First publ. in: On Educational Testing. San Francisco: Jossey-Bass, 1983, pp. 252-270

12

Wilhelm Kempf

Some Theoretical Concerns about Applying Latent Trait Models in Educational Testing

If there is one crucial problem with the application of so-called psychological test theories in education (or in any other context), it is that both test developers and test users just don't know what they are doing. Let me give you two examples:

First, we have been testing intelligence for three quarters of a century now, but there are virtually no test psychologists who can tell what intelligence is or what they are talking about when they use the word *intelligence.* Still, each of us has a more or less vague feeling that this or that achievement is a mark of intelligence. Although such a feeling does not provide us with any sound terminological basis for constructing a psychological theory of intelligence or for its assessment, it does keep us going. Why not just compare subjects

Konstanzer Online-Publikations-System (KOPS)
URL: http://www.ub.uni-konstanz.de/kops/volltexte/2008/6361/
URN: http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-63616

with respect to a more or less arbitrarily chosen sample of achievements?

Second, when we test someone we know quite well, his or her achievement in the test situation is often not what we would have expected. Since we won't give up our expectations so easily, we say, "No, this is not his 'real achievement.'" Or, taking account of the fact that we have condensed the achievement into a single number, called a *score,* we say, "The test score is not in accordance with his 'true score.'" In any argument based on methodology, of course, our opponent would reply, "How can you defend this assertion, and, anyway, what do you mean by a 'true score'?" However, there was no such opponent in the early days of psychological test theory, and thus nobody forced us to come up with a clearly defined terminology.

We then proceeded to argue as follows: If a subject (v) has a score (x_{vt}) on a test (t) that is not in accordance with his true score (τ_{vt}), then there must be a difference between the two

$$f_{vt} = x_{vt} - \tau_{vt} \tag{1}$$

that we can call the *error of measurement.* But this amounted to the most ingenious fraud ever invented in order to provide a new discipline—test psychology—with high scientific honor.

Seemingly tautological and thus not disputable, equation (1) is neither a definition of the error of measurement nor a mathematical equation that allows for computing the error from given data. It would be so if and only if the concept of *true score* were defined in advance. Nonetheless it has the form of a mathematical equation and thus was highly instrumental in giving test psychology a scientific image: What could better satisfy scientific standards than the mathematical formulation of a theory? And even more: Did we not call the term on the left-hand side of the equation the *error of measurement*? And did this not imply that our tests were measuring instruments?

From here on, the door was closed for constructing a *psychological* test theory, but it was opened wide for constructing *statistical* theories of psychological test scores. There were several paths that could then have been followed, but the most naive one became the most popular; it led to classical test theory, and, beyond that, to the theory of generalizability and the method of factor analysis. The landmarks along the way were: (1) to keep the crucial terms of the theory undefined ("true score," "measurement") and (2) to compensate for the lack of definitions by a habit of inventing mathematically convenient but psychologically unfounded assumptions. These assumptions were often untestable (like the so-called axioms of classical test theory), or never tested in practice and not likely to withstand a critical test (like the assumption of the parallelism of tests), or, even worse, in contradiction to well-known statistical laws (like the assumption of homoscedasticity of error scores) if we take into account how psychological test scores are composed. Finally, some of these assumptions were simply not in accordance with psychological intuition.

An Attempt at Reformulation

Among the few attempts to put classical test theory on a sound methodological basis there is the reformulation by Novick (1966). As it turns out, however, this is the attempt of a statistician, not of a psychologist. A psychologist might ask, "How can the concept of true score be defined so that it will coincide with psychological intuition, that is, will explicate our implicit usage of the phrase *real achievement*?" But Novick posed the question, "How can the concept of true score be defined in order to transform the so-called axioms of classical test theory into mathematically derivable sentences?"

Novick's answer to the question was to state that to each subject v and to each test t there corresponds a random variable X_{vt} with finite expectation $E(X_{vt})$ and variance $\sigma^2(X_{vt})$. (Since any random variable defined on a finite interval has finite moments of every [positive] order and since test scores usually are constructed as the sum [or finite weighted sum] of finite scored responses to a finite number of test items, the latter is not really an assumption but noted just for the sake of completeness.) The true score of subject v on test t is then defined as the expected value of the observed score; that is

$$\tau_{vt} = E(X_{vt}). \qquad (2)$$

Although Novick's definition is a suitable basis on which to found the axioms of classical test theory, there are at least two good reasons why we cannot agree with it: First, Novick's definition does not coincide with psychological intuition. This will become clear, as soon as we explicate our everyday usage of the term *true achievement.* Second, Novick's definition cannot be understood. To state that to each subject and to each test there corresponds a random variable implies that a subject's test score is the result of the application of a random generator or, at least, that it can be treated *as if* it were. This is not the case, since any random generator must satisfy the principle of repeatability; that is, after each application the random generator must be in the same state as before. Only if the principle of repeatability is satisfied can the concept of probability be defined so that the law of large numbers will hold and probability can be interpreted as the limiting value of relative frequency in an increasing number of experiments (that is, applications of a random generator).* As all of us know, psychological tests cannot be administered to the same subject an arbitrary number of times. Any subject may learn or remember something while working on a test that will influence performance when the test is administered the next time. If this were not the case, we would not need to construct special statistical theories of psychological test scores but could just administer the test repeatedly and apply the usual statistics. The only way around this would be to brainwash our subjects after each test administration. However, this would very quickly land us in the realm of science fiction.

