DOCUMENT RESUME

ED 237 514 TM 820 818

AUTHOR Wainer, Howard
TITLE Testing and Test Theory: Whither and Whence.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RM-82-1
PUB DATE Jan 82
NOTE 27p.; Presented at the National Relations Office of the Educational Testing Service (Washington, DC, January 9, 1981).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; *Latent Trait Theory; Metaphors; *Testing; Test Items; Test Reliability; *Tes Theory

ABSTRACT

This paper is the transcript of a talk given to those who use test information but who have little technical background in test theory. The concepts of modern test theory are compared with traditional test theory, as well as a probable future test theory. The explanations given are couched within an extended metaphor that allows a full description of the concepts and implications of test theory without utilizing any mathematics. (Author)

RM-82-1

RESEARCH MEMORANDUM

TESTING AND TEST THEORY: WHITHER AND WHENCE

Howard Wainer

A talk given at the National Relations Office: Educational Testing Service, Washington, D.C.

Testing and Test Theory: Whither and Whence[1]

Howard Wainer

ETS Research Memorandum
RM-82-1

Educational Testing Service
Princeton, New Jersey 08541

January, 1982

[1]A talk given at the National Relations Office of the Educational Testing Service in Washington, D.C. -- January 9, 1981. My thanks to Thomas F. Donlon and Norman Frederiksen for their helpful comments on an earlier draft.

Abstract

This paper is the transcript of a talk given to those who use test information but who have little technical background in test theory.

The concepts of modern test theory are compared with traditional test theory, as well as a probable future test theory. The explanations given are couched within an extended metaphor that allows a full description of the concepts and implications of test theory without utilizing any mathematics.

Table of Contents

"while you and i oices which are
for kissing and o cares if some
one-eyed son-of-a ve s an instrument
to measure spring

..e. cummings, 1926)

INTRODUCTION

On the fifth floor of this building there are two groups of researchers. The first group is in complete sympathy with the notions expressed by Cummings, the second also enjoy kissing and singing, but feel that in many circumstances it is important to measure certain aspects of Spring. They are currently engaged in identical research programs, namely the study of hurdle jumping ability in humans. The strategies these two groups use are quite different, and it is instructive to consider them. The first group had panels of expert coaches studying the movements and builds of various subjects. They had them running and jumping, and they kept copious notes. Later the notes were compared and attempts were made to arrive at a consensus about each subject. The second group had a long runway constructed with a sequence of hurdles spaced out along it; the hurdles started out very low and gradually got higher and higher. Each subject was instructed to run along the runway and jump over each hurdle as it came. I noticed that most would get over the lowest hurdles

easily, and would knock over the highest ones. They had a single clerk recording the pattern of hurdles (standing or knocked over) after each runner completed the course. The runner's hurdling ability was somehow tied to the height of the hurdles that were successfully negotiated.

After observing this for a while I spoke with the directors of these two projects regarding their aims and the probable outcomes. It turned out that this was but one part of a much larger enterprise. Similar sites are to be set up all over the world, and careful records are to be kept of hurdling ability so as to keep track of both individual improvement and any changes that might occur in the general hurdling ability of people over time. The director of the first project told me that they were training experts to make judgements about hurdling ability based upon in-depth study of each runner. I asked how they would measure change in hurdling ability over time both within the same individual and over all children of the same age. He replied that a good coach could make such judgements accurately. As I turned to go to speak to the other study director, I overheard two of the coaches arguing heatedly about whether Joe Louis could have beat Mohammed Ali when both were at their peaks.

The second study was quite different. They had a mimeograph machine working at top speed turning out an instruction sheet that specified the number of hurdles, the distance between them, the height of each hurdle,

their order, the kind of track they must be on, allowable bounds of temperature, wind, etc., etc. They were also preparing a package of hurdles to be mailed out to the other testing sites. At the coffee machine I overheard a discussion between two of the researchers regarding how amazing the progress of women swimmers had been over the last ten years -- that the top women of today would have dominated the men's olympic team just twenty years earlier.

