distribution of the children who happened to appear in the standardizing sample.

If all of a specified set of items have been tried by a child you wish to measure, then all you can obtain is his percentile position among whatever groups of children were used to standardize the test. How now do you interpret this measure beyond the confines of that set of items and those groups of children? Change the children and you have a new yardstick. Change the items and you have a new yardstick. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test-takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardsticks?

Objectivity in Mental Measurement

The growth of science depends on the development of objective methods for transforming observation into measurement The physical sciences are a good example. Their hallmark is the development of methods for measuring which are specific to the measurement intended and independent of variation in the other characteristics of the objects measured or the measuring instruments used. When we want a physical measurement we do not worry about the individual identity of the measuring instrument. We do not concern ourselves with what objects other than the one we want to measure might sometime be or once have been measured. It is sufficient to know that the instrument is a member in good standing of the class of instruments appropriate for the job.

When a man says he is at the ninetieth percentile in math ability, we need to know in what group and on what test before we can make any sense of his statement. But when I say I'm 5'11" do you ask to see my yardstick? You know yardsticks differ in color, temperature, compositions, weight, even size. Yet you assume they share a scale of length in a manner sufficiently independent of these secondary characteristics to give the measurement 5'11" objective meaning. I may be at a different ability percentile in every group I compare myself with. But I am the same 175 pounds in all of them.

Let me call measurement that possesses this property "objective" (Rasch, 1960, 9-12, 109-125; 1966a, 104-105; 1966b, 55). Two conditions are necessary to achieve it. First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. We can't have the instrument changing every time we use it. Second, the measurement of objects must be independent of which instrument happens to be used for measuring.* In practice these conditions can only be approximated. But their approximation is what makes objective measurement possible.

Object-free instrument calibration and instrument-free object measurement are the conditions which make it possible to generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to combine or partition instruments to suit new measurement requirements.**

The guiding star toward which models for mental measurement should aim is this kind of objectivity. Otherwise how can we ever achieve a quantitative grasp of mental abilities or ever construct a science of mental development. The calibration of test item easiness must be independent of the particular collection of persons used for the calibration. The measurement of person ability must be independent of the particular selection of test items used for measuring.

* There is a third condition which follows from the first two. The evaluation of how well a given set of observations can be transformed into objective measurements must be independent of which objects and which instruments are used to produce the observations. It must also be reasonable to hypothesize that objects and instruments have stable characteristics which do not interact with each other.

** Were it useful to glue three twelve inch rulers together to make a thirty-six inch yardstick or to saw a thirty-six inch yardstick in three to make some twelve inch rulers, we would retain our confidence in the objective meaning of length measurements made with the resulting new instruments.

When we compare one item to another in order to calibrate items, it should not matter whose responses to these items we use for the comparison. This means that our method for test calibration should give us the same results regardless of whom we try the test on. That's the only way we will ever be able to construct tests which have uniform meaning regardless of whom we measure with them.

When we test a person it should not matter which selection of items we happen to have found convenient to measure him with or which items he happens to have found time to complete. We should be able to arrive at statistically equivalent measurements of his ability, whatever selection of items happen to have been used.

An Individualistic Approach to Item Analysis

Well, exhortations about objectivity and sarcasm at the expense of present practices are well and good. But can anything be done about it? Is there a better way?

In the old way of doing things we calibrate a test on a standard sample of persons. Item easiness is defined by the proportion of correct responses in the sample. Person ability is defined by percentile standing in the sample. The approach leans entirely on the appropriateness of the standardizing sample of persons.

A different approach is possible, one in which no assumptions are made about the persons used. This approach assumes instead a very simple model for what happens when any person encounters any item. The model just says that the outcome of the encounter is governed by the product of the ability of the person and the easiness of the item. That's all, nothing more. The more able the person, the better his chances for success with <u>any</u> item. The more easy the item, the more likely <u>any</u> person is to solve it.