* It is often claimed that probability can be defined "implicitly" by stating the axioms of probability theory. As it turns out, this is not the case, since the relations stated in the axioms hold for any relative frequencies as well. Thus, the so-called axiomatic definition cannot make clear what the difference between a frequency distribution and a probability distribution is. For the sake of mathematical reasoning the axioms of probability theory may suffice. They do not suffice, however, for transforming the application of probability predicates on psychological (or any other) events into meaningful sentences. It is also claimed that probability can be defined as the subjective chance of an event. As it turns out, such a definition would leave the axioms of probability theory without foundation and thus cannot help us understand the meaning of the computations based on these axioms. In other words, an intuitive definition of probability cannot make clear what mathematical probability theory and its application have to do with probability.

Latent Trait Theory

But if we decide to forget about classical test theory, what is the alternative? At first glance, the situation with latent trait theory seems to be even worse. The basic concept of latent trait theory as defined by Lazarsfeld (1950) is the item characteristic function. It states that to each subject (v) and to each test item (i) there corresponds a random variable (A_{vi}) with a characteristic probability function:

$$P(a_{vi}) = f_i(a_{vi}, \xi_v), \tag{3}$$

where ξ_v is a vector of parameters that measure the latent abilities of the subject that are involved in his or her response to the item. Implicit to this definition is the assumption of "local independence," which states that a subject's response to an item does not depend on his or her responses to prior items. Although local independence is an essential assumption of most latent trait models, it is not a necessary assumption but can be replaced by specified forms of what I call *local serial dependence* (Kempf, 1977), in which case the concept of item characteristic functions has to be replaced by the concept of conditional item characteristic functions:

$$P(a_{vi} | a_{v1}, \ . \ . \ . \ , a_{vi-1}) = f_i(a_{vi}, \xi_v | a_{v1}, \ . \ . \ . \ , a_{vi-1}).$$

As we can see from Lazarsfeld's definition, the concepts of randomness and probability are once again applied to nonrepeatable events, and—as it has turned out in the context of the Birnbaum (1968) model—this lack of repeatability not only hinders us in understanding the meaning of the probability predicate $P(a_{vi})$ but also involves us in serious statistical problems. (The Birnbaum model is taken as an example only. The same criticism holds for the whole class of latent trait models that do not allow for conditional inference.)

The Birnbaum model was constructed for the analysis of binary items with

$$a_{vi} = \begin{cases} 1 \text{ if S } v \text{ gives a `correct' response to item } i \\ 0 \text{ if S } v \text{ gives a `false' response to item } i. \end{cases} \tag{4}$$

It is defined by the item characteristic function

$$f_i(a_{vi},\xi_v) = \exp(a_{vi}\alpha_i(\xi_v - \sigma_i))/(1 + \exp(\alpha_i(\xi_v - \sigma_i))) \tag{5}$$

where σ_i is an item-difficulty parameter and α_i is an item discrimination parameter. ξ_v is the subject's ability parameter (a scalar, not a vector). As Figure 1 and Figure 2 show, the probability of a correct response $a_{vi} = 1$ is a strictly increasing function of the subject's ability parameter and a strictly decreasing function of the item difficulty parameter. The higher the item discrimination power is, the steeper the ascent of the function.

For any practical application of the model it is crucial to be able to estimate its parameters. Here the statistical problems begin because the well-known methods of parameter estimation (such as the maximum likelihood method) break down. They do not produce consistent estimators (see Neyman and Scott, 1948) because the number of parameters to be estimated does not converge toward a fixed number while the number of observations increases to infinity. This is just another way to state that the repeatability criterion for random generators is violated. Is there a way out of these problems? Can repeatability be established? There are two ways out: a simple one that gives up the concept of latent trait, and an ingenious one that has provided the basis for highly important developments within mathematical statistics.

The simple way out was chosen by Lazarsfeld and led to latent class analysis, which is based on the assumption that each subject belongs to one of a finite number of classes $L = 1, \ldots, M$ and that

$$P(a_{vi}|v\epsilon L) = \pi_{Li}{}^{a}vi(1 - \pi_{Li})^{1-{}^{a}vi} \text{ for all } i = 1,k \text{ and } L = 1, M \tag{6}$$

where a_{vi} are binary responses as defined in (4), and π_{Li} denotes the probability of a response $a_{vi} = 1$ for a randomly chosen subject from class L. The method of latent class analysis then allows for estimation of the parameters π_{Li} and for assigning each subject v to the class that he or she most probably belongs to.

