Conceptual issues in psychological measurement

Borsboom, D.

Citation for published version (APA):
2. TRUE SCORES

Nothing, not even real data, can contradict classical test theory...
– Philip Levy, 1969

2.1 Introduction

In September 1888, Francis Ysidro Edgeworth read a paper before Section F of the British Association at Bath, in which he unfolded some ideas that would profoundly influence psychology. In this paper, he suggested that the theory of errors, at that point mainly used in physics and astronomy, could also be applied to mental test scores. The paper's primary example concerned the evaluation of student essays. Specifically, Edgeworth (1888, p. 602) argued that “...it is intelligible to speak of the mean judgment of competent critics as the true judgment; and deviations from that mean as errors”. Edgeworth's suggestion, to decompose observed test scores into a 'true score' and an 'error' component, was destined to become the most famous equation in psychological measurement: $O_{observed} = T_{true} + E_{error}$.

In the years that followed, the theory was refined, axiomatized, and extended in various ways, but the axiomatic system that is now generally presented as classical test theory was introduced by Novick (1966), and formed the basis of the most articulate exposition of the theory to date: The seminal work by Lord & Novick (1968). Their treatment of the classical test model, unrivalled in clarity, precision, and scope, is arguably the most influential treatise on psychological measurement in the history of psychology. To illustrate, few psychologists know about the other approaches to measurement that are discussed here: You may be able to find a handful of psychologists who know of latent variables analysis, and one or two who have heard about fundamental measurement theory, but every psychologist knows about true scores, random error, and reliability – the core concepts of classical test theory.

The main idea in classical test theory, that observed scores can be decomposed into a true score and an error component, has thus proved a very attractive one. Actually, what was once an idea seems to have been transformed into a fact: There is no psychological measurement without error. This seems to be a safe position to take when applying psychological tests – after all, who would be so overconfident to claim that he could measure perfectly – but it is also counterintuitive. That
is, if I endorse the item ‘I like to go to parties’, why would there necessarily be measurement error involved? Could it not be that I truly like to go to parties? What, then, is measurement error? Now, this question is not as easily resolved as it may seem to be. It is seductive to conclude that random error, for example, represents the impact of unsystematic, transient, factors on the observations (e.g., the subject had a headache at the testing occasion, or she was distracted by noise, etc.). However, we will see that this interpretation is not without problems in the classical test theory framework. More generally, it is exceedingly difficult to reconcile the formal content of classical test theory with common interpretations of terms such as ‘random error’. The friction between intuitive interpretations of terms, and the way they are formally conceptualized, is particularly salient in the interpretation of classical test theory’s central concept, the true score.

The true score is commonly introduced by using phrases such as “the true score is the construct we are attempting to measure” (Judd, Smith, & Kidder, 1991, p.49), or by stressing the distinction “between observed scores and construct scores (true scores)” (Schmidt & Hunter, 1999, p.189). This interpretation of true scores, as ‘valid’ or ‘construct’ scores, has been called the platonic true score interpretation (Lord & Novick, 1968, p. 39 ff.). Of course, the use of the adjective ‘true’ strongly invites such an interpretation, and as a consequence it is endorsed by many researchers and students. However, problems with the platonic interpretation of true scores have been exposed by several authors (Klein & Cleary, 1967; Lord & Novick, 1968; Lumsden, 1976). In particular, cases can be constructed where equating the true score with the construct score leads to violations of basic theorems in classical test theory. In these cases, the identification of true and construct scores will, for example, lead to correlations between true and error scores (Lord & Novick, 1968; Lumsden, 1976), while in the classical test theory model, these correlations are zero by construction.

These observations point to the conclusion that the conjunction of the platonic true score interpretation with the axiomatic system of classical test theory is, at least for some cases, untenable. The implication of such a conclusion would be that, in general, the true score does not admit a realist interpretation. It is argued here that this is indeed the case. Further, the factors that preclude such an interpretation are elucidated. It is argued that the problems can be traced back to the fact that the true score is syntactically defined in terms of a series of observations. This severely restricts the interpretation of the concept; for instance, the true score does not lend itself to an identification with Loevinger’s (1957) traits, which are presumed to exist independently of the test scores. The reason for this is that true scores are conceptualized in terms of observed scores, and, as a result of the way classical test theory is constructed, have a highly restricted domain of generalization – namely, the domain of parallel tests. It is, however, also argued in this chapter that the entire idea, that two distinct tests could be parallel, is inconsistent. This essentially forces the conclusion that the true score can only apply to the test in terms of which it is defined. This, in turn, implies that a conceptualization of psychological constructs as true scores requires an operationalist position with regard to such constructs.
2.2 Three perspectives on the true score

The psychometric models discussed in this book are viewed from three perspectives: formal, empirical, and ontological. The formal perspective consists of two parts. First, the model formulation, or syntax, is discussed. Second, the interpretation of the formal terms in the model, i.e., the model semantics, is evaluated. After clarifying the syntax and semantics of the model, I discuss it from an empirical perspective, by examining the way the model handles data in actual research. Finally, the ontological stance evaluates whether psychometric concepts such as the true score can be taken to refer to an external, objective reality, or must be considered to be products of the imagination of the researcher.

In the context of classical test theory, the formal stance will focus mainly on the syntactical definitions of true and error scores, which form the basis of the theory. The semantic interpretation of these concepts immediately takes us into philosophical territory, because it must be framed in terms of counterfactual premises. Specifically, classical test theory must rely on a thought experiment to establish a version of probability theory that applies to the individual subject; this version of probability theory is needed for a consistent interpretation of the true score. From an empirical perspective, the thought experiment does heavy work in the interpretation of concepts such as reliability. But from an ontological perspective, the fact that the true score is defined in purely syntactic terms, and moreover requires an interpretation in terms of counterfactuals, severely limits the interpretation of the concept. It is argued here that the true score is better conceptualized as an instrumental concept, that governs the interpretation of data analytic results in test analysis, than as an entity that exists independently of the researcher’s imagination.

2.2.1 The formal stance

Syntax Classical test theory is syntactically the simplest theory discussed in this book. Virtually all theorems follow from just two definitions. First, classical test theory defines the true score of person $i$, $t_i$, as the expectation of the observed score $X_i$ over replications:

$$ t_i = \mathbb{E}(X_i). \quad (2.1) $$

Second, the error score $E_i$ is defined as the difference between the observed score and the true score:

$$ E_i = X_i - t_i. \quad (2.2) $$

The notation emphasizes that, while $X_i$ and $E_i$ are considered random variables, the true score $t_i$ is by definition a constant. Note that the error scores have zero expectation by construction, since $\mathbb{E}(E_i) = \mathbb{E}(X_i - t_i) = t_i - t_i = 0$.

An extra source of randomness is introduced by sampling from a population of subjects. As a result, the true score also becomes a random variable and the theory generalizes to the familiar equation

$$ X = T + E. \quad (2.3) $$
Lord & Novick (1968, p.34) note that no assumption concerning linearity needs to be made in order to derive Equation 2.3. The linear relation between true scores and observed scores follows directly from the definitions of true and error scores. Novick (1966) showed that all other required assumptions follow from the definitions of true and error scores for the individual, as given in Equations 2.1 and 2.2. For example, the above definitions ensure the independence of true and error scores, and imply that the error scores have zero expectation in the population (Mellenbergh, 1999).

**Semantics** The true score is defined as the expected value of the observed scores. However, the interpretation of the expectation operator immediately yields a problem, because the expected value of the observed score is conceived of at the level of the individual. This conceptualization is borrowed from the theory of errors (Edgeworth, 1888; see also Stigler, 1986, and Hacking, 1990), which has been fruitfully applied, for example, in astronomy. It is useful to briefly summarize this theory.

The theory of errors works as follows. Suppose that one wants to determine the position of a planet, and that the planet is sufficiently distant for its position to be considered a constant. Suppose further that multiple measurements of its position are made. These measurements, if made with sufficient precision, will not yield identical values (for most readers, this will not come as a surprise, but it was originally considered to be a tremendously shocking discovery; see Stigler, 1986). Now, the deviations from the true value may be interpreted as accidental disturbances, that is, as the aggregated effects of a large number of independent factors (e.g., weather conditions, unsystematic fluctuations in the measurement apparatus used, and the like). It is intuitively plausible that, if this is indeed the case, the observations will tend to produce a symmetrical, bell-shaped frequency distribution around the true value: Because they are accidental, deviations to either side of the true value are equally likely, and, further, larger deviations are less likely than smaller ones. A formal justification for this idea can be given on the basis of the central limit theorem, which states that the sum of independently distributed variables approaches the normal distribution as the number of variables of which it is composed gets larger. Indeed, in the context of astronomical observations, the repeated measurements were often observed to follow such a bell-shaped frequency distribution. The theory of errors combines these ideas: It conceptualizes accidental disturbances as realizations of a random error variable, which will produce a normal distribution of the observations around the true value. If this conceptualization is adequate, then it follows that random errors will tend to average out as the number of observations increases. Thus, in such a case it is reasonable to assume that the expectation of the errors of measurement equals zero. This, in turn, supports the use of the arithmetic mean over a series of measurements as an estimate of the true position, because the mean is defined as the point for which the sum of the deviations from that point equals zero. It takes but a small step to conceptualize the true position of the planet as the expected value of the measurements, for which the arithmetic mean is a maximum likelihood estimator.