This simple model has a surprising consequence for item analysis. When measurement is governed by this model, it is possible to take into account whatever abilities persons in the calibration sample happen to have and to free the calibration of the test from the particulars of these abilities. The scores persons obtain on the test can be used to remove the influence of their abilities from the item analysis.

I learned this kind of item analysis from Georg Rasch. But comparable suggestions have been made by others. Some of the ideas have been in print for years. I don't understand why this powerful method is not used in practice.

Perhaps too few recognize the importance of objectivity in mental measurement. Perhaps too many despair that it can ever be achieved. Well, it can, and I am here to prove it.

The crucial questions are: Can test calibration really be independent of the ability characteristics of the persons used to make the calibration? and Can person measurement, the estimation of a person's ability from a score on some selection of test items, really be independent of which items are used for the measurement?

I have placed some data in your hands which illustrate that both of these ideals can be lived up to in practice. These data happen to come from the responses of 976 beginning law students to 48 reading comprehension items on the Law School Admission Test. But they are only one illustration.

Person-Free Test Calibration

In order to examine the dependence of test calibration on the abilities of these law students let us construct the worst possible situation. Into a Dumb Group, we will put the 325 students who did worst on the test. The best of them got a score of 23. Into a Smart Group, we will put the 303 students who did best. The worst of them got a score of 33. We have two groups dramatically different in their ability to succeed on this test of reading comprehension. There are ten points difference between the smartest of the Dumb Group and the dumbest of the Smart Group.

Now for the acid test. How would a test calibration based on the Dumb Group compare with one based on the Smart Group?

Figures 1 and 2

To remind us of how things look using the old way of doing things I made up these calibrations in terms of sample percentiles. Each curve in Figure 1 represents a person-bound test calibration. The curve on the left is the calibration produced by the Dumb Group. The curve on the right is the calibration produced by the Smart Group.

Obviously any person-bound calibration based on the Dumb Group is going to be incomparable with one based on the Samrt Group. From the Dumb Group we can only set up percentile ability measures for students who score between ten and twenty-three. From the Smart Group we can only set them up for students who score between thirty-three and forty-six. These two calibrations do not even overlap. And what about all the scores outside the range covered by either group?

Of course Figure 1 describes an exaggerated situation. No one in his right mind would attempt to base a test calibration on two such different groups. But this exaggeration has a purpose. It is aimed at bringing out a treacherous property of person-bound test calibration and at providing an acid test for any method which claims to be person-free.