Since latent class analysis is rarely applied within the context of educational testing, let me give an example from clinical psychology: Suppose there are certain clinical syndromes (for example,

Figure 1. Item Characteristic Function of the Birnbaum Model—Ability Parameter.

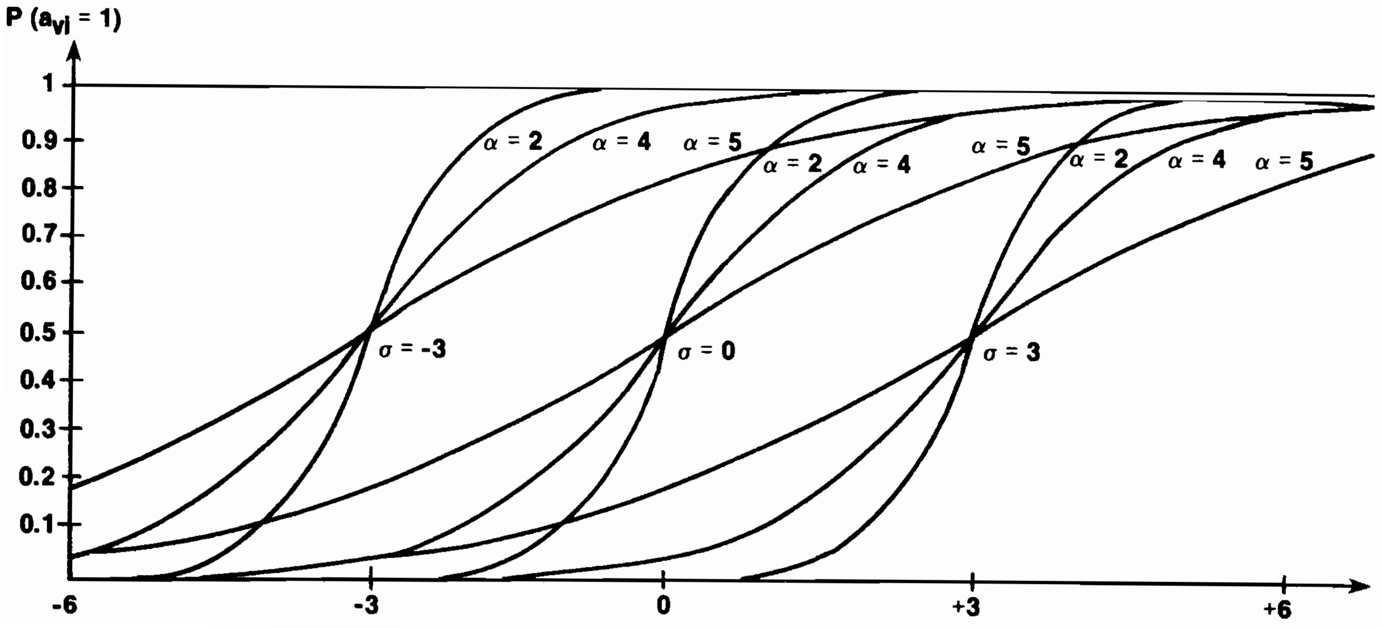
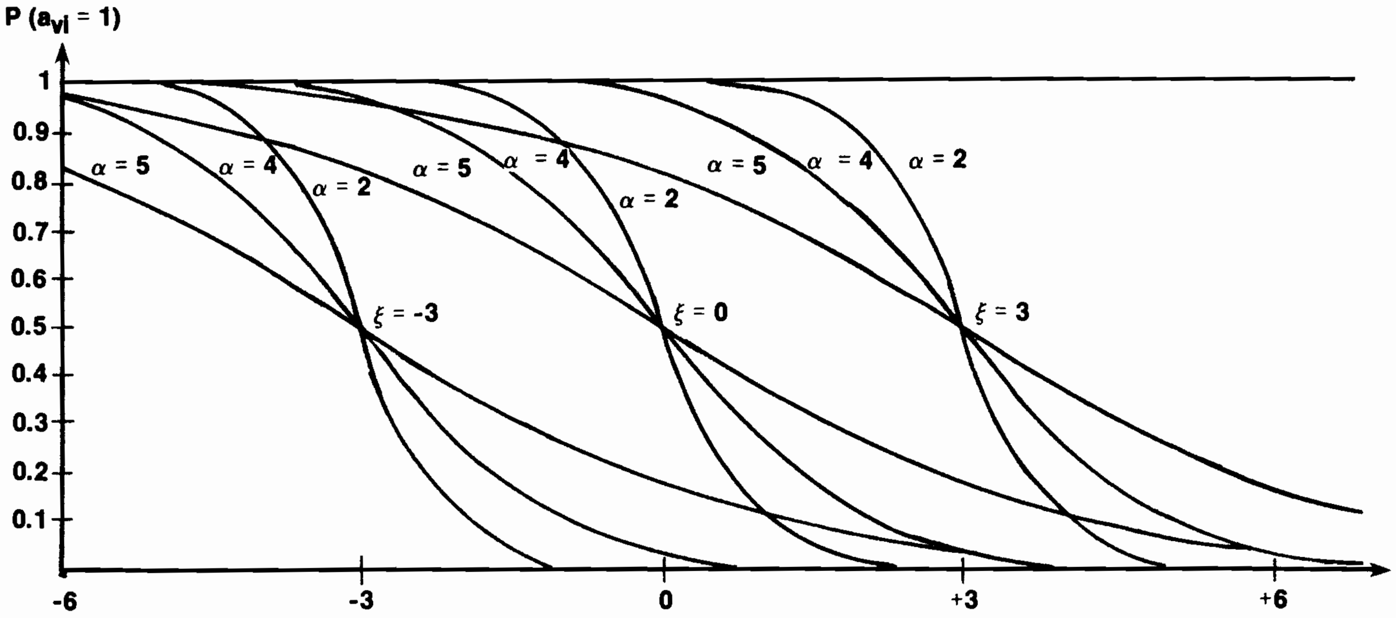