If classical test theory dealt with series of repeated measurements for which an analogous line of reasoning could be maintained, there would be few problems in the
interpretation of the theory. However, classical test theory does not deal with such series of measurements, but with measurements on a single occasion. Moreover, series of measurements for which the theory holds are not to be expected in psychological measurement. Such series must satisfy the axioms of classical test theory, which require that the replications are parallel. In a realistic interpretation, this would mean that replicated observations should be considered to originate from a stationary random process; Molenaar (personal communication) has observed that, in the terminology of time series analysis, one would refer to the observed score as a ‘white noise’ variable with nonzero expectation. A procedure that would approximately satisfy the assumptions involved could, for example, consist in repeatedly throwing dice. That throwing dice would conform to the requirements of classical test theory is no coincidence, for what is in fact required is a procedure that allows for the application of the probability calculus in a frequentist sense. In the context of psychological measurement, the stated assumptions are unrealistic, because human beings will remember their previous response, learn, get fatigued, and will change in many other ways during a series of repeated administrations of the same test. Thus, even if the observed scores could be appropriately characterized as originating from a random process (which could be doubted in itself), this random process would not be stationary, which implies that the repeated measurements would not be parallel. It is clear, therefore, that classical test theory a) is not concerned with series of measurements, and b) could not concern itself with such series in the first place, because actual repeated measurements cannot be expected to conform to the assumptions of the theory. Still, the syntactical formulation of the theory uses the expectation operator at an essential point in the development of the theory – namely in the definition of its central concept, the true score. What is to be done about this awkward situation?

**Introducing Mr. Brown**  It is useful to put oneself in Lord & Novick’s shoes in order to appreciate the problems at hand. First, Lord & Novick want to use a probability model based on Kolmogorov’s (1933) axioms, but are unable to give this model a strong frequentist interpretation, which would make it comply with the dominant view of probability at the time (e.g., Neyman & Pearson 1967), because no actual series of repeated measurements will allow for such an interpretation. A subjectivist interpretation (De Finetti, 1974) is conceptually difficult; of course, the true score of subject i could be conceptualized as the expected value of the researcher’s degree-of-belief distribution over the possible responses of subject i, but this view will not match the average researcher’s idea of what constitutes a true value. For example, in psychological testing, the researcher will often not have any knowledge of subject i prior to test administration. In such cases, the Bayesian view would motivate the use of a noninformative prior distribution, which would moreover be the same across subjects. But this would imply that every subject has the same true score prior to testing. This is not unreasonable within the Bayesian paradigm, but it is squarely opposed to the way the average researcher thinks of

---

1 The development commented on here can be found in Lord & Novick, 1968, Chapter 2.
measurement\(^2\). As a consequence, the application of the probability calculus has to be justified in a different manner.

Second, Lord & Novick want to reason along the lines of the theory of errors, but they cannot do this because the assumption that errors will average out in an actual series of repeated observations, and that the arithmetic mean of that series will therefore be a reasonable estimate of the theoretical construct in question, is in flagrant contradiction with the basic fact that human beings, unlike coins and dices, are capable of learning and inclined to do so. Moreover, Lord & Novick do not want to restrict the theory to continuous variables with normally distributed error scores, which, in the theory of errors, are critical for motivating the interpretation of the expected value as the true score. On the contrary, they want to generalize the theory to categorical observed variables, because, in psychological testing, these are far more common than continuous observed variables. For example, intelligence tests work with items that are scored dichotomously (as correct of incorrect), and Lord & Novick surely want their theory to cover such situations.

Third, Lord & Novick need to do something with the individual, but this does not mean that they want to take such an undertaking serious. Classical test theory has no business with the peculiar idiosyncratic processes taking place at the level of the individual: The probability model is merely needed to allow for the formulation of concepts such as reliability and validity, both of which are defined at the population level. A serious attempt at modeling individual subjects (e.g., through time series analysis) would, in all likelihood, not even yield results consistent with classical test theory. So, the subject must receive a probability distribution, but only in order to make him disappear from the analysis as smoothly as possible.

Lord & Novick's response to these problems may either be characterized as a brilliant solution, or as a deceptive evasion. In either case, their approach rigorously disposes of all problems in a single stroke: Lord & Novick simply delete subjects' memory by brainwashing them. Naturally, they have to rely on a thought experiment to achieve this. This thought experiment is taken from Lazarsfeld (1959):

‘Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favour of the United Nations; suppose further that after each question we ‘wash his brains’ and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations.’ (Lazarsfeld, 1959; quoted in Lord & Novick, 1968, pp. 29-30)

Through the application of this thought experiment, the replications are rendered independent as a result of the brainwashing procedure. The resulting hy-

\(^2\) Application of Bayes' theorem would also involve a term denoting the expected value of the observed score, conditional on the true score, and it is not unlikely that the interpretation of this term would still require a thought experiment similar to Lord & Novick's (1968, p.29), to be described hereafter. In this context, it is interesting that Novick, Jackson, & Thayer (1971) do not address this issue, while Novick & Jackson (1974) seem to retain this thought experiment in their Bayesian account of test theory.
2.2 Three perspectives on the true score

A hypothetical series of observations allows for the application of standard probability theory, a quasi-frequentistic conception of probability, and a syntactical definition of the true score which has at least a semantic interpretation: In the particular case of Mr. Brown, the true score equals the probability of him giving a favorable answer, which is estimated by the proportion of times he was in favor of the United Nations.

Propensities? Interestingly, Lord & Novick call the probability distribution characterizing this counterfactual series of replications a propensity distribution. This may be after Popper (1963), who proposed the propensity theory of probability as an objectivist alternative to Von Mises’ conception of probability as relative frequency (Van Lambalgen, 1990). The propensity view holds that probability is not a relative long run frequency, but a physical characteristic of an object like a coin, or, more accurately, of the object and the chance experimental setup (Hacking, 1965). Lord & Novick’s reference to the propensity view is remarkable because, in the thought experiment, they seem to introduce a limiting frequency view of probability. However, the limiting frequency and propensity interpretations of probability do not, in general, coincide. This is because propensities, by themselves, do not logically entail anything about relative frequencies. For example, a coin may have a propensity of .5 to fall heads; then it is possible, although perhaps unlikely, that it will forever fail to do so. In this case, the limiting relative frequency equals zero and thus deviates from the propensity. Because propensities are, in contrast to relative frequencies, logically disconnected from empirical observations, but are nevertheless supposed to conform to Kolmogorov’s axioms, they have been said to operate under the ‘conservation of mystery’ (Kelly, 1996, p. 334). So, strictly speaking, the true score as a limiting frequency in the thought experiment is not logically connected to the true score as a propensity, because the propensity view and the relative frequency view are not logically connected.

Thus, Lord & Novick’s reference to the propensity interpretation of probability is intriguing, especially in view of the fact that they are going through so much trouble in order to generate a relative frequency interpretation for the observed score distribution. One reason for their referencing the propensity view may be that it is the only objectivist theory of probability that allows one to ascribe probabilities to unique events. It is not improbable that Lord & Novick mention the term ‘propensity’ because they are aware of the fact that they are actually doing just this, and therefore cannot use a relative frequency account. But why, then, introduce the thought experiment in the first place? Why not settle for the propensity interpretation and let the relative frequencies be?

My guess is that the reason for this move is twofold. First, propensities are logically disconnected from relative frequencies (i.e., they are not defined in terms of such frequencies), but they are not fully disconnected either. It is in fact obvious that the propensity of a coin to fall heads is related to its behavior in repeated coin tossing. One could say that propensities should be viewed as dispositions to behave in a certain way; a propensity of .5 to fall heads, as ascribed to a coin, could then be viewed as expressing the conditional ‘if the coin were tossed a large number
of times, the relative frequency of heads would approximately be \( .5 \)'. Because ascribing a disposition generally involves a prediction of this kind, Ryle (1949) has called dispositional properties ‘inference tickets’. So, if Mr. Brown’s true score is to be conceptualized in a similar way, the frequency behavior for which it would be an inference ticket must involve replicated measurements. Actual replicated measurements, however, are not generated by stationary random processes, and so it is likely that the propensities will not predict the actual relative frequencies at all. This would render Ryle’s inference ticket useless. The inference ticket would, however, apply to the replicated measurements with intermediate brainwashing.

Second, we must not forget that Lord & Novick are forging an account of psychological measurement; and although they know that they cannot follow the line of reasoning that is the basis for the theory of errors, they do want to stay close to it. The theory of errors is clearly based on an observation concerning the behavior of scores in a long run of replicated measurements. Moreover, it is essential for these series themselves that they are unsystematic, i.e., that they are random. If they were not, there would be little reason to attribute the fact, that repeated measurements are not identical, to unsystematic fluctuations, and to view such disturbances as random error. Again, actual replications are unlikely to produce such series; these will neither be stationary, nor random. Hence, the need for Mr. Brown’s being brainwashed inbetween the replications.