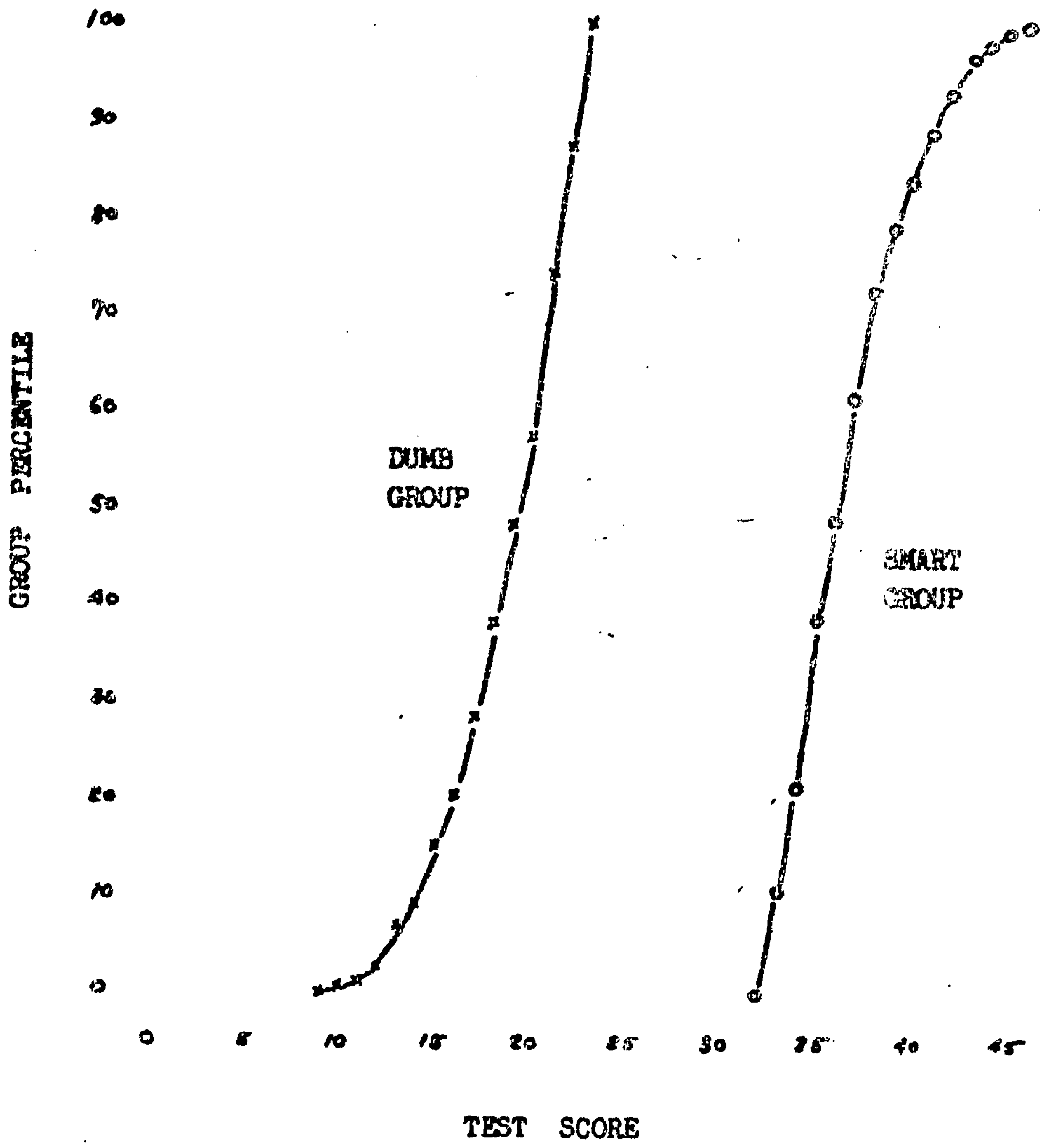
Now let us see how well the new way of test calibration handles this exaggerated situation. I will not burden you with mathematical details. They are covered in the references. Should you become interested in applying the method, let me know. I have a dandy computer program which does it nicely, and a technical write-up which describes every step. Let us look at the results.

Please examine Figure 2. Same data. Same test. Same students. [illegible] different method of calibration. As