My visit to the fifth floor reassured me of the importance of the work that I would like to discuss with you today.

SOME AIMS OF TESTING AND HOW THEY CAN BE ACCOMPLISHED

The hurdling test described above is a very good one. It embodies much about what we think is sensible about the measurement of human ability. It also presents a context for our discussion of test theories past, present and future.

Often what we are interested in when we give a test is a measure of the ability of the examinee. In the hurdling test just described the measure of ability relates to the heights of the hurdles successfully cleared. This can be operationalized in a variety of ways. If someone performs exactly according to expectation they will have a response pattern like:

11111100000

in which a '1' represents successfully clearing a hurdle, and a '0' means

knocking it over. If the hurdles are in order of increasing size we could then characterize a person by the height between the highest hurdle cleared and the lowest one missed, i.e.

Heights (cm)	10	20	30	40	50	60	70	80	90	100	110
Performance	1	1	1	1	1	0	0	0	0	0	0

This would imply that we would estimate this person's ability as about 55 cm. The second question we would ask is the accuracy of this estimate. In this situation we might say that we were accurate to within 5 cm (i.e. between 50 and 60 cm). Of course, this presupposes tremendous consistency on the part of the hurdler, but if we found that each time the hurdler 'took the test' the same result occurred we would arrive at the same conclusion, and our confidence in this conclusion would increase. Of course, if we wanted more accuracy in our estimate we would have to insert more hurdles between 50 and 60 cm for this person. This is a common problem in testing, for the increased accuracy obtained with a more finely graduated test has an increase in labor for the examinee (i.e. to measure to the nearest cm we would need 110 hurdles rather than the 11 used here). If all hurdlers have to attempt all hurdles, the increased labor is of little help for an individual whose ability is far from the heights being used (i.e. having successfully negotiated ten hurdles between 10 cm and 20 cm does not tell us much more about our 50 cm hurdler than noting that he cleared the 10, 20, and 30 cm hurdles).

Traditional practice increases the number of hurdles in the area in which the greatest discrimination is required. Often we can get a feeling

for the structure of the test by drawing a graph that indicates what the error of measurement is at each jumping level. Shown below is a plot of the error for the test as it is now made up, and another for

---

Insert plot here of error as a function of ability for two tests

---

one in which there were ten hurdles (items) inserted between hurdle 50 and 60 (yielding errors of .5cm).

Adaptive testing tries to solve this problem in another way which will be detailed later.

Suppose our jumper runs down the runway again, and this time clears the 60 cm hurdle but knocks over the 50 cm. This tells us that his ability is still likely to be in the 50-60 range, but our error range is expanded a bit. This points out two components of error in ability estimation --

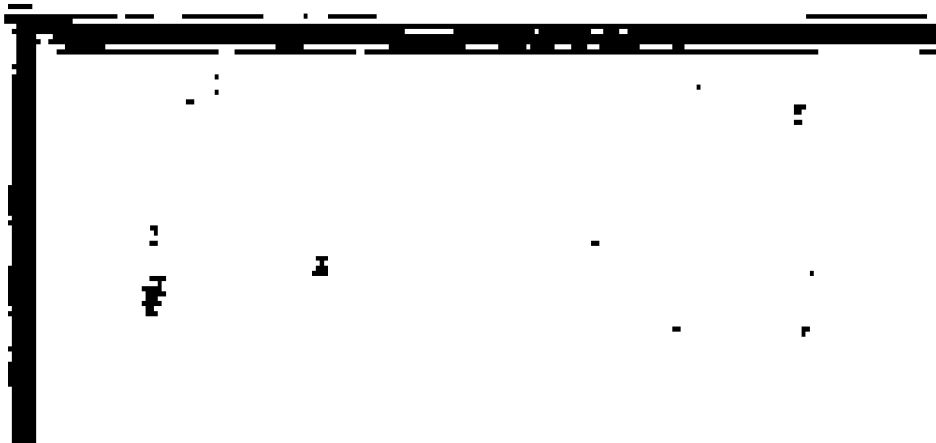
1) the accuracy limitations of the test construction, and
2) the variability of human performance.