Figure 2. Item Characteristic Function of the Birnbaum Model—Item Difficulty Parameter.



forms of depression) that are made up of certain symptoms (for example, lowered psychomotor functioning, inhibited thinking, anxious-distressed-complaining mood, melancholy, and hypochondria). Then each syndrome (each *class* of depressive patients) will be defined by a typical, "ideal" pattern of syndromes (items). Such an ideal pattern will be made up of some symptoms shown by all patients, some symptoms shown by none, and some symptoms shown by some of the patients but not by others. Now, in reality not all the patients will show such an ideal pattern of symptoms; instead, many will show one or the other of the symptoms that do not belong to the syndrome of their illness and/or will not show certain symptoms that do belong. In such a situation the question arises: To which syndrome should the patient be assigned? A simple and straightforward solution would be to choose that syndrome for which the number of deviations from the ideal is a minimum. So far we do not need any special statistical theory.

The situation becomes more complicated, however, if we do not yet have a theory that tells us what the ideal syndromes look like and if we are still in the process of constructing such a theory. Then it is essential to have some experience with respect to which symptoms typically go together and which do not. This is the situation where latent class analysis applies: as a heuristic method for structuring data in order to get empirically based hypotheses about possible syndromes. These hypotheses can then be used as starting points for theoretical reasoning that in turn will result in the construction of ideal classes.

The second and more complex way to reestablish repeatability was chosen by Rasch (1960, 1965) and is closely related to the concept of measurement, where measurement is defined as a mapping of objects into the set of real numbers so that there is a one-to-one correspondence between certain empirically stated relations and corresponding numerical relations. From this definition it becomes clear that the empirical basis of measurement is the comparison of objects. Now, let us have a set of random generators R_{vi} ($v = 1,2, \ldots ; i = 1,2, \ldots$), each of which defines a random variable A_{vi} with the probability function

$$P(a_{vi}) = f_i(a_{vi}, \xi_v). \tag{7}$$

Let us further suppose that we want to compare these random generators with respect to the parameters ξ_v so that none of the random generators may be applied repeatedly. Then we might ask what form the probability function must take for there to exist a conditional distribution for which the repeatability criterion holds and on which the (empirical) comparison of the random generators can be based.

As Rasch has shown, such a conditional distribution exists in the case of binary items if and only if the probability function has the form

$$f_i(a_{vi},\xi_v) = \exp(a_{vi}(\xi_v - \sigma_i))/(1 + \exp(\xi_v - \sigma_i)). \tag{8}$$

If (8) holds, then the conditional probabilities of the vector variables

$$P(a_{1i}, \ldots ,a_{ni})| \sum_{v=1}^{n} a_{vi} = r_i) = f_r((a_{1i}, \ldots ,a_{ni}),(\xi_1, \ldots ,\xi_n)) \tag{9}$$

are independent of the parameters σ_i; that is, they are the same for all $i = 1,2, \ldots$ and thus pertain to repeatable events.