Benjamin D. Wright

FIGURE 1

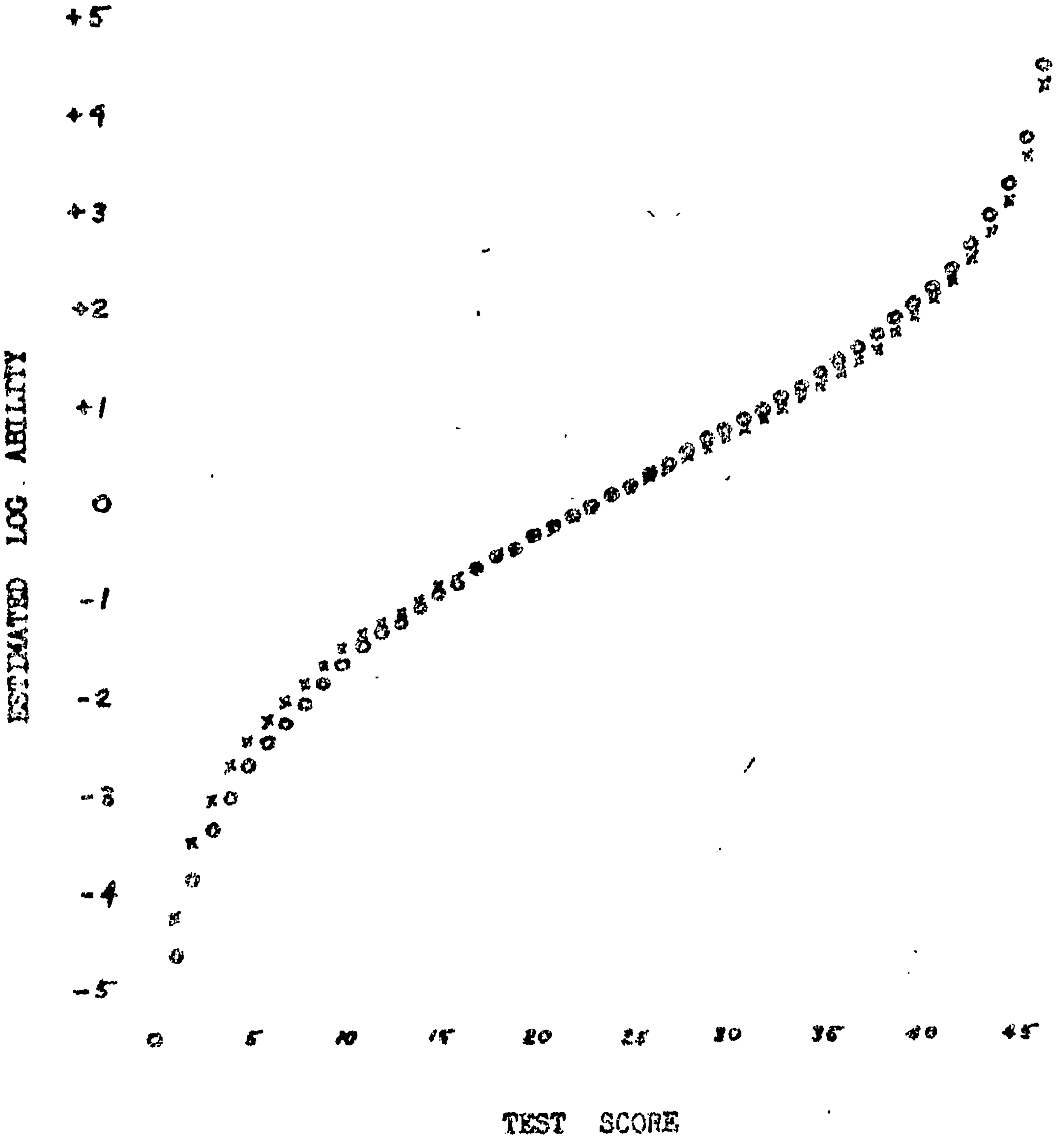
PERSON-BOUND TEST CALIBRATION



Benjamin D. Wright

FIGURE 2

PERSON-FREE TEST CALIBRATION



in Figure 1 the x's mark the test calibration based on the Dumb Group. The o's mark the calibration based on the Smart Group. But now, in Figure 2, how different are the two calibration curves?

At this point you may have a question about how calibration curves work to turn test scores into ability measurements. Each curve represents a conversion table. When a person gets a score on the test, then you enter the graph along the bottom at that score, look up vertically to a calibration curve and then across to the left horizontally to read off his ability. In Figure 1 you would read ability in group percentiles, if you could decide which curve to use. In Figure 2 ability is expressed in logs. If you do not like logs you can take the antilog and get an ability measure on a ratio scale. This may interest you because ability is then measured on a scale where zero means exactly no ability and for which a regular meaningful useful unit can be defined. *

In Figure 1 the calibrations curves do not come close to each other. In Figure 2 they are almost indistinguishable. ** Would you say that the difference between the two calibrations in Figure 2 was of practical significance? How much would you care which of these calibration curves you used to make the test a measuring instrument for you? And yet the two groups on which they are based were constructed to make it as hard as possible to achieve person-free test calibration.

* For a score of 15, the estimated log ability is about -1.0 and the ratio scale ability is about 0.4. A score of 33 indicates a log ability of about +1.0 and a ratio scale ability of about 2.7. Thus a score of 35 indicates about 7 times more ability than a score of 15.