We can control the first but not the second.

Enough concepts have now been illustrated so that we can compare these aspects of testing as they were operationalized in traditional test theory (e.g. Gulliksen, 1950), modern item response theory (e.g. Lord, 1980), and in a Gedanken test theory of the future.

TH

6



ABILITY

Traditional Test Theory operationalizes the examinee's ability by the number right. The problem with this concretization is that although it works well when a person performs as he ought, it offers few obvious clues when something is amiss. For example, a hurdler who gets a response vector of 1111100000 would have a score of 5, as would someone with a vector of 1111000001. The latter response seems clearly erroneous, and one would suspect that there was either a clerical error in recording or the runner had sidestepped the last hurdle (this corresponds to successful guessing on a test). It would also give a score of 5 to someone who scores 0000011111. With a response pattern like this we clearly ought to suspect something peculiar is going on with the person taking the test, and a sensible result ought to be to require a retesting. But with traditional test theory a score of 5 is a score of 5, and that's it.

A second shortcoming is that the ability scale derived is only ordinal. That is, that if one person scores 4 and another 5 we perceive that the difference between them is the same as that between someone whose scored ability is 9 and another 10. Clearly changes in raw score do not have the same meaning all along the scale, even if the items are all evenly spaced. In addition, suppose we had expanded the test so that we now had 10 items between 50 and 60 cm; would an increase in score from 14 (hurdles 10 cm - 59 cm) to a score of 15 (jumping over all hurdles from 10 cm through 60 cm) mean the same as an increase from 15 to 16 (clearing 70 cm)? Of course not!

Item Response Theory uses a nonlinear transformation of the proportion correct as an estimate of ability, and centers this estimate on the

difficulties of the items. Further, it yields a goodness-of-fit test that will clearly indicate when an unusual response pattern appears. Thus, the ability given to someone who jumps over all hurdles up to and including 60 cm will be essentially the same regardless of how many hurdles were intervening between the 50 and 60 cm hurdle. The only difference will be the error estimate that is obtained. It also stretches out the ability scale at the ends in such a way so as to yield ability estimates on an interval scale (i.e. observed changes have the same meaning anywhere on the scale - an increase of .5 has the same meaning whether it is from 1 to 1.5 or from 6 to 6.5). Most importantly, by keying ability estimates to the difficulty of the items we obtain a test whose parameters do not change with the norming sample. This is a major advance.

Future Test Theory as I envision it will hold that ability estimates have essentially the same structure as those of current IRT, except that they will be directly referenced to material. In the case of the hurdling example, this means that a person's ability is directly related to the height of the hurdles jumped, and even if no one else ever took the test (how would one compute percentiles?) the results are valid and of interest (i.e. a 70 cm hurdler is a 70 cm hurdler no matter what else happens), and we can measure progress in a metric that makes good sense.

ACCURACY OF MEASUREMENT

Traditional Test Theory - Traditionally, we would measure the accuracy of the assessment of a person's ability estimate by looking at how the test

orders a group of people who have taken the test twice (or manufactured two versions of the test artificially using, say odd and even items). If a test orders the people essentially the same way on two testings we say it is 'reliable'. The extent to which it does this is its 'reliability'. Let us look at the components of reliability. First, it depends upon the discriminating ability of the items (if the test had only two heights of hurdles, 20 cm and 20 feet, we would find that the test was not particularly reliable, since virtually everyone would perform the same way on the test). Second, it depends upon the inherent variability of the individuals being tested (if on one administration a person cleared every hurdle and on the next administration cleared none, we would be hard pressed to assign an ability estimate). Third, it depends upon the variability of ability in the sample being tested (if everyone had the same ability their ordering from one administration to the next would vary enormously thus indicating an unreliable test, when, in fact, the test could be quite good). This is one of the gravest problems with traditional test theory - the reliance on reliability. It can make the character of the test appear to change with changes in the population being tested -- should the accuracy of a scale change depending upon who was being weighed the day you stepped on it?