The conclusion must be that Lord & Novick do not need the thought experiment for the application of the probability calculus itself; this could be done solely on the basis of the propensity view. Moreover, the propensity view seems more appropriate because classical test theory is largely concerned with probability statements concerning unique events. Lord & Novick need the thought experiment to maintain the connection between probability theory and the theory of errors, that is, to justify the definition of the true score as the expected value of the observed scores, and to defend the view that deviations from that value are to be interpreted as to random error.

**Thought experiments** The brainwashing thought experiment could be called successful, for it is used in many psychometric models. Models that use it are said to follow a *stochastic subject* interpretation (Holland, 1990; Ellis & Van den Wollenberg, 1993). A stochastic subject interpretation of psychometric models must, in general, rely upon a thought experiment like the above. The thought experiments are needed to provide an interpretation that is in line with both the probability calculus and the typical idea of random error, and could be said to function as a ‘semantic bridge’. This property distinguishes them from other kinds of thought experiments, which are usually directed at a theory, rather than part of a theory (Brown, 1991; Sorensen, 1992). For this reason, it has been proposed to treat these thought experiments as a distinct class of ‘functional’ thought experiments (Borsboom, Mellenbergh, & Van Heerden, 2002-a³).

Classical test theory requires such a functional thought experiment, but this does not mean that it must take the particular form in which Lord & Novick present

---

³ this paper is included in this dissertation as Appendix A
it. Any thought experiment that provides an interpretation consistent with the syntax of the theory could, in principle, do. Rozeboom (1966-a, p.387) considers, for example, that “we may fantasize an experiment in which each member \( i \) of \( P \) has been replicated \( p \) times and each replica (…) is tested (…), so that if \( p \) is large the frequency of a particular observed value \( X \) among among \( i \)’s replicas approaches the probability of this observed score for \( i \)”.

This thought experiment thus considers a probability distribution over a very large number of replicas of Mr. Brown, every one of which is asked whether he is in favor of the United Nations. Still another form of the thought experiment is in terms of an infinite series of administrations of distinct parallel tests. In this case, we would not ask Mr. Brown the same question repeatedly, but we would present him with different questions that are parallel to the original question, that is, with a series of questions that all have the same expected value and error variance as the original question. Probably, many other forms of the thought experiment could be imagined. These thought experiments have in common that, as Rozeboom (1966-a, p. 385) puts it, they “try to convey some feeling for how sense can be made of the notion that a given testing procedure determines a probability distribution over potential test scores specific to each individual who might so be tested”. It should be noted, however, that such thought experiments do little more than convey some feeling. Basically, the classical test theorist is trying to sell you shoes of which it is already obvious that they are three sizes too big.

**How definitions replaced assumptions**  Lord & Novick swiftly go over the construction of true and error scores based on this thought experiment, and manage to dispose of the individual subject in exactly six pages (Lord & Novick, 1968, p. 28-34). In the remainder of their treatment of classical test theory, the focus is on between-subjects results and techniques. At the basis of the theory, however, remains the true score, defined through this peculiar thought experiment.

It is illustrative to recapitulate what has happened here. Lord & Novick have managed to put the theory of errors on its head. Recall that this theory is based on the idea that accidental errors will average out in the long run. The statistical translation of this notion is that accidental error scores can be viewed as realizations of a random variable with zero expectation. The zero expectation of measurement errors must therefore be viewed as an assumption (i.e., its truth is contingent upon the actual state of affairs in the world). On the basis of this assumption, the expectation of the measurements can be conceptualized as an estimate of the true score. Since Lord & Novick are not in a position to use anything resembling an actual series of replications, and therefore are not in possession of a suitable long run, they create one for themselves. However, because their long run is constructed on counterfactual premises, it must remain thought experimental. It is obvious that, upon this conceptualization, the zero expectation of error scores can no longer be taken serious as an assumption, because it applies to a counterfactual state of affairs. As a result, there is no empirical basis for taking the expected value of the measurements as an estimate of the true score. Now, Lord & Novick’s response to this problem is remarkable. Instead of taking the zero expectation of errors as
an assumption on which one can base the hypothesis that the expectation of the observed scores is equal to the true score, they define the true score as the expected value of the observed scores and then derive the zero expectation of errors as a consequence. Where the theory of errors observes irregularities in measurement, and then proposes statistical machinery to deal with those, classical test theory proposes the statistical machinery, and then hypothesizes the irregularities that would conform to it. The identity of expected observed score and true score is thus transformed from a hypothesis into a definition; and the assumption that error scores have zero expectation becomes a necessary truth. Following these moves, one can see the circle close: The theory becomes a tautology. The price that is paid consists in the fully syntactical definition of the true score.

2.2.2 The empirical stance

If the applications of classical test theory were as esoteric as its theoretical formulation, nothing could be done with it. However, classical test theory is without doubt the most extensively used model for test analysis. What, then, does it actually do in test analysis? How does it relate to empirical data?

At this point, it is important to distinguish between how the classical model could be used in test analysis, and how the model is typically used. The basic axioms of classical test theory imply nothing about the data, and are therefore permanently immune to falsification: The adequacy of the posited decomposition of observed scores in true and error scores cannot, for any given item, be checked. Thus, this part of the model is untestable. This does not mean, however, that classical test theory could not be used to formulate testable hypotheses at all. However, to formulate such hypotheses requires extending the model with additional assumptions. These additional assumptions concern relations between true scores on different test forms, or items. Three such relations are commonly distinguished: parallelism, tau-equivalence, and essential tau-equivalence. Two tests \( x \) and \( x' \) are parallel in a population if they yield the same expected value and the same observed score variance for every subpopulation (including subpopulations consisting of a single subject). If distinct tests are assumed to be parallel, they must have equal means and variances; in addition, all intercorrelations between tests must be the same. Two tests are tau-equivalent if they yield the same expected values, but different error variances; and they satisfy essential tau-equivalence if they neither yield identical expected values, nor identical observed score variances, but the expected values are linearly related through the equation \( E(X) = c + E(X') \), where \( c \) is constant over persons. For a given set of items, all three of these relations can readily be tested. For example, as Jöreskog (1971) has observed, when the classical model is extended with any one of the above relations, the model can be formulated as an identified factor model, and the implied covariance matrix can be fitted to the observed covariance matrix. Thus, commonly invoked assumptions about relations between true scores do have testable consequences. At least some parts of the so extended model could be tested.

This is how the model could be applied. It is safe to say, however, that classical test theory is never applied in this way. The common applications of classical test
2.2 Three perspectives on the true score

theory do not involve testing the model assumptions. The cause of this neglect is probably historical, but will not concern us here. Rather, we will be concerned with the function classical test theory fulfills in applications. The strategy that is followed is highly indirect, and works via the estimation of reliability. It is important to review this process extensively, for it contains the basis for many misinterpretations of what classical test theory is about.

Reliability

Reliability is a population dependent index of measurement precision (Mellenbergh, 1996). It indicates the fraction of observed variance that is systematic, as opposed to random, in a given population. In classical test theory, reliability is the squared population correlation, $\rho_{XT}^2$, between true and observed scores. This equals the ratio of true score variance to observed score variance:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}.$$  

This equation has intuitive appeal: In a given population, the value of the reliability coefficient will decrease as the error variance increases. If there is no error variance, reliability is perfect and equals unity. Note that this definition of reliability is population dependent (Mellenbergh, 1996). The reason for this is that reliability is defined in terms of the population model in Equation 2.3. This is reflected in the random variable notation for the true score in the definition of reliability, i.e., in Equation 2.4 the true score is denoted as $T$ and not as $t$. A well-known implication of this definition is that reliability becomes smaller, if the true score variance in a population approaches zero while the error variance remains constant. As a consequence, for any individual subject $i$ the reliability of a test equals zero, because by definition $\sigma_T^2$ equals zero for all $i$. Because reliability is a population dependent concept, it can be meaningfully considered only when interpreted in terms of individual differences in a specific population.

Of course, the formula for reliability contains the true score, which is unobservable. The conceptual strategy of classical test theory consists in rewriting the formula for reliability in terms of potentially observable terms. Lord & Novick (1968) discuss the matter on p. 58-59; what follows here could be viewed as a conceptual reconstruction of this development.

First, suppose that we had the ability of brainwashing subjects inbetween measurements. In this case, determining reliability would pose no difficulties. The determination of true score variance would still be impossible at any given time point, but because replications would be parallel by definition, we could use the correlation between the observed scores on two administrations, $X$ and $X'$, as follows. Assume, without loss of generality, that the expected value of the test scores in the population is zero. The correlation between the observed scores at two time points would equal:

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\mathcal{E}(TT')}{\sigma_X \sigma_{X'}}.$$
See Lord & Novick, 1968, p. 58, for the details of the derivation. This almost equals Equation 2.4, which defines reliability. All that remains to be done is to rewrite the term $E(TT')$ as $\sigma^2_T$, and the term $\sigma_X\sigma_X'$ as $\sigma^2_X$. If this step can be justified, the quantity $\sigma^2_T/\sigma^2_X$, which is unobservable in principle, has been rewritten as the quantity $\rho_{XX'}$, which is observable in principle. This would create a possible connection to the analysis of empirical data. Thus, what we have to do is to interpret a covariance between two variables as the variance of a single variable, and the product of two standard deviations of different variables as the variance of a single variable. This requires that the two variables in question are one and the same. That is, we need to be able to say not only that $T = T'$, in the sense of being numerically equal, but that $T \equiv T'$, in the sense that $T$ and $T'$ are synonymous. The reason for this is not primarily syntactical: $\rho_{XX'}$ will be numerically equal to $\rho^2_{XT}$ as soon as the true scores and error variances on two tests $x$ and $x'$ are numerically equal for each subject, even if this is by accident. For a consistent interpretation of the theory, however, these quantities have to be equal by necessity.