** There is a slight systematic difference. But this reading comprehension test was taken as it stood without any modifications in favor of fitting the item analysis model. When test items are chosen to conform to the statistical requirements of the model then no systematic differences between calibrations are discernible.

One thing that may puzzle you about Figure 2 is the range of test calibration. Either calibration curve provides ability measures for all raw scores on the test from 1 to 47. How can that be done when neither group obtained more than a few of the scores possible?

The answer lies in the measuring model on which these calibration curves are based. Remember that this model uses no assumptions about the abilities of the calibration sample. Its only assumption is what happens when any person encounters any item. Out of this assumption it is possible to calibrate a test over its entire range of possible scores even when everyone in the calibration sample happens to get exactly the same score.

That sounds impossible. But it follows directly from this new item analysis model. The important idea is that even with the same total score persons differ in which items they succeed on. When the calibration sample is large these differences can be used to calibrate the items, and hence the test over its entire range of possible scores, even though only one score has actually been observed.

How would you do that with the present methods of item analysis?

Comparing the calibrations shown in Figures 1 and 2, then, we can see the contrast between the present way of doing things, calibration based on the ability distribution of a standardizing sample, and a new way of doing things, calibration which is free from the effects of the ability distribution of the persons used for the calibration. Which do you prefer? *

* Even though you used this new way as your basis for calibration you could still construct all the percentile standardizations you wanted. Nothing would prevent you from embedding your ability measures in as many sample contexts as you liked. But, and this is the vital point, you would not be bound by those contexts. You would have an ability measure which was invariant with respect to the peculiarities of the persons used to establish the test calibration. If you were a test manufacturer, you would not have to worry over whether you had obtained the right standardizing samples to suit your customers. Your test would be equally valid for all situations in which the test was appropriate. At the same time, since the calibration was person-free, you would be able to use new data as it came in to verify and improve item calibration to add to the item pool and to document the scope of situations in which the test was functioning properly.

Item-Free Person Measurement

So much for person-free test calibration. Now, how about the companion question. Can ability be measured in a fashion that frees it from dependence on the use of a fixed set of items? Is item-free person measurement possible? If a pool of test items have been calibrated on a common scale can we use any selection we want from that pool to make statistically equivalent ability measurements?

In order to judge whether person measurement can be independent of item selection we want a situation that will make it as difficult as possible for person measurememt to be item-free. For this we will divide the 48 items on the original test into two subtests of 24 items each with no items in common between them.

It would be tempting to make these subtests equal in overall easiness. Then they would be parallel forms. But that would be too tame to challenge a scheme for item-free person measurement. Instead the two subtests will be made as different as possible. The 24 easiest items will be used to make an Easy Test. The 24 hardest items will be used to make a Hard Test. Now, under these circumstances, what is the evidence that ability measurement can be item-free? In other words, what is the evidence that the ability estimates based on the Easy Test are statistically equivalent with those based on the Hard Test?

Why do I say statistically equivalent? We know that there are a wide variety of factors at work when a person takes a test. Even knowing a person's ability and an item's easiness will not tell us exactly how he will do on the item. At most we can say what his chances are. This uncertainty follows through into his test score. Even if we could give a person the same test twice, wiping all memory of the first exposure from his mind before his second trial, we would not expect him to get the same score both times. We know there will be some variation. This uncertainty is an inevitable part of the situation. It is the error of measurement.

In finding out how item-free person measurement can be we must make allowance for this. We cannot ask whether estimates of ability based on the Easy Test are identical with those based on the Hard Test. But we can ask whether the two estimates are close enough so that their differences are what we expect from the uncertainties in the testing situation. Are they close enough in the light of their error of measurement to be considered statistically equivalent?

To answer this question we will examine the test responses of the 976 law students to the 48 item test. The score each student earned on the whole test can be split into a subscore on the Easy Test and a subscore on the Hard Test. This gives each student a pair of independent scores each of which should provide an independent estimate of his reading comprehension ability. In order to convert these scores into ability measures on a common scale we will calculate calibration curves like the one in Figure 2 for each of the subtests. To do this we will use item calibrations on a scale common to all 48 items. Then the separate calibration curves for the Easy and Hard tests will convert scores on these different tests into ability estimates on a common scale. If the data fit the item analysis model, then the independent results from these two different tests should produce statistically equivalent ability estimates.