The effect of the norming group on test reliability and validity is not merely a statistical curiousity. It occurs often, and can stir up trouble when it is not understood. For example, it is common to find

among first year law students that the correlation between the student's LSAT scores and their undergraduate grade point averages is zero, or even negative. This is sometimes used as evidence against the validity of the LSAT. The actual reason is that the LSAT was used (properly) as an admission device along with UGPA. This makes the admitted group much more homogeneous and so reduces the correlation between the measures. One can understand why it goes negative by thinking about the bivariate distribution. Some students will do poorly on both measures; and, therefore, not be admitted. Others will do well on both - they go to Harvard. Thus, the students who attend most law schools have done relatively better on one than on the other, and so were admitted. Surely, a measure of a test's efficacy should not depend upon who is being measured.

Item Response Theory - IRT deals with estimating the accuracy of a test through the standard error of estimate. This is essentially a function of the item structure. Thus, if we have items spread every 10 cm at one part of the test, a person whose ability falls in that part will be accurate to within 5 cm. If in another part of the test we have hurdles every centimeter then the error at that part of the test is of the order of .5 cm. This is regardless of the variability of ability of those people taking the test. In fact, there need only be one person taking this test. The problem with this is that it does not take into account the variability of the person taking the test (actually it does, but not as an individual, only on average) Thus, the error estimate is sometimes a bit on the optimistic side.

Future Test Theory - This sort of test theory would expand or contract the error estimate currently in use on IRT's by the variability of the person taking the test, therefore yielding a more honest estimate of variability, but nonetheless one that is still independent of the group taking the test. Thus, we ought to be able to give a better estimate for a person who responds 1111100000 than one who responds 1111010000, or 1101001010.

ITEM DIFFICULTY

Traditional Test Theory - The difficulty is not well defined in the classic treatment (Gulliksen, 1950), in fact it is not listed in the index as a term! A careful search turns up (p. 367 ff) a variety of definitions that relate the difficulty of an item to the proportion of individuals that get an item correct within a particular sample. This is the grist that will eventually be made into a measure of difficulty, but at the time it was merely one way of doing it. A problem with this formulation is that difficulty changes from one norming group to another. A more serious problem is that the concept of difficulty is not functionally tied to the concept of ability.

Item Response Theory - The most fundamental concept of IRT is the functional relationship between the ability of a person and the difficulty of the item. Referring back to the hurdling test, this means that we can describe the likelihood of a person successfully negotiating a hurdle of a particular height as a function of the difference between their jumping ability and the height of the hurdle. The way that difficulty is defined is as a function of the proportion of individuals who answer it correctly. A hurdle that is almost

never scaled is called 'difficult'. One that is almost never toppled is considered 'easy'. These same designations are possible with traditional test theory, but IRT defines 'difficulty' quantitatively, and unifies it with ability within the context of a theoretical structure.

There is an apparent circularity here that needs to be explicated. Item difficulty is defined by the proportion of people who answer that item correctly. Person ability is defined by the proportion of items that a person answers correctly. Yet I stated that IRT was relatively unrelated to such things as norming groups, and that the accuracy of estimate didn't depend upon who else took the test. How does this follow? The critical concept here is one of the difference between ability and difficulty being the variable of interest. Suppose we gather a group of individuals to take the hurdling test. We have no idea of the difficulty of the items nor the ability of the people. Quickly, we find that some hurdles are easier to jump than others, and some people are more skilled jumpers. Through the intervention of the IRT model we obtain ability estimates for the people and difficulties for the items - of course, they are not correct with respect to origin, but they are correct relative to one another. If we now give the same items to another group of subjects we can calculate their ability on the same scale as the first group. Or we can have the same group jump different hurdles and calibrate those hurdles on the same scale as the original ones. Note that we can choose a subset of the original hurdles to measure some new group of people and do it on the same