So far, the statistical problems have been solved: The empirical comparison of subjects can be based on the conditional probability distribution in (9). But what about the problem of psychological meaning? Here it is obvious that the existence of repeatable conditional events cannot establish the psychological meaning of the item characteristic function, which still pertains to unrepeatable events. Thus the item characteristic function has the status of an auxiliary construction that allows for the generation of repeatable and thus meaningful conditional events. For the comparison of any two subjects v and w, for instance, we get the conditional distribution

$$P(a_{vi},a_{wi}) = (1,0)|a_{vi} + a_{wi} = 1) = \frac{\exp(\xi_v)}{\exp(\xi_v) + \exp(\xi_w)}. \tag{10}$$

Hence, if k items are administered to the subjects and if m of these items are solved by exactly one of the two subjects, then the number of items m_v solved by subject v but not by subject w will follow a binomial distribution with

$$P(m_v|m) = \binom{m}{m_v}\left(\frac{\theta_v}{\theta_v + \theta_w}\right)^{m_v}\left(\frac{\theta_w}{\theta_v + \theta_w}\right)^{m-m_v} \text{ with } \theta_v = \exp(\xi_v)$$

$$\text{and } \theta_w = \exp(\xi_w) \qquad (11)$$

so that the ratio $\theta_v/(\theta_v + \theta_w)$ can be estimated by the relative frequency m_v/m. Similarly, the ratio $\theta_w/(\theta_v + \theta_w)$ can be estimated by the relative frequency m_w/m so that

$$\frac{\theta_v}{\theta_w} \approx \frac{m_v}{m_w}. \qquad (12)$$

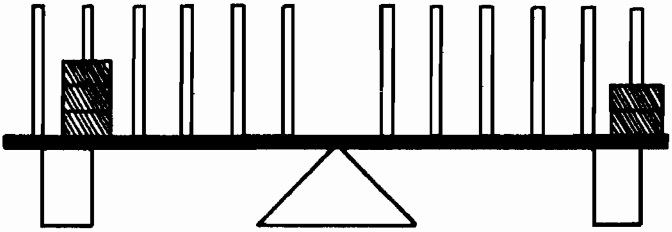
This gives us a straightforward interpretation for the comparison of subjects as carried out by the Rasch model: The ratio of any two antilog ability parameters θ_v/θ_w is an idealization of the ratio based on the numbers of items solved by one of the two subjects but not by the other.

Educational Testing

So far the problems of psychological meaning also seem to be in process of solution. To apply the Rasch model implies that for any two subjects v and w and for any sample of test items $i = 1, \ldots, k$, there will be a more or less stable ratio m_v/m_w and that subjects will be compared with respect to this ratio. But in educational testing is this the kind of comparison that we are really interested in? Are we really interested in finding out that there are x times as many items that subject v can solve but subject w cannot solve as there are items that subject w can solve but that subject v cannot solve? And what are the educational goals for which such knowledge could be useful? To quote Bob Dylan, "The answer is blowin' in the wind." But let me offer one example from educational testing that might be typical of many others.

Suppose we try to teach students a certain body of knowledge. Then it may be essential to find out what the students already know (or believe) in advance and what prior opinions we can build upon. Consider children's understanding of a two-arm balance, for instance, as shown in Figure 3. Here, according to studies by Siegler

Figure 3. The Balance Scale—"Which Side Will Go Down When the Blocks Are Moved?"



(1976), Klahr (1978), and May (1979), the typical development of children's understanding can be understood as a stepwise accumulation of knowledge and hypotheses.

Thus, on the lowest level, children know only that the weights are relevant for balance. Their balance scale predictions based on this knowledge may be described by a flow diagram as shown in Figure 4. Later on children come to realize that the scale is in balance with the same weights on both sides only if the distances are equal too and that otherwise the side with the greater distance will go down (see Figure 5). On level 3 children have realized that the distances are relevant when different weights are put on the two sides, but they don't yet know what will happen when the greater weight is not on the same side as the greater distance; that is, when there is a "conflict" between weights and distances (see Figure 6). On level 4 children come up with the hypothesis that in case of

Figure 4. Flow Diagram of Balance Scale Predictions of a Level-1 Child.

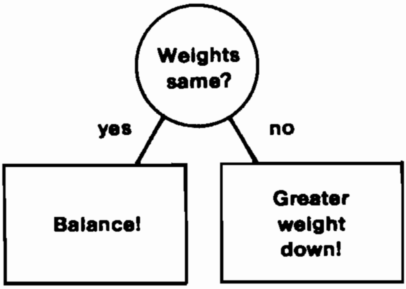
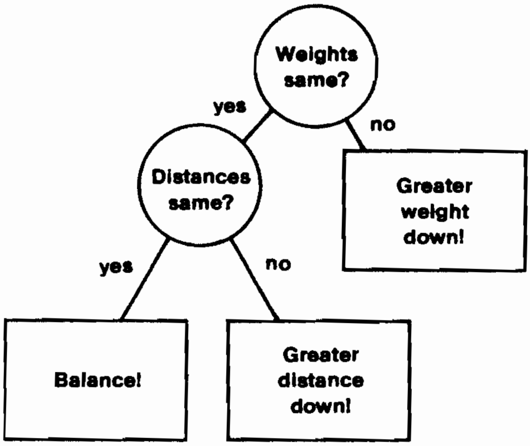


Figure 5. Flow Diagram of Balance Scale Predictions of a Level-2 Child.