Table 1

The data are in Table 1. The upper half of the table is an obvious example of item-bound person measurement. The 976 law students average 6.78 points more on the Easy Test than they do on the Hard one. How can two tests which lead to such different scores be equated to yield comparable ability estimates?

This problem has been handled in the past by referring test scores back through a percentile table based on some well chosen standardizing sample who have taken both forms. That is one way to equate two tests that are

Table 1

ITEM-FREE PERSON MEASUREMENT

	Test Score		
	Easy Test	Hard Test	Difference
Mean	17.16	10.38	6.78
Std. Error	0.13	0.14	0.11
Std. Deviation	3.93	4.29	3.30

	Estimated Log Ability			Standardized Difference
	Easy Test	Hard Test	Difference	
Mean	.464	.403	.061	0.003
Std. Error	.032	.028	.024	0.032
Std. Deviation	.997	.868	.749	1.014

supposed to measure the same ability. The trouble is that this equation depends on the characteristics of the sample of persons used to equate the tests. We know that an equation based on one group of persons is not in general appropriate for equating measurements made on persons from another group.

Is there a better way to equate tests? Can we go directly from a test score and a person-free calibration of the test items to a measure of ability that does not lean on any particular standardizing sample and that is statistically invariant with respect to those calibrated items that are actually used to obtain the score?

The lower half of Table 1 shows how the new approach equates the Easy and Hard tests. For each person we have his score on the Easy Test and his score on the Hard Test. For each score we look up the corresponding estimated log ability on calibration curves like the ones in Figure 2. For each pair of scores we obtain a pair of estimated log abilities. They will not be identical. But how do they compare statistically?

The distribution of score differences with a mean of 6.78 and a statdard deviation of 3.30 is almost entirely above zero. But the distribution of ability differences with a mean of .061 and a standard deviation of .749 is nicely centered right at zero. On the average there alternative estimates of ability seem to be aiming at the same thing.

How does the variation around zero compare with what would be expected from errors of measurement alone? To examine this we will standardize the differences in ability estimates. For each test score there is not only its corresponding ability estimate but also the measurement error which goes with that ability estimate. The difference between the Easy Test and Hard Test ability estimates can be divided by the measurement error of this difference to produce a standardized difference.

It is the distribution of these standardized differences that will show us whether or not the two ability estimates are statistically equivalent. If they are, then this standardized variable should have a *mean* of zero and a standard deviation of one. That would mean that the only variation observed in ability estimates was of the same magnitude as that expected from the error of measurement in the test. Table 1 shows that, for these 976 students, the standardized differences in ability estimates between the Easy and the Hard tests have a mean of 0.003 and a standard deviation of 1.014. Is that close enough to zero and one to suit you?

What does item-free person measurement mean for test constructors and test users? If you can make statistically equivalent person measurements from any selection of items you wish, then all the tricky and difficult problems of equating parallel forms, connecting sequential forms, and relating short and long forms disappear. Incomplete data ceases to be a problem. You can measure a person with whatever item he takes.

Once you have developed a pool of items which conform to this item analysis model and have calibrated these items, then you are free to make up any tests you wish out of any selection from this item pool. On the basis of these item calibrations alone and without any further recourse to standardizing samples you can compute a calibration curve or a table of estimated abilities along with their errors of measurement for every possible score on any subtest you want.

All such abilities will be on the same ability scale whatever subset of items they were estimated from. You can measure John on an Easy Test and Jim on a Hard Test and be able to compare their resulting estimated abilities on the same ratio scale. That means you can say how many times more of less able John is than Jim in a precise quantitative and meaningful way.