scale. Then if we suspect that a particular group of new examinees may have greater ability (for hurdling they may be taller or older) we might give them fewer short hurdles and more high ones (and so increase the accuracy of the test in the area of their anticipated ability). This aspect of IRT is called "Item Independent Person Measurement" -- being able to measure individuals on the same scale in a way that is independent of the precise set of items chosen. Thus, we can choose items in such a way so as to reduce error of measurement. This is important in computerized adaptive testing (more later).

In addition, because the items are calibrated by the difference between their difficulty and the ability of the norming group we can estimate item difficulty from any group, if we are sensible about matching items with people so that they are reasonably suitable.[2] IRT can give us protection from our ignorance, but not from stupidity. The characteristic of IRT that allows us to calibrate items on any group of individuals is called "Sample-free item calibration".

ASIDE - Actually if we do make a mistake and use the wrong calibration sample the model will tell us so. Consider what the ability estimate would be for a person whose hurdling vector is 1111111111. We know that he can clear 110 cm, and we presume that he will miss one of infinite height, so that we can assign an ability estimate between those two extremes with a huge error. Such a huge error of estimate tells us that we don't have enough hurdles in the

[2] As long as we acquire usable information from them -- i.e. if the hurdles are so much beyond the ability of the norming group that they are almost never scaled we cannot get an estimate of their difficulty other than that they are too difficult for this group -- similarly, if they are almost never missed we cannot estimate their difficulty.

appropriate range for this person. Symmetric arguments follow for item calibration. Of course, we needn't have been so extreme as to choose a hurdle of infinite height, perhaps we could have chosen one of height 300 cm as one that would not have been scaled, and consequently calculate a more realistic estimate of ability and error for the 'perfect hurdler'. The insertion of plausible bounds on ability distributions aids us in estimation; such so-called "Bayesian" methods appear to be a fruitful path for future methodology.

Future Test Theory - The shortcoming of the estimation of difficulty within the context of IRT is that it is operationally related to the people. If, somehow, this could be separated from the people taking the test and assessed independently it would allow us to make much more powerful conclusions. Let us consider the hurdling test again. Suppose we kept track of how often each hurdle was successfully jumped, and then we began to make careful physical measurements of the hurdles themselves -- their height and color, the distance between them, where they occur in the test, etc. Suppose we then tried to correlate these physical properties with the observed difficulties. It might be that we would find that we could predict the difficulty of a hurdle from these physical measurements. We could then produce a new hurdle, and before anyone had actually tried it be able to predict those individuals who would or would not be successful in jumping it. Obviously, in this example height is the crucial variable, and we have faith that if we have someone whose hurdling ability is estimated to be 60 cm $\pm$ 5 and we present this person with a hurdle of 47.3 cm we can be pretty well assured that his chances of

successfully negotiating it are greater than someone whose ability was measured to be 50 cm. How much greater, and precisely what is each person's probability of clearing this untried hurdle requires the actual use of the IRT model. The use of expert judges might also enable qualitative judgements to be made; it is the precise statements of the likelihood of success and error bounds around these statements that are the strength of the mathematical model.

SOME EXAMPLES

Traditional Test Theory - Most tests are still scored using traditional true score theory, among them the SAT, the LSAT, the GRE, and virtually all of ETS's tests. There are many reasons for this. Three factors that come to mind are: inertia (theories don't die, just the people who believe in them), a desire to maintain comparability with past performance, and some technical problems associated with the use of IRT on large scale tests.