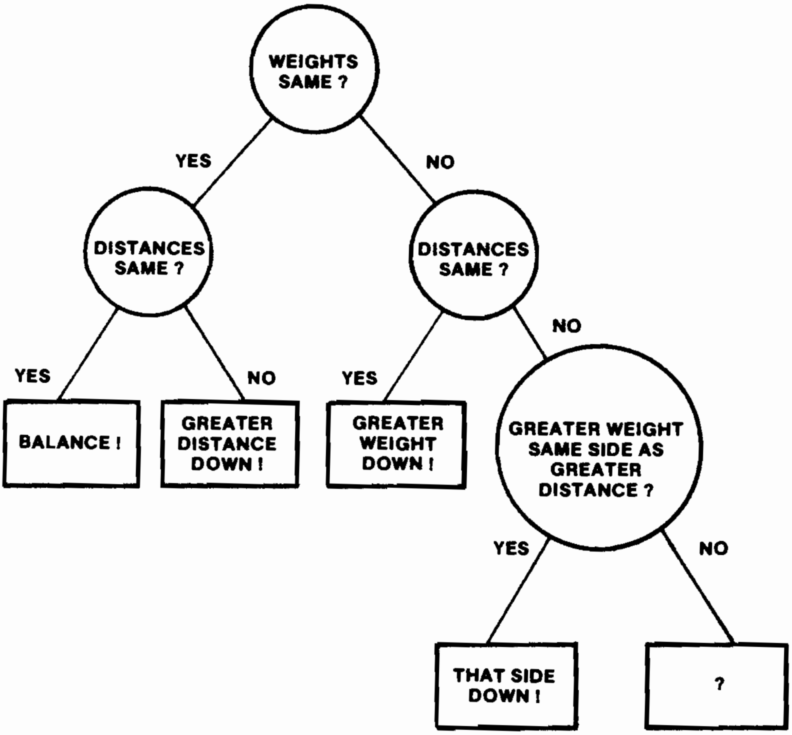


conflict between weights and distances the behavior of the balance will depend on whether there is a greater difference between the weights or between the distances (see Figure 7). On level 5 children will come to realize that it is not the differences but the ratios of weights and of differences that are relevant in case of conflict (see Figure 8).

If we now want to assess a student's prior opinions or knowledge by means of a test, we can do so because there is an analytical connection between a subject's opinion and his or her actions. Given his or her goals and his or her view of the situation, a subject will do exactly what he or she believes to be appropriate in that situation (see Kempf, 1978). In other words, if a subject follows the test administrator's instruction and if he or she does not misjudge the task (for example, he or she does not make a mistake in counting weights and measuring distances), then we can predict the response to the item if we know what the level of knowledge is. This allows us to distinguish several types of items and to contruct ideal response patterns as shown in Table 1.

In order to assess a subject's knowledge, we thus have to compare the pattern of his or her responses with the ideal response patterns and to decide which of these patterns it is in accordance

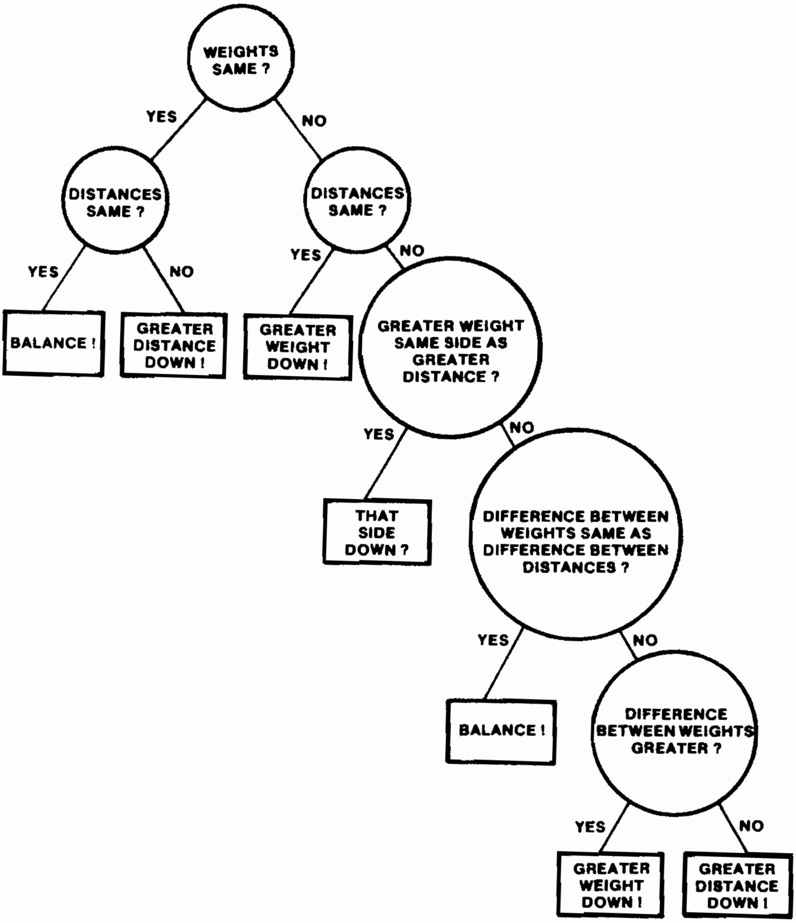
Figure 6. Flow Diagram of Balance Scale Predictions of a Level-3 Child.



with. This means that we need a *qualitative* analysis of a subject's knowledge, not a *quantitative* measurement. With the construction of scores, such as the number of correct responses, the relevant information in the data may get lost. In our example, for instance, the number of correct responses could not discriminate between levels 2 through 4 (see Table 1).