You can measure many children with a short test and a few with a longer more precise test and have all the measures on the same ability scale. Think of how this would expedite screening and selection procedures. The number of items you gave a child could depend on how close he came to the point of decision. Children far away on either side would be quickly detected with a few items. Only children very near the decision point would require longer tests in order to estimate more precisely on which side of the criterion their ability lay.

You would let the required precision, the acceptable error of measurement, determine test length. You would not be bound to any particular predetermined set of items. You could select items from a calibrated pool and compose test forms extemporayeously to suit your measurement needs. * Yet all the measurements made with selections of items from this pool would be located on one scale and used to define whatever norms you or your friends desired. Indeed, since item analyses would be both person and item free, it would be easy to construct tests so that all new data which came in could be used directly to verify and improve item calibration, to add new items to the item pool, to document the range of persons with whom the test was functioning satisfactorily and to establish and extend ability norms for whatever groups were being tested.

* The most important criterion for item selection is the magnitude of measurement error. This is minimum when the person being measured has even odds to succeed on the item. That means that we would like to choose items just right for the person being measured, items just as easy as the person is able. In individual or computerized testing where it is possible to choose the next item on the basis of information gathered from the persons's performance up to that point, this rule specifies exactly what item to use next.

The Item Analysis Model for Measuring Ability Objectively

By now I hope I have whetted your appetite to know more about the item analysis model which made these person-free test calibrations and item-free person measurements possible? The measuring model contains just two parameters. One of these belongs to the person and represents the amount of his ability, Z_n. The other belongs to the item and represents the degree of item easiness, E_i. The model combines these two parameters to make a probabilistic statement about what happens when the person tries the item.

Here is the measuring model: The odds in favor of success, O_{ni}, are given by the product of the person's ability Z_n and the item's easiness E_i.*

$$O_{ni} = Z_n E_i$$

This is the same as saying that: The probability P_{ni} that a person with ability Z_n will succeed on an item with easiness E_i is the product $Z_n E_i$ of his ability and the item's easiness divided by one plus this product.**

$$P_{ni} = Z_n E_i/(1 + Z_n E_i)$$

This is the measuring model used to analyze the forty-eight reading comprehension items on the Law School Admission Test.

* This can equally well be expressed in terms of log odds L_{ni}, log ability X_n and log easiness D_i as

$$L_{ni} = \log O_{ni} = \log Z_n + \log E_i = X_n + D_i \ .$$

The log odds form brings out the simple linear structure from which this model derives its optimal measuring properties.

** This can equally well be expressed in terms of the logistic function as

$$P_{ni} = 1/(1 + \exp(-(X_n + D_i))) \ .$$

What does this simple model say about the scale on which person ability and item easiness are measured? Odds vary from zero to infinity. Since this model gives the odds in favor of success as the product of person ability and item easiness, the natural scale on which to define ability and easiness also varies between zero and infinity.

What does that mean? When a person has no ability then his zero ability will give him zero odds in favor of success no matter what item he tries. With no ability he has no chance of succeeding. On the other hand, if an item has no easiness, then it is infinitely hard and no one can solve it. Measurements made on these scales of ability and easiness have a natural zero.

What about the unit of measurement? Reconsider the product of person ability and item easiness. There is an indeterminancy in that product. We can multiply ability by any factor we like and not change the product, as long as we divide easiness by the same factor. This shows us that if we want to make measurements, we will have to define a measurement unit.

How can such a unit be defined? One way is to select a special group of items as standard. These items can be chosen on theoretical or normative grounds. They can be chosen because they represent a minimal level of ability or an optimal level. Once chosen the combined easiness of these items is set at one. This calibration will then define a person's ability as his odds for success on these standard items.

When a person is functioning at about the level of easiness of these items, then his ability is about one. If he is below the level of these items, then his ability is less than one. If in the course of development or education he doubles his odds for success, that will mean he has doubled his measured ability. Thus one way a unit or measurement can be defined is in terms of even odds to succeed on items selected to be standard.

Another way to define a unit of measurement is in terms of standard persons. These persons can be chosen because they are typical, or because they are liminal for some criterion or because they are the dumbest persons you can find. Now the ability unit is the ability of these standard persons. If you are just at their standard then your ability is one. If your odds to succeed on any item are twice those of a standard person then your ability is two.