As an illustration of this point, consider the following situation. Suppose that it were the case that height and weight correlated unity in a population of objects, and that these attributes were measured on such a scale that the expected value of the measurement of weight with a balance scale, and the expected value of the measurement of length with a centimeter, happened to always be numerically equal. One could then use the correlation between height and weight as an estimate of the reliability of the balance scale. As a pragmatic empirical strategy, this could work. But theoretically, one cannot admit such a situation in definitions and derivations like the above, because it would not be a necessary, but a contingent fact that the expectations of the measurement procedures were equal; they might very well not have been. Thus, from a semantic perspective, equating the correlation between parallel tests with the reliability of a single test makes sense only if the two tests measure the same true score. This requires that the true scores on the first and second administration are not merely numerically equal, but synonymous.

Can we take the required step while retaining a consistent semantic interpretation of the theory? It is one of the intriguing aspects of classical test theory that this can be done. The reason for this is that the true scores in question are not only syntactically, but also semantically indistinguishable. This is because, for subject $i$, both $t_i$ and $t_i'$ are defined as the expected value on test $x$, where the expectation is interpreted in terms of repeated administrations with intermediate brainwashing. It may seem that, because $t_i$ is the expected value of the observed scores on the first administration of test $x$, and $t_i'$ is the expected value of the observed scores on the second administration of test $x$, $t_i$ and $t_i'$ are distinguishable with respect to their temporal position. But the role of time in the brainwashing thought experiment is a peculiar one. The thought experiment uses the term ‘replications’ in order to make the application of the expectation operator to the individual subject a little more digestible than it would otherwise be, but the idea that we are talking about replications in the actual temporal domain is an illusion. This may be illustrated through the classical test theory models for change (Mellenbergh & Van den Brink, 1998). In such models, the difference between subject $i$'s observed scores on administrations 1 and 2 of the same test, $X_{i2} - X_{i1}$, must be considered.
to be an estimator of $i$'s true gain score, defined as $t_{i2} - t_{i1}$. Each of the true scores is thus defined as the expected value at a single time point. Although the thought experiment creates the impression that the expectation can be interpreted in terms of temporally separated replications of the same test, the term 'brainwashing' must be taken to mean that the subject is restored to his original state — not only with respect to memory, learning, and fatiguing effects, but with respect to time itself. Otherwise, classical test theory concepts such as the true gain score would be completely uninterpretable. Within the brainwashing thought experiment, the true scores on replications must be considered synonymous. Thus, Lord & Novick are justified in stating that $T = T'$, and are able to write

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT_X},$$

which completes the first part of their mission.

Obviously, the development sketched above only takes us halfway in making the connection between classical test theory and the analysis of empirical data. What we want is not to express reliability in terms of counterfactual relations, which involve brainwashing entire populations, but to express it in terms of actual relations between observed variables in real data. So, Lord & Novick's brainwash has had its best time; it has been crucially important in deriving the main psychometric concepts in classical test theory, but now it has to go. Can we get rid of it? The answer is: yes and no. An exact estimate of reliability cannot be obtained from empirical data, so in this sense there is no way to get around the issue. We can, however, settle for lower bounds on reliability, which can be estimated from the data under rather mild conditions. In the final analysis, however, the true score must be invoked again to conceptualize what such a lower bound is a lower bound for.

### Constructing empirical estimates of reliability

The first option for constructing estimates of reliability is to neglect the conditions, that preclude the interpretation of actual repeated measurements as identical with the thought experimental replications, by simply ignoring the problem. This can be done in two ways: either we may assume that two actual replications of the same test are parallel, or we may assume that two distinct tests are parallel. The first of these methods is known as the test-retest method, and the second forms the basis of the parallel test method, the split-halves method, and the internal consistency method.

#### Test-retest reliability

The test-retest method is based on the idea that two administrations of the same test may be regarded as one administration of two parallel tests. If this were the case, the population correlation between the scores on these administrations would be equal to the reliability of the test scores. However, the assumption that repeated administrations are parallel introduces a substantial assumption into the technicalities of classical test theory, namely that the trait in
On the basis of this observation, it has been suggested that the test-retest correlation should be called a ‘stability coefficient’. It should be noted, however, that the between-subjects correlation cannot distinguish between situations where individual true scores are stable and situations where they increase or decrease by the same amount. Therefore, the term ‘stability’ can only be taken to refer to the stability of the ordering of persons, not to the stability of the construct itself. Note also that the method necessarily confounds differential change trajectories and unreliability. We do not know, for most constructs, whether change trajectories are homogeneous or heterogeneous across subjects. This, of course, poses a problem for the interpretation of the test-retest correlation as a reliability estimate.

A second problem is that, in contrast to the thought experimental replications, actual replications are temporally separated, which creates the problem of choosing an appropriate spacing of the replications. Is reliability to be estimated by test-retest correlations based on immediate retesting? Retesting after a day? A month? A year? Since classical test theory cannot provide an answer to these questions, the test-retest scheme must introduce decisions which are, from a methodological perspective, arbitrary. However, these arbitrary decisions concerning the spacing of the replications will generally influence the value of the test-retest correlation. Does this mean that there is a distinct reliability for each choice of temporal spacing? Or should we consider the approximation to reliability to be systematically affected by temporal spacing, so that, for example, the estimate becomes better as we wait longer with retesting? Or does the approximation decrease with the time elapsed since the first administration? Or is this relation curvilinear so that, for example, the approximation is optimal after 1.2 weeks? And should we consider the relation between the quality of the reliability estimate and elapsed time to be the same across testing situations? Across groups? Across constructs? Why? It seems that these issues cannot be satisfactorily addressed, either from a psychological, a philosophical, or a methodological perspective.

In view of these issues, it is interesting that the test-retest method has recently been defended by Brennan (2001), on the grounds that reliability is intelligible only when interpreted in terms of replications of full test forms. This is plausible, but the concept of reliability should be considered within the definitions of classical test theory. Classical test theory defines the true score in terms of a thought experiment, and since the syntactical notation of reliability contains the true score as one of its elements, this definitional issue carries over to the interpretation of reliability. Upon a consistent interpretation of classical test theory, reliability is the proportion of variance in observed scores that would be attributable to variance in true scores; for the test-retest correlation to be an estimate of this proportion, the entire population of subjects must be brainwashed inbetween repeated administrations. Therefore, reliability must conceptually be interpreted in terms of the brainwashing thought experiment; it cannot be defined in terms of actual replications because these simply will not behave according to the axioms of classical test theory. Practically, of course, one may suppose that the actual test-retest correlation is an estimate of the thought experimental one, but in this case it has to be assumed that relevant characteristics of the thought experimental replication are
2.2 Three perspectives on the true score

retained in an actual replication. Unfortunately, the essential characteristics involve parallelism and independence of repeated measurements, i.e., the assumption that the replications could be viewed as realizations of a stationary random variable. This is extremely unrealistic. Thus, the interpretation of the test-retest correlation as reliability (i.e., as the concept is defined in classical test theory through equation 2.4) requires a substantial leap of faith.

Using correlations between distinct tests The second strategy, which encompasses the methods of parallel tests, split-halves, and internal consistency estimates, is based on the idea that two distinct tests could be parallel. First, consider the parallel test method. This method assumes that a simultaneous administration of two different tests could be viewed as approximating two thought experimental replications of a single test. In case we had distinct parallel tests, the correlation between them could then be taken to be a direct estimate of the reliability of the test scores. There are two problems with this method.

The first is a practical problem, namely that the search for parallel test forms has been unsuccessful to date; this is not surprising, because the empirical requirements for parallelism (equal means, variances, and covariances of observed scores) are rather demanding. Further, there is no substantial psychological reason for assuming that two tests for, say, spatial reasoning, should have equal means and variances; nor is there a reason for regarding such tests as theoretically superior to tests that are not parallel.

The second problem is of a theoretical nature, namely that the idea that two distinct tests could be parallel seems semantically inconsistent. We have seen, in Section 2.2.1, that classical test theory interprets the true score on a test $x$ as the expected value on a number of repeated independent administrations of that test. That is, the true score is explicitly defined in terms of the test in question. If we now turn to a distinct test $y$, the true score on this test is semantically interpreted in terms of repeated independent administrations of test $y$. Earlier in this section, we have seen that, to interpret the correlation between parallel test scores as a reliability estimate, the covariance between the two true scores on these measures must be interpreted as the variance of one true score, that is, it must be assumed that $T = T'$. This can be done within the counterfactual state of affairs, defined in Lord & Novick’s brainwashing thought experiment, exactly because $T$ and $T'$ are synonymous. However, the true scores on distinct tests $x$ and $y$ are semantically distinguishable, simply because they are defined with respect to different tests. They may be empirically equal, but this does not make them logically identical. This is to say that the identity of the true scores on repeated administrations with intermediate brainwashing, as used in the derivation of equation 2.4, is a necessary truth; but the empirical equality of expected values on distinct tests is a contingent truth (if it is a truth at all). This may be illustrated by noting that the former equivalence will hold by definition (one does not even have to administer the test to find out), while the observation that the latter holds in the present testing occasion does not guarantee that it will hold tomorrow.