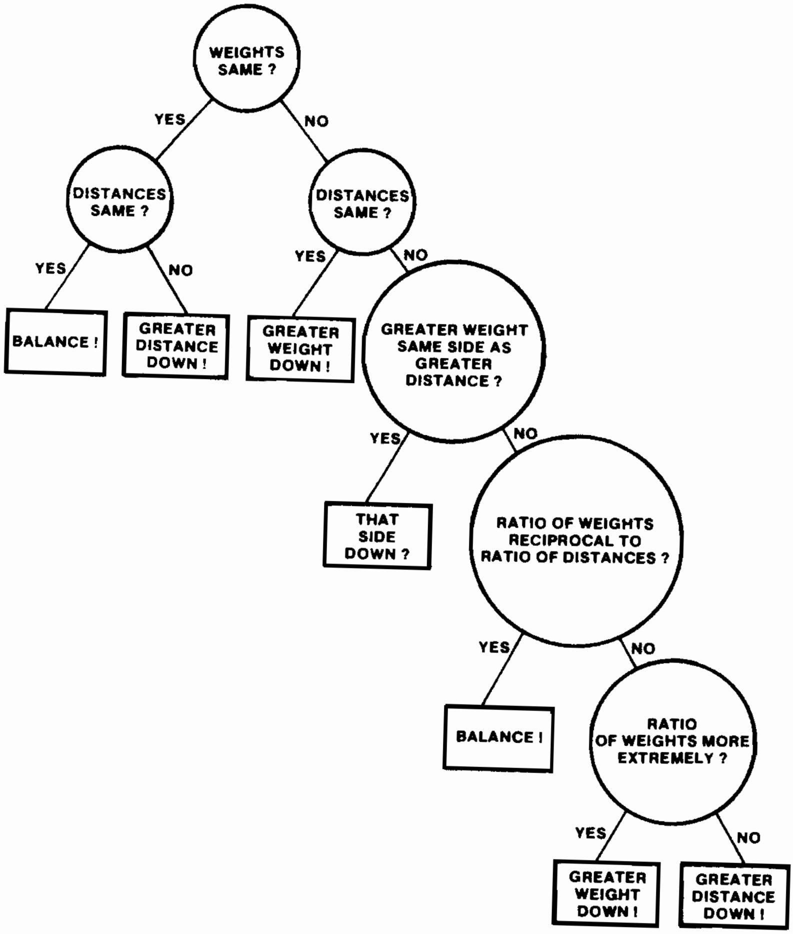
Now, in reality the response patterns usually will deviate more or less from the ideal patterns. As the empirical experience available has shown, however, in a well-designed study the percentage of responses deviating from the ideal will not be more than 6 or 7 percent even with a paper and pencil test. Nonetheless, there arises the question of what to do with subjects whose response patterns are not identical to any of the ideal patterns. Would it be helpful in such a situation to apply the Rasch model and to describe the subjects'

Figure 7. Flow Diagram of Balance Scale Predictions of a Level-4 Child.



knowledge by means of their ability parameters as estimated from the model? Obviously it will not. Since a subject's score is a sufficient statistic for his or her ability parameter in the Rasch model, and since the parameter estimates are nothing more than a strictly increasing transformation of the scores, the ability parameter estimates cannot describe relevant information in the data; for example, a level 2 child and a level 3 child, both of whom give 6 percent

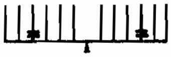
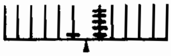
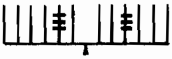
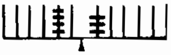
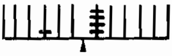
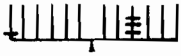
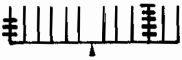
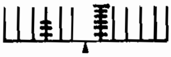
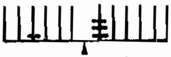
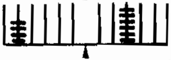
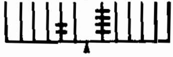
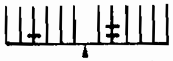
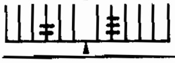
Figure 8. Flow Diagram of Balance Scale Predictions of a Level-5 Child.



atypical responses. The expected score will still remain the same for both children.