In our exploration into what zero means and how to define a unit of measurement we have uncovered the sense in which measures made with this item analysis model are on a ratio scale. When one item is twice as easy as another, then any person's odds for success on the easier item are twice his odds for success on the harder one.

Finally, and most important, this simple item analysis model has a mathematical property which is vital to objectivity in mental measurement. When observations are made in terms of dichotomies like right/wrong, success/failure, then it is a mathematical fact that this is the only model which leads both to person-free test calibration and to item-free person measurement. When observations are dichotomous, the simple form of this item analysis model is the sufficient and necessary condition for objective mental measurement.

Test Construction and the Future of Item Analysis

What bearing does this model for measuring ability objectively have on the construction of mental tests? The model is so simple that those of you who have worried about how to do item analysis may cry out, "What about guessing? What about item discrimination? What about the influence of one test item on another?"

It is obvious that in any real testing situation all of these factors play a part. But rather than "What about them?" I prefer to ask, "What do we want to do with them?

We can construct tests in which guessing plays a big part, in which items vary widely in their discrimination and in which the answer to one item prepares for the next. But do we want to? Not if we aspire to objective mental measurements. If we value objectivity, we will employ our test constructing ingenuity in the opposite direction.*

If we use multiple chice items, we will devise distractors that make guessing infrequent, and we will select items easy enough so that the motivation to guess is slight. When we pilot study the characte stics of potential items, we will select items for the final pool which discriminate equally and fit an objective measuring model.

* Most item analysis models use at least two parameters to describe items. In addition to the item easiness which is part of the simple model presented here, there is also item discrimination. This represents the item's power to magnify or attenuate the extent to which ability is expressed. The discovery of item discrimination was an important step toward understanding how items behave. But as a parameter in the final measuring model it is fatal to objectivity.

If item discrimination is allowed to remain as an active parameter in the measuring model, if variation in item discrimination is tolerated in the final pool of test items, then the possibility of person-free test calibration is lost.

It may be useful to estimate item discrimination when constructing an item pool in order to bring it under control through item selection. But there are more general statistical tests for whether an item or a set of items fit this simple item analysis model. These more general tests are more generally useful.

You might complain that this nice advice is impossible to follow. Do not despair. The reading comprehension items on the Law School Admission Test were not constructed for equal discrimination or item independence. They are multiple choice items with five alternatives. They differ considerably in discrimination and they are grouped around common paragraphs of text to be read for comprehension. Yet the simple item analysis model without guessing, without discrimination and assuming item independence succeeded quite well even with these unfit data. This shows that the measuring model is robust with respect to departures from its assumptions. We do not have to create a perfect test in order to use the model. Nevertheless, if we are really interested in objective mental measurement, then the ideals of no guessing, equal discrimination and item independence can guide us toward constructing better tests. And the kind of item analysis I have illustrated can transform observations made with these tests into objective mental measurements.

References

Loevinger, J. "Person and Population as Psychometric Concepts." Psychological Review, 1965, Vol. 72, pp. 143-15

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960. Chapters V-VII, X.

Rasch, G. "On General Laws and the Meaning of Measurement in Psychology." In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics. Berkeley: University of California Press, 1961, Vol. IV, pp. 321-334.

Rasch, G. "An Individualistic Approach to Item Analysis." In Readings in Mathematical Social Science. Edited by Lazarsfeld and Henry. Chicago. Science Research Associates Inc., 1966, pp. 89-107.

Rasch, G. "An Item Analysis which takes Individual Differences into Account." British Journal of Mathematical and Statistical Psychology. London, 1966, Vol. 19, Part 1, pp. 49-57.

Sitgreaves, R. "Review of Probabilistic Models for Some Intelligence and Attainment Tests" Psychometrika, 1963, Vol. 28, pp. 219-220.

Wright, B. and Panchapakesan, N. "A Procedure for Sample-Free Item Analysis" Department of Education, University of Chicago, January, 1968.