The problem here is not so much that, as a hypothesis formulated independently
of the classical test theory model, two distinct tests could not be taken to measure
the same attribute; this hypothesis could certainly be added, and would in effect
specify a latent variable model. The problem is rather that classical test theory itself
has insufficient conceptual power to do the trick. The syntax of classical test theory
cannot express what it means for two distinct tests to measure the same attribute,
if the attribute is identified with the true score. It is only possible to write down,
syntactically, that two tests measure the same true score. However, semantically,
this makes sense only if these two ‘tests’ are in fact replicated administrations of
the same test, as they are in the brainwashing thought experiment. But of course
the brainwashing thought experiment is completely unrealistic. This is why the
theory must take recourse to the strange requirement of tests that are distinct
and yet parallel. What the syntactical derivations, as well as the semantics, of
classical test theory imply is that parallel measurements consist in two independent
administrations of the same test. A procedure that could reasonably be said to
conform to the requirement of parallelism is, for example, the replicated length
measurement of a number of rods with the same centimeter. With two distinct
psychological items or test scores, however, this logic is, at best, artificial and
contrived; at worst, it is inconsistent. Thus, it is difficult to see how the method
could yield theoretically interesting results, since it seems built on a contradiction.
It is also obvious that the method has no practical value, because tests that satisfy
at least the empirical equivalence needed for exact reliability estimates to work,
are hard to come by. The parallel test method is thus useful for only one purpose,
namely for the derivation of reliability formulae. It cannot be taken serious as an
empirical method.

In the pursuit of exact reliability estimates, two methods have been proposed
that may serve as alternatives to the parallel test method. These are the split-
halves and internal consistency methods. The split-halves method splits a test in
two subtests of equal length, assumes that the subtests are parallel (or constructs
them to be nearly so; Gulliksen, 1950; Mellenbergh, 1994), computes the correla-
tion between the total scores on subtests, and yields an estimate of the reliability of
total test scores by using the Spearman-Brown correction for test lengthening. In-
ternal consistency formulae such as the $KR_{20}$ and coefficient $\alpha$ extend this method.
They can be interpreted as the average reliability coefficient as derived from the
split-halves correlation, where the average is taken over all possible split-halves. If
the split-halves are parallel, the resulting quantity yields an exact estimate of the
reliability of the total test scores. Since parallelism is as troublesome for split-halves
as it is for full test forms, these methods fail for the same reasons as the parallel
test method.

Lower bounds The exact estimation of reliability from observed data is thus im-
practical and theoretically questionable. This has prompted classical test theo-
rists to look at worst-case scenarios, and to search for lower bounds for reliability
(Guttman, 1945; Jackson & Agunwamba, 1977). For instance, it can be proven
that, if test forms are not parallel, but satisfy weaker assumptions such as essential
tau-equivalence, reliability estimates like Cronbach’s $\alpha$ provide a lower bound on
2.2 Three perspectives on the true score

reliability. Thus, if $\alpha$ equals .80 in the population, then the reliability of the test scores is at least .80. This is a clever strategy, and the researcher who follows it seems to be fairly safe. In essence, the reasoning which could be followed is: no matter how bad things may be, the reliability of my test is always higher than the (population) value of the lower bound that is computed. This is probably the most viable defense that could be given for the standard practice in test analysis.

Note, however, that the true score does not do any work in the computation of any of the statistics discussed. The test-retest correlation is, well, a test-retest correlation, and internal consistency is just a transformation of the average split-half correlation. Both could be used in test analysis, and judged for their merits, without recourse to classical test theory as a theory of measurement. The statistical machinery will do just fine. However, this does not mean that classical test theory is irrelevant to the way the analyses are used. For the interpretation of test-retest correlations or average split-halves correlations as reliability estimates does involve classical test theory. What is obtained in the analysis is a test-retest or average split-halves correlation, but when these are interpreted in terms of reliability, they are interpreted as estimates of, or lower bounds for, the quantity denoted as $p^2_{XTX}$, and this quantity does involve the true score as defined in classical test theory. Thus, what we observe here is an inference from empirical relations (involving only observables) to theoretical relations (involving observables and unobservables). This type of inference is, of course, nothing new, for it is the gist of science. What is typical and unusual here, is that the inference does not come at a price. The researcher gets the theoretical interpretation in terms of unobservable true scores for free. The question, however, is what this theoretical interpretation is worth: What is it exactly, that we are informed about? What is the status of the true score?

2.2.3 The ontological stance

Of all psychometric concepts, reliability plays the most important role in practical test analysis. Of course, all researchers pay lip service to validity, but if one reads empirical research reports, reliability estimates are more often than not used as a primary criterion for judging and defending the adequacy of a test. In this sense, reliability is the poor man's validity coefficient, as Rozeboom (1966-a) has observed. I think that the analysis presented above casts doubt on whether reliability deserves this status. The theoretical acrobatics necessary to couple empirical quantities, like test-retest correlations, to reliability, as defined in classical test theory, are disconcerting. Coupled with the fact that these coefficients are used and interpreted rather uncritically, the observation that "classical measurement theory [is] the measurement model used in probably 95% of the research in differential psychology" (Schmidt & Hunter, 1999, p. 185) seems to be a cause for concern, not for celebration. The problems grow even deeper when one considers that 95% of the researchers involved in research in differential psychology are probably not doing what they think they are doing. For no concept in test theory has been so prone to misinterpretation as the true score.

As has been noted earlier in this chapter, it is tempting to think that the distinc-
tion between true scores and observed scores is the same as the distinction "between observed scores and construct scores" (Schmidt & Hunter, 1999, p.189), or that "the true score is the construct we are attempting to measure" (Judd, Smith, & Kidder, 1991, p.49), or that it is the score "that would be obtained if there were no errors of measurement" (Nunnally, 1978, p. 110). This is the way the matter is often explained to students, and it is the way many researchers think about psychological measurement. However, the identification of the psychological construct with the true score of classical test theory is not without problems.

There are two problematic assumptions underlying the platonic interpretation of the true score. The first assumption underlying the idea that the true score is the real score on a psychological construct is the result of a confound of unreliability with invalidity. This is a recognized fallacy, but it is so common and persuasive that it deserves a thorough treatment. The second assumption concerns the ontological status of the true score itself. It will be argued here that the entire idea that a person has a true score, as defined in classical test theory, is unintelligible – except when interpreted in a thought experimental sense. So interpreted, it has the status of a dispositional concept, but, oddly enough, it specifies dispositional properties with respect to an impossible sequence of situations; namely, the thought experimental replications. The true score is therefore best thought of as a fiction. Finally, in contrast to psychological constructs, the true score cannot be conceptualized independently of the test in question. This is why the true score must be seen as a concept that is best interpreted in an operationalist sense.

**True scores as construct scores**

The idea that true scores are valid construct scores can be seen as a confound of reliability and validity. These are qualitatively different concepts: Reliability has to do with the precision of the measurement procedure, while validity involves the question whether the intended attribute is indeed being measured. For the simple reason that no formal model can contain its own meaning (it cannot itself say what it is a model for), it seems obvious that this interpretation is incorrect from the outset. However, although various authors have warned against it, the platonic true score interpretation is like an alien in a B-movie: No matter how hard you beat it up, it keeps coming back. A recent revival has, for example, been attempted by Schmidt & Hunter (1999; see Borsboom & Mellenbergh, 2002, for a criticism). True scores are not valid construct scores, and neither do they necessarily reflect construct scores.

At the present point in the discussion, the concept of validity is introduced, and therefore the relation of measurement has become important. In itself, it is interesting that, in the entire discussion so far, the term 'measurement' has remained unanalyzed. We have been able to review the assumptions, semantics, and empirical applications of classical test theory without making the meaning of this concept explicit. This is typical of classical test theory and contains an important clue as to why the identification of true scores with psychological constructs is so problematic. To see this, take it as given that the objective of psychological testing is to measure constructs, or, if you like, the phenomena to which constructs refer. If true scores
could be taken to be identical to construct scores, then it should be possible for classical test theory to rewrite the relation of measurement, interpreted as a relation between observed scores and construct scores, as a relation between observed scores and true scores. It turns out that classical test theory cannot do this. The reason for this is that, because the theory is statistical in nature, it is natural to conceive of the relation between observed scores and construct scores statistically. This is also the position taken by Lord & Novick (1968, p. 20), who say that ‘... an observable variable is a measure of a theoretical construct if its expected value is presumed to increase monotonically with the construct’ and ‘... to be primarily related to construct being defined’. This is similar to the measurement relation as conceived in item response models, where the expected value on items is related to the position on the latent variable. It follows from this conceptualization, however, that true scores cannot play the role of construct scores. This is because the true score is itself defined as the expected value on a test, so that identifying true scores with construct scores and substituting this in Lord & Novick’s conception of measurement leads to the following definition: ‘... an observable variable is a measure of a [true score] if its [true score] is presumed to increase monotonically with the [true score]’. This can hardly be considered enlightening.