IRT - TOEFL (Test of English as a Foreign Language) is equated using IRT; a qualifying examination given to prospective physicians by the National Board of Medical Examiners uses IRT for both scoring and equating; many small scale tests (Wainer, 1980; Bock & Fitzgerald, 1972) are examples. At ETS a variety of careful studies are underway that were designed to explore the efficacy of changing to this model on some of the large testing programs (the GRE and SAT are currently being scrutinized as to their suitability for the application of IRT) and a new International Aptitude Test (sort of an SAT given in other countries) is being considered for IRT use for calibrating and equating.

Test Theory of the Future - I have brought with me one test that I consider to be a model of one type of a test of the future. It is the Degrees of Reading Power (the DRP) test currently being used in New York State to measure reading comprehension. It has a variety of characteristics that make it very special. A sample 'item' is shown below. It uses a

---

Insert a Sample DRP Passage Here

---

modified 'Cloze' procedure for testing, and it has been found that the difficulty of the questions is almost perfectly predicted by the readability of the passage. The readability is obtained through a weighted combination of several physical characteristics of the prose (mean sentence length, mean word length, and the mean frequency of occurrence of the words used in ordinary English prose). Thus, we can measure the 'height of the hurdle' in that we can score any piece of expository prose for readability with a computer program, and then predict rather accurately how well someone whose ability has been assessed with the DRP can read it. Further, it means that we can criterion-reference the ability estimates by showing the height of the hurdle that a person with a particular ability can successfully scale -- i.e. 'your child can read The Daily News with 90% comprehension, the Times with 70% comprehension, and the New York Review of Books with 50% comprehension.' Therefore, the teacher with the aid of the DRP can assign reading materials

Dogs help blind people. Many blind people stay home. They will not go outside alone. They can not see. So they are afraid. They think they might fall. Or get hurt. Or get lost.

Such fears are not foolish. There really are many _____1_____. Blind people often need help. But they may not ask people for it. They may get a dog. It is a seeing eye dog. It sees for them.

1 a) jobs b) masters
c) dangers d) expenses
e) tests

The _____2_____ helps a lot. It is a guard. It is a friend. It is a leader. Man and dog go out together. They come to a corner. The dog stops. He looks. He listens. He thinks. He crosses when it is safe. The dog sees if anything is the matter. There may be a fence. Or a hole. Or water. He stops. Then he shows the safe way. Then they go on.

2 a) animal b) doctor
c) exercise d) sound
e) color

The dog must obey. But he must also know when not to obey. Good _____3_____ is important. The man may say "Go." But a car may be coming. Then the dog must not go.

3 a) progress b) health
c) company d) food
e) sense

(Go to the next page)

to children that are suitable, and the parent can more readily understand the progress that his/her child has been making.

Such schemes as that of the DRP seem possible with tests of skills like reading and arithmetic, but less so with tests of knowledge like history and economics. Nonetheless, it seems that future tests should aim toward assessing the difficulty of their items independently of the examinees. Such assessments are usually called 'content validity', but through the use of IRT models we are able to parameterize this concept and specify the relationship between the content validity of the tests and the ability of the examinees.

A second area of improvement in future testing is in the determination of which items will be presented to an examinee. As was pointed out earlier, the accuracy of assessment of ability is partially dependent upon the fineness of the difficulty gradations in the vicinity of each examinee's ability. To make the entire test finely gradated can make the test overlong, tedious, and introduce extraneous (albeit perhaps interesting) factors into the determination of success (grit, determination, endurance, etc.). The usual alternative is to 'peak' the test in the same area as the ability distribution (i.e. have more items in the middle of the ability range than in the extremes) or, if the test is to be used for selection, peak the test in the crucial area of selection (i.e. if a child has to read at a particular level of competency to be promoted into the next grade, have most items at that level of competency).