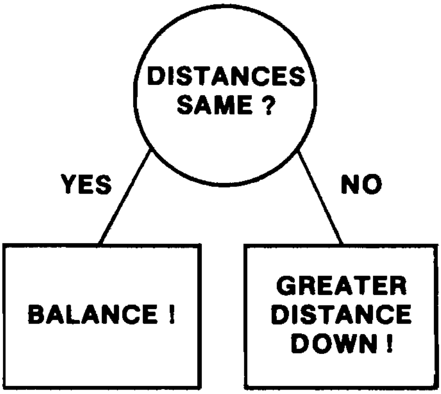
So what do we do with a subject whose response pattern is not identical with any of the ideal patterns? The only reasonable approach is to search for reasons for such a deviation. There might be children, for instance, whose knowledge is entirely different from any of the typical levels previously described and who will therefore

Table 1. Type of Items and Ideal Response Patterns in a Balance Scale Test.

Type of Item		Level of Knowledge						
		1	2	3	4	5	atypical	
	Balance	+	+	+	+	+	+	
	Weight	+	+	+	+	+	−	
	Distance	−	+	+	+	+	+	
	Harmony	+	+	+	+	+	+	
	Conflict-weight/weight	+	+	?	+	+	−	
	Conflict-weight/distance	+	+	?	−	+	−	
	Conflict-weight/balance	+	+	?	−	+	−	
	Conflict-distance/weight	−	−	?	−	+	+	
	Conflict-distance/distance	−	−	?	+	+	+	
	Conflict-distance/balance	−	−	?	−	+	+	
	Conflict-balance/weight	−	−	?	−	+	−	
	Conflict-balance/distance	−	−	?	−	+	−	
	Conflict-balance/balance	−	−	?	+	+	−	
Expected score if test contains h items of each type and if random guessing is assumed in case of "?":		6	7	7	7	13	6	xh

have an entirely different ideal response pattern. In this situation we have to reconstruct the argument that may lead to the observed response pattern. If, for instance, we observe an atypical response pattern as shown in the last column of Table 1, this response pattern could be explained by stating that the child knows only that the distances are relevant for balance (see Figure 9).

Figure 9. Flow Diagram of Balance Scale Predictions of an Atypical Child.



But what shall we do if we cannot reconstruct the knowledge so that the observed response pattern can be explained completely? Here we might be forced to retreat to the assumption of random errors and assign the subject to that class for which the number of deviations from the corresponding ideal response pattern is a minimum. Doing this also makes it clear what we mean when we talk of a subject's "true achievement": The term refers to nothing more than the ideal response pattern a subject would show if his or her knowledge really were what we suspect it to be.

In considering the application of latent trait models in an educational context and the question of which issues should be further developed from an applied point of view, my response was to give an example from educational psychology. The fact that latent trait models do not apply in this example does not necessarily mean that there is no field of application for them in educational research. But it does mean that latent trait models must not be applied routinely. Statistical theories of psychological test scores cannot substitute for psychological theorizing and for a clearly defined psychological terminology. And these are the issues that should be further developed: psychological terminology and psychological theorizing.

References

Birnbaum, A. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In F. M. Lord and M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley, 1968.

Kempf, W. "Dynamic Models for the Measurement of 'Traits' in Social Behavior." In W. Kempf and B. Repp (Eds.), *Mathematical Models for Social Psychology.* New York: Wiley, 1977.

Kempf, W. 'Rule Learning as a Methodological Principle." In J. M. Scandura and C. J. Brainerd (Eds.), *Structural/Process Models of Complex Human Behavior.* Alphen aan den Rijn: Sijthof and Noordhoff, 1978.

Klahr, D. "Information-Processing Models of Cognitive Development." In J. M. Scandura and C. J. Brainerd (Eds.), *Structural/Process Models of Complex Human Behavior.* Alphen aan den Rijn: Sijthoff and Noordhoff, 1978.

Lazarsfeld, P. F. "Logical and Mathematical Foundations of Latent Structure Analysis." In S. A. Stouffer and others (Eds.), *Studies in Psychology in World War II.* Vol. 4 : *Measurement and Prediction.* Princeton, N.J.: Princeton University Press, 1950.

May, R. "Wie Entwickelt sich das Verständnis von Proportionalität." Psychologische Diplomarbeit, Universität Konstanz, 1979.

Neyman, J., and Scott, E. L. "Consistent Estimates Based on Partially Consistent Observations." *Econometrika,* 1948, *16*(1), 1–32.

Novick, M. R. "The Axioms and Principal Results of Classical Test Theory." *Journal of Mathematical Psychology,* 1966, *3,* 1–18.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute of Educational Research, 1960.

Rasch, G. "Kolloquium über Messmodelle." Unpublished paper, 1965.

Siegler, R. S. "Three Aspects of Cognitive Development." *Cognitive Psychology,* 1976, *8,* 481–520.