In contrast to, for example, latent variable models, classical test theory does not have the conceptual power to represent the construct in the model. The relation of measurement must thus be seen as a relation between true scores and something else. This is in perfect accordance with the way validity is treated in classical test theory, namely as the correlation between the true scores on the test in question and an external criterion. However, it is inconsistent with the idea that true scores are construct scores. It is actually rather strange that this misconception occurs at all, because classical test theory defines the true score without ever referring to psychological constructs or a measurement relation. The theory does not contain the identity of true scores and construct scores - either by definition, by assumption, or by hypothesis. Moreover, it is obvious from the definition of the true score that classical test theory does not assume that there is a construct underlying the measurements at all. In fact, from the point of view of classical test theory, literally every test has a true score associated with it. For example, suppose we constructed a test consisting of the items “I would like to be a military leader”, “$.10/\sqrt{.05} + .05 = .. ”, and “I am over six feet tall”. After arbitrary - but consistent - scoring of a person’s item responses and adding them up, we multiply the resulting number by the number of letters in the person’s name, which gives the test score. This test score has an expectation over a hypothetical long run of independent observations, and so the person has a true score on the test. The test will probably even be highly reliable in the general population, because the variation in true scores will be large relative to the variation in random error (see also Mellenbergh, 1996). The true score on this test, however, presumably does not reflect an attribute of interest. The argument shows that it is very easy to construct true scores that have no substantial meaning in terms of scientific theories, and are therefore invalid upon any reasonable account of validity.

It is also very easy to construct situations in which there is a valid construct score, while that score differs from the true score as classical test theory defines
it. Consider, for example, the following example, which is based on an example by Lord & Novick (1968, p.39 ff.). At present, whether a patient has Alzheimer’s disease or not cannot be determined with certainty until the patient is deceased and autopsy can be performed. In other words, the diagnostic process, taking place while the patient is still alive, is subject to error. We can conceptualize the diagnostic process as a test, designed to measure a nominal variable with two levels (‘having the disease’ and ‘not having the disease’). Because this variable is nominal, we may assign an arbitrary number to each of its levels. Let us assign the number ‘1’ to a patient who actually has Alzheimer’s, and the number ‘0’ to a patient who does not. This number represents patient i’s construct score $c_i$ on the nominal variable ‘having Alzheimer’s’. Thus, a patient who actually has Alzheimer’s has construct score $c_i = 1$, and a patient who does not have Alzheimer’s has construct score $c_i = 0$.

In practice, the construct score cannot be directly determined. Instead, we obtain an observed score, namely the outcome of the diagnostic process. This observed score is also nominal, so we may again assign an arbitrary number to each of its levels. Let us code patient i’s observed score $X_i$ as follows. The value $X_i = 1$ indicates the diagnosis ‘having Alzheimer’s’, and the value $X_i = 0$ indicates the diagnosis ‘not having Alzheimer’s’.

The diagnostic process is imperfect and therefore the test scores are subject to error. Now suppose that the test is valid, so that misclassifications are due solely to random error, for example, to equipment failures that occur at random points in time. This renders the observed score a random variable $X$. What is the true score on the test? It is tempting to think that patient i’s true score, $t_i$, on the diagnostic test is equal to the construct score (i.e., $t_i = c_i$). Specifically, the infelicitous use of the adjective ‘true’ suggests that a patient who actually has Alzheimer’s, i.e., a patient with construct score $c_i = 1$, also has a true score of $t_i = 1$ on the test. For this indicates the diagnosis ‘having Alzheimer’s’, and it is, after all, true that the patient has that disease.

This interpretation of the true score is not, in general, consistent with classical test theory. For suppose that the sensitivity of the diagnostic test is .80. This means that the probability that a patient who actually has Alzheimer’s will be correctly diagnosed as such is .80. Now consider the true score of a patient who has Alzheimer’s, i.e., a patient with construct score $c_i = 1$. This patient’s true score is not $t_i = 1$, because the true score of classical test theory is equal to the expectation of the observed score, which is $t_i = E(X_i | c_i = 1) = .80$. Suppose further that the sensitivity of the test is .70. This means that the probability that a patient who does not have Alzheimer’s will be correctly diagnosed is .70. For a patient who does not have Alzheimer’s (i.e., a patient whose construct score is $c_i = 0$), the true score is equal to $t_i = E(X_i | c_i = 0) = .30$. In both cases the true score and construct score yield different values.

It can now be seen why the identification of true scores with construct scores is logically inconsistent with classical test theory in general. If the test in the

---

4 Note that the argument implicitly uses a latent class formulation, where the construct score indicates class membership; this suggests that latent variables can be used to extend the model in the required direction. It will be argued in the next chapter that this is indeed the case.
example contains error, this means that there is misclassification; and if there is misclassification, the expected value of the observed score can never be equal to the construct score. So, if measurements contain random error, the identification of true scores with construct scores is logically inconsistent with classical test theory in general. It should be noted that Lord and Novick (1968) themselves were thoroughly aware of this, since they explicitly state that "in general the two concepts and definitions [of true scores and construct scores] do not agree" (p. 41).

It is clear that the identification of construct scores with true scores is fundamentally incorrect. The objective of psychological measurement is to measure psychological constructs, but classical test theory cannot express the relation of measurement as a relation between observed and true scores. Rather, the theory must conceptualize the measurement relation as a relation between true scores and psychological constructs, which shows that these should not be considered identical. This conclusion is strengthened by the observation that we can easily construct cases where a true score 'exists', but where it is invalid in that it does not have substantial meaning in terms of a theory. We can also construct cases where there is a valid score, but where that score is not the true score.

In view of these problems, it is interesting and elucidating to inquire under what conditions the true score and the construct score could be taken to coincide. It seems that the situation, in which this would be the case, is exactly the situation as the theory of errors portrays it. Namely, if the validity of the test has been ascertained, the observations are continuous, the attribute in question is stable, and deviations from the true value over actual replications are produced by a large number of independent factors. In this case, the axioms of classical test theory will be satisfied by actual, rather than thought experimental, replications – in fact, there would be no need for a thought experiment. It also seems that the number of psychological measurement procedures, for which these assumptions could be taken to hold, equals zero. Thus, it is safe to conclude that, in psychological measurement, the true score cannot be taken to coincide with the construct score.

**Do true scores exist?**

The identification of true scores with constructs is a serious mistake that, unfortunately, permeates much of the literature on psychological measurement. The fact that true scores cannot be considered in this way does not, however, entail that true scores cannot exist. We may suppose that true scores and construct scores both exist, but are not identical; for example, we could imagine true scores to exist quite independently of the construct, but to be systematically related to that construct. This is the way Lord & Novick construct the relation of measurement, as we have seen, and it is also the way that latent variable models sometimes formulate the situation. The question then becomes how the existence of true scores could be interpreted. Is there a plausible interpretation that could locate the true score in reality, i.e., conceive of it as an objectively existing entity, without becoming inconsistent or downright absurd? It is argued in this section that such a realist interpretation is unreasonable. When the classical test theorist invites us to imagine the existence of a true score, most of us will be inclined to grant him this much. We
will see, however, that it is completely unclear what we are supposed to imagine. The reason for this is that it is difficult, or even impossible, to give a serious account of the distribution on which the true score is defined. The problem is that the thought experiment, that is supposed to define this distribution, does not specify sources of random error, and that the almost universally endorsed interpretation of random error is circular. Moreover, the assumption that true scores exist in reality does not lead to testable predictions, which strongly invites the application of Occam’s razor – especially because the true score leads to a needless multiplication of theoretical entities, which is undesirable.

**Where does error come from?** The conceptualization of the true score as an expected value is ill-defined. For it is entirely unclear under what circumstances the replications mentioned in Lord & Novick’s brainwashing thought experiment should occur. The primary problem is that it is unclear where the random variation is supposed to come from. This issue is usually circumvented in treatises on psychological measurement. These suggest that random error is due to unsystematic factors affecting the observations. For example, the typical examples of unsystematic errors are: Mr. Brown had a headache at the particular testing occasion; Mr. Brown accidentally filled in “yes”, while he intended to fill in “no”; Mr. Brown was distracted by the noise of schoolchildren playing nearby, etc. However, identifying this, in itself reasonable, conceptualization of random error with the formal term indicated by \( E_i \) is circular.

To see this, first recall that the true score cannot be conceptualized as the average score over actual replications. This would violate the basic assumptions of the model, especially those concerning independence and parallelism of repeated measurements. For the same reason, error cannot be conceptualized as the lump sum of all variables that cause variation in the observed scores over actual replications: The true score is defined through a thought experiment, and so is the error score. Further, we have seen that the semantics of classical test theory do not only require that Mr. Brown is brainwashed inbetween measurements, but also that Mr. Brown takes a trip in a time-machine inbetween measurements, because the true score must be conceptualized as being instantiated at a particular time point. What, then, is supposed to cause the fluctuations, that might generate the probability distribution on which the true score is defined, on this particular time point? In other words: What varies in the replications under consideration?

There are three possible answers to this question. The first is: Nothing. In this interpretation, we have a quite mysterious source of randomness, which is supposedly inherent to Mr. Brown himself. Test theorists holding this interpretation should definitely have a chat with people working in quantum mechanics, for it would follow that human beings and quarks have more in common than one might think. But certainly, the random error would not come from variations in ‘irrelevant’ variables, because there would not be variation at all. This interpretation does therefore not return the typical idea of random error as discussed above.

The second answer to the question is: Everything. Now we imagine Mr. Brown taking his United Nations test not only in the original testing situation, but also
2.2 Three perspectives on the true score

in the jungle, in space, under water, while playing a game of tennis, and so on. This interpretation, however, neither returns the typical idea of random error. For nothing prohibits Mr. Brown’s constitution to be changed in such a way that, say, his social desirability level goes down, or, more drastically, he turns deaf, or, still more dramatically, he becomes identical to a different person (say, Kofi Annan). Therefore, this interpretation forces us to include under the header ‘random error’ factors that we do not usually view as such – social desirability, for instance, is the classic example of a variable that is supposed to influence test scores systematically, not randomly.

The third answer that we may give is: Some things will change, and some will not. This, however, requires that we distinguish between factors that are variable across replications, and factors that are constant. Doing this allows us to create the desired interpretation of random error, but at the price of circularity. For which things are supposed to change in order to return the desired interpretation of random error? Well, those things that are supposed to be unsystematic. Which things are that? Supposedly, Mr. Brown’s headache, schoolchildren playing nearby, etc. But why these things? Because they are influential and change across replications. And why do they change? Because we have included them as varying in the thought experiment. Now we are back at square one. Thus, a platonic conception of error, as reflecting unsystematic influences on the observed test score, involves a circularity in reasoning. It actually allows us to create any interpretation of random error we desire, by incorporating the factors we want to subsume under that header as variable in the thought experimental replications. Nothing is gained in this interpretation.

Clearly, the true score is ill-defined as an expected value, because the distribution that is supposed to generate it cannot be characterized - not even roughly. The thought experiment that should do this does not specify the conditions under which replications should occur, except for the fact that these should be statistically independent, which, ironically, is exactly the reason that such replications cannot in general be equated with actual replications. Moreover, there is a serious problem in the interpretation of the thought experimental replications. Not only does classical test theory fail to provide grounds for choosing between the above accounts of random error, but the available accounts are either mysterious, inadequate, or circular. The thought experiment does not elucidate the situation. Mr. Brown’s brainwash adds little to the syntactic formula \( t_i = \mathcal{E}(X_i) \), but rather obscures the fact that taking the expectation of a distribution, which is defined at a particular moment on a particular person, is a doubtful move. Thus, when Lord & Novick invite the reader to assume the existence of a true score, it is not at all clear what the reader is supposed to believe in.

The multiplication of true scores The true score is ill-defined, but this, in itself, is not sufficient reason for rejecting the realist interpretation. Many concepts lack an unambiguous definition; surely, most psychological constructs do. The inability to define a construct unambiguously does not force us to the conclusion that the phenomena denoted by that construct therefore cannot exist. In many cases, definitions are the result of doing research, not a prerequisite for it. Indeed, much
True scores

scientific progress can be described in terms of a continuous redefining of scientific constructs.

However, what we may require from a realist interpretation of true scores is some kind of testability. This does not mean that theories must be falsifiable in the strict sense of Popper (1959) - in psychology, this would probably leave us with no theories at all - but there must be some kind of connection to observations that takes the form of a prediction. In theories of psychological measurement, this connection usually takes the form of discriminative hypotheses. For example, the intelligence tester may concede that he cannot give a definition of intelligence, but he can formulate the hypothesis that the number series ‘1 1 2 3 5 8 ..’ does measure intelligence (in a population of normal adults), while the item ‘I like to go to parties’ does not. This is, for example, the way that constructs are related to testable predictions in latent variable modeling. In the case of true score theory, no such connection can be made. There are two reasons for this. First, according to the classical test model, a distinct true score exists for literally every distinct test. Second, the theory cannot say what it means for two distinct tests to measure the same true score, except through the awkward requirement of parallelism. Therefore, the true score hypothesis does not yield testable predictions in the discriminative sense discussed above.

Consider the first point. The definition of the true score as an expected value leaves no room for saying that some tests do measure a true score, and some do not: We may always imagine a series of thought experimental replications and define the true score as the expected value of the resulting distribution. This means that every imaginable test has an associated true score, as has been illustrated in the previous section. Admitting the true score into reality thus forces the conclusion that every person is a walking collection of infinitely many true scores - one for every imaginable testing procedure. It would seem that, in this way, reality gets rather crowded.

Second, classical test theory cannot posit the true score as a hypothesis generating entity. This could, in principle, be done if it were reasonable for, say, the intelligence tester, to say that a number series item measures the same true score as a Raven item, similar to the way different items can be related to a single latent variable in item response models. Within true score theory, the only way to say that two tests measure the same true score is by saying that the tests are parallel. However, there is absolutely no reason to suppose that two distinct items that measure the same construct should be empirically parallel. Moreover, it has been shown in section 2.2.2 that the very idea, that two items that are empirically parallel measure the same true score, is inconsistent in its own right: The only item that could be said to measure the same true score as the number series item ‘1 1 2 3 5 8 ..’ is the number series item ‘1 1 2 3 5 8 ..’ itself. Of course, one could reason that two items that measure the same construct should have, for example, perfectly correlated true scores. This does yield testable predictions, but these do not result from the true score hypothesis itself. Rather, they result from a hypothesis concerning relations between true scores; a hypothesis that, in turn, is based on the idea that the items measure the same construct – in fact, it is based on a latent variable hypothesis and specifies Jöreskog’s (1971) congeneric model. The construct theory can specify
testable discriminative hypotheses ("these items measure intelligence, but those do not"), but the hypothesis that there exists a true score for a given measurement procedure cannot.

Thus, upon a realist interpretation, the true score is a metaphysical entity of the worst kind: Posing its existence does not lead to a single testable hypothesis. This does not mean that true scores, or classical test theory, are useless; obviously, the true score may figure in a set of hypotheses based on substantive theory, as it does in the congeneric model. It means that the true score hypothesis in itself is not capable of generating testable predictions.

Operationalism and true score theory

Two conclusions must be drawn. First, it is unclear what a true score is, because the probability distribution that is supposed to generate it lacks sufficient specification. Second, the true score hypothesis, in itself, does not lead to predictions. Note that these conclusions are not problematic for the true score concept, or for classical test theory in general. They are only problematic for a full-blown realist interpretation of classical test theory. It seems that such an interpretation is untenable.

However, just like the lack of correspondence between construct scores and true scores should not, in itself, bother the classical test theorist, the fact that a realist conception of true scores is problematic does not pose a problem for classical test theory either. It does suggest that we consider different ways of conceptualizing the true score's ontological status. Since there are enough alternatives to realism, the question becomes within which of these the true score could find a home. An adequate account of the theoretical status of true scores also illuminates what kind of philosophical outlook would be consistent with an identification of true scores with construct scores. The observation, that such an identification is completely unreasonable, suggests that the philosophical viewpoint that is consistent with it will also be completely unreasonable. In fact, the philosophical viewpoint that is consistent with classical test theory (as well as with the identification of constructs with true scores) is the most unreasonable of all, namely operationalism.

Operationalism (Bridgman, 1927) holds that the meaning of a theoretical term is synonymous with the operations by which it is measured. Interestingly, we have seen that the true score is defined without reference to anything but a measurement process. The true score is thus completely defined in terms of a series of operations: It is the proportion of times Mr. Brown would be in favour of the United Nations if he were tested infinitely many times. That the operations in question are hypothetical, and cannot be carried out, is a peculiar feature of the true score, but it does not preclude the conclusion that the true score is defined in terms of these operations, which is consistent with operationalism.

The true score also has some typical problematic aspects that are essentially identical to those faced by the operationalist philosophy of measurement. It has been argued against that view, for example, that it leads to a multiplication of theoretical terms (Suppe, 1977). For example, suppose that the meaning of the theoretical term 'intelligence' is equated with the set of operations that lead to an IQ-score on the Stanford-Binet. It immediately follows that the WAIS, the Raven,
or any other intelligence test cannot also measure intelligence, because each test specifies a distinct set of operations. So, each measurement procedure generates a distinct theoretical concept. It is therefore conceptually difficult, if not impossible, for an operationalist to say what it means for two tests to measure the same construct.

In classical test theory, we face essentially the same problem. The true score is defined in terms of the expected value on a particular test, and since each test generates a distinct expected value, it generates a distinct true score. Moreover, when the classical test theorist tries to express the idea that two tests measure the same true score, he runs into troubles that are comparable to those facing the operationalist. The only option, that comes close to saying that two tests measure the same true score, is to invoke the idea of distinct yet parallel tests. This is not only a highly contrived and unnecessarily strict requirement, that in no way matches the intended meaning of the proposition that the Stanford-Binet and the WAIS measure the same construct; it is essentially a concealed way of saying that the only test that measures the same construct as test \( x \) is test \( x \) itself. The same conclusion would be reached by an operationalist.

The operationalist view also resolves some of the problems surrounding the true score, as exposed in previous sections. For instance, the operationalist does not refer to a construct or attribute score as something that exists in objective reality, and certainly not as something that exists independent of the measurement process. Thus, we cannot speak of length without mentioning a centimeter; we cannot speak of weight without referring to a balance scale; and we cannot consider intelligence apart from the IQ-test. To do this is, in the eyes of the operationalist, to indulge in intolerable metaphysical speculation. A difficulty for operationalism may be created by countering that objects would still have a definite length if there were nobody to measure it, and that people already possessed intelligence before the advent of IQ-tests. This argument, which of course invites realism about such attributes, can be deflected by invoking scores on attributes as dispositional properties. Rather than accepting my height, which is about 185 cm, as a property which I have independent of any measurement apparatus used, the proposed solution is that height is a dispositional property with respect to a measurement procedure: if I were measured with a centimeter, then the indicated height would be about 185 cm.

It is interesting to inquire whether the true score could also be interpreted as a disposition, and, if so, what kind of disposition it is. In this context, Van Heerden & Smolenaars (1989) propose a taxonomy of dispositions by cross-classifying them with respect to the question whether they concern heterogeneous or homogeneous behavior, and whether this behavior is recurrent or unique. An example of a disposition that refers to the unique, non-recurrent, presentation of a specific piece of homogeneous behavior is ‘being mortal’. Saying that John is mortal expresses the fact that, upon the appropriate operations (e.g., poisoning him), John will decease. Dispositions may also refer to the recurrent presentation of homogeneous behavior. For example, the proposition ‘Professor G. likes to go out with his Ph. D. students at conferences’ refers to professor G.’s tendency to end up in a bar, in the company of his Ph. D. students, at conferences – usually at closing time. A proposition that
refers to the recurrent presentation of a heterogeneous set of behaviors is ‘Professor J. is forgetful’. This proposition refers to professor J.’s tendency to display a wide variety of behaviors recurrently; for example, professor J. regularly finds himself in the secretary’s office without remembering what he went there for, he forgets to remove the keys from his bicycle, etc. Finally, a disposition can refer to the unique presentation of a heterogeneous set behaviors. Van Heerden & Smolenaars (1989) give the example of ‘being intelligent’, which can be viewed as a disposition that may manifest itself once by completing the different (heterogeneous) tasks of an intelligence test.

The latter example is interesting, because Rozeboom (1966-a) treats test scores in a similar manner. He observes that, in order to conceptualize the idea that ‘every person has an IQ-score’, we must not interpret this sentence realistically (in which case it is obviously false), but in terms of the dispositional ‘for every person it is true that, if that person were tested, he or she would obtain an IQ-score’. In a similar vain, we can interpret the sentence ‘every person has a true IQ-score’ as ‘for every person it is true that, if he or she were tested infinitely many times (with intermediate brainwashing and time-travel), the resulting observed score distribution would have an expected value’. In the terminology of Van Heerden & Smolenaars (1989), the behavior under consideration could be heterogeneous (if it refers to total test scores) or homogeneous (if it refers to item scores), but it would certainly be recurrent, although with respect to thought experimental replications. The true score, as classical test theory defines it, may thus be considered to be a disposition, which specifies how a person will behave in a counterfactual state of affairs.

The overburdening of reality, which follows from the realist interpretation of true scores, dissolves upon this view. I am not a walking collection of true scores on all imaginable tests; but it is the case that, if I were repeatedly administered an arbitrary test (with the appropriate brainwashing and time-travel), then my true score on that test would take on a definite value. Such dispositions are thus not located in reality, but rather specify characteristics of subjunctive, or in our case counterfactual, conditionals (Rozeboom, 1966-a). It is certainly the case that the true score, with its thought experimental character, must be considered an oddball among recognized dispositional properties (e.g., solubility, fragility), because these usually specify responses, of the object to which the dispositional property is ascribed, in situations that could actually occur. The concept of fragility is considered to be dispositional, because it is characterized by conditionals such as ‘this vase is fragile, for if it were dropped, it would break’; and we may check this by actually dropping the vase. Similarly, all of the examples of Van Heerden & Smolenaars (1989) refer to actually realizable behaviors in realistic situations. This is the reason that Ryle (1949) characterizes dispositions as inference tickets. Rozeboom (1973) argues that such dispositional properties involve a realist commitment to underlying characteristics that generate the dispositions, and Van Heerden & Smolenaars

---

5 It is interesting to observe that, upon a dispositional interpretation of true scores, the stochastic aspect of classical test theory may receive a new interpretation. For instance, the value of the observed score variance for a given subject (normally interpreted as an index of measurement precision) would in this case reflect the strength of the disposition.
(1989) interpret dispositions as promissory notes (i.e., they involve a promise that, if one looks into dispositional properties, one could discover more fundamental laws that govern dispositional behavior).

The true score is not open to such an interpretation as long as the thought experiment is retained. For it refers to what would happen in an impossible state of affairs, namely to the observed score distribution that Mr. Brown would generate in the brainwashing thought experiment. So, the thought experimental character of the true score makes it quite useless as an inference ticket in the common interpretation of that term. A possible response to this problem, is to dispose of the thought experiment and to interpret true scores as propensities, without reference to limiting frequencies. I have already discussed this possibility in 2.2.1. Doing this, however, would preclude an interpretation of true score theory as a theory of measurement, for the only remaining connection to the theory of errors would disappear. In fact, classical test theory would become a statistical theory of dispositions; interpretations of observed score variance as 'random error', for example, would be out of the question, and the interpretation of the squared correlation between observed and true scores as 'reliability' would hardly make sense. Certainly, the founders of the theory did not intend classical test theory in this way, and its users do not interpret it so; in fact, it is likely that, interpreted as a statistical theory of dispositions for unique events, classical test theory would not appeal to those involved in psychological testing at all.

2.3 Discussion

Classical test theory was either one of the best ideas in 20th century psychology, or one of the worst mistakes. The theory is mathematically elegant and conceptually simple, and in terms of its acceptance by psychologists, it is a psychometric success story. However, as is typical of popular statistical procedures, classical test theory is prone to misinterpretation. One reason for this is the terminology used: If a competition for the misnomer of the century existed, the term ‘true score’ would be a serious contestant. The infelicitous use of the adjective ‘true’ invites the mistaken idea that the true score on a test must somehow be identical to the ‘real’, ‘valid’, or ‘construct’ score. This chapter has hopefully proved the inadequacy of this view beyond reasonable doubt.

The problems with the platonic true score interpretation were, however, seen to run deeper than a confound of validity and unreliability. It seems that the entire idea, that true scores are real entities, leads to a metaphysical explosion of reality. It was therefore argued that true scores are not to be granted a place in reality, but rather should be seen as a particular kind of score. And just as we do not say that every person has an IQ-score, but rather that every person would receive an IQ-score if tested, we have to refrain from saying that every person has a true score. What we can say is that every person would have an expected value, if he or she were repeatedly tested in a long run of testing occasions with intermediate brainwashing and time travel. The true score must thus be considered to be a dispositional concept, in particular as a disposition to generate specific long run frequencies.
The fact that not even such frequencies can be granted a place in reality, but must be considered in terms of a thought experiment, further degenerates the connection that the true score may bear to the real world.

A philosophy of measurement that could accommodate for these problems is operationalism. The true score has some problems that are similar to those of operationalism, and, at the same time, other problems can be resolved by introducing arguments that are similar to the operationalist defense. An operationalist interpretation could also restore the identity of constructs and true scores; this time, however, not by upgrading true scores, but by degrading constructs. Such an interpretation should therefore not be considered a platonic true score interpretation, but rather a nonplatonic construct interpretation. That is, if one is willing to give up the realist semantics of psychological constructs such as intelligence, extraversion, or attitudes, and to conceive of them in an operationalist fashion (with a touch of fictionalism to accommodate for the thought experimental character of the true score), then the true score could be a candidate for representing these constructs in a measurement model. The fictionalist element, which must be introduced because the true score, as a disposition, generalizes over a domain of impossible replications, precludes the interpretation of the true score as an inference ticket (Ryle 1949), or a promissory note (Van Heerden & Smolenaars, 1989). It also deprives the concept of the possible realist semantics that may be introduced for dispositions (Rozeboom, 1973), unless the entire idea that we are dealing with a theory of measurement is given up, by dismissing the thought experiment and disconnecting the theory from the theory of errors. I suspect that no classical test theorist will be willing to do this; classical test theory is intended, formulated, and used as a theory of measurement, and I do not expect classical test theorists to revert their self-image from ‘expert in measurement theory’ to ‘expert in dispositional psychology’. However, retaining the idea that we are dealing with a theory of measurement requires abandoning a realist interpretation of the true score, and taking an operationalist perspective.

Of course, since the true score is appropriately characterized as a product of the test theorists imagination, and therefore does not obtain a realist ontology, this is not a particularly pressing philosophical problem. At least, it is better than the kind of realism needed to localize the true score in the world. It is a pressing theoretical problem for psychology, however, because I do not think that many researchers in psychology are particularly interested in a true score, which specifies a disposition with respect to a set of impossible situations. So, once again we see the fundamental tension that Lord & Novick have introduced through their axiomatic treatment of test theory: The theory is constructed in