A future improvement is what has been called "Computerized Adaptive Testing"

(CAT). In this application a computer presents items roughly in the middle of the ability range. If an individual gets them right it presents more difficult items; however, if he gets them wrong easier items are presented. This minimizes the number of too-easy items that could bore the examinee and too-difficult items that can frustrate him/her. It also reduces the likelihood of blind guessing, since the examinee will only rarely be facing an item entirely beyond his ken. To see how this sort of scheme would work, suppose we presented a hurdler with a 50 cm hurdle. If this was cleared, the next one would be 75 cm. If this was missed, one of 62.5 cm would follow. If this was cleared, one of 68.75 cm would be given; and if this was missed, one of 65.625 cm, and so on. The distance between what was passed and what was failed was continually halved until the ability of the individual was estimated with acceptable accuracy. Note that in the example above the hurdler had faced only 5 hurdles, and we were able to estimate his ability to within 3 cm. This would have been the case for anyone whose ability lay in the range 0 - 100 cm. Note that to have done this with a conventional test we would have required hurdles every 6 cm and the hurdler would have faced 11 hurdles before missing. Thus, the length of the test has effectively been halved, and for greater accuracy the savings would have been greater. A further advantage is that the examinee who behaves in a regular way gets a very short test, and can leave. Someone who behaves in an irregular way stays longer. Thus, the length of the test required to obtain a fixed amount of accuracy varies with the examinee. If a test is inappropriate for a

particular examinee, this procedure will not converge within a reasonable amount of time and so cue the tester to the problem.

ENDING

As I left the fifth floor to come down here to present this material, I paused to watch a very athletic young man run over the hurdles. I marveled at his grace and elan, as he cleared all but the highest hurdles. We chatted briefly, while the clerks were tallying up his score. He told me that he was in the other study as well, and the coaches there had classified him a THREE. I said that must indicate a very high rating because he ran the hurdles so well. He replied modestly that the only reason I thought he was so good was because a THREE was the best I had seen. But that my judgement would be quite different had I ever met a FOUR. I replied that most of this talk was a metaphor.

Gulliksen, H. (1950) Theory of Mental Tests. New York: John Wiley & Sons. This is the first complete treatment of true score theory. It explains the various concepts of mental testing clearly and unambiguously, and provides many examples.

Rasch, G. (1960) Probabilistic Models for some intelligence and attainment tests. Copenhagen: Nielson and Lydiche (for Denmarks Paedagogiski Institut). Republished in 1980 by the University of Chicago Press: Chicago. This is a complete statement of the simplest item response theory, the one parameter model, often called "The Rasch Model" after its originator. Besides detailing a test theory model he also explains why this model must be the one employed on measurement theory grounds.

Thurstone, L.L. and Chave E.J. (1929) The measurement of attitude. Chicago: The University of Chicago Press. In this book Thurstone (the originator of virtually all of modern psychometrics) develops much of the methodology that will eventually be called Item Response Theory. He does it in the context of attitude measurement, but it is all there.

Lord, F.M. and Novick, M.L. (1968) Statistical Theories of Mental Test Scores. New York: Addison-Wesley. This book puts the capstone on classical true score theory, developing it from first principles, and showing all of its useful aspects. It also presents the details of item response theory in the four chapters by Birnbaum. Thus, in one sense it can be thought of as ending one era and beginning the next.

Lord, F.M. (1980) Applications of item response theory to practical testing problems. New York: Lawrence Erlbaum Associates. The next twelve years of IRT from the leading expert in the field. This describes the developments of IRT since the publication of Lord and Novick, and shows how to use IRT to solve practical problems.

Wright, B. D. and Stone, M. (1979) Best Test Design. MESA Press: Chicago, A practical handbook on Rasch Analysis. It is to Rasch's book and the 1 parameter model what Lord's book is to the multiparameter IRT.

References

Bock, R. D. and Fitzgerald, W. (1972) The B-F spacial visualization test (unpublished).

Wainer, H. (1980) A test of graphicacy in children. Applied Psychological Measurement, 4, 331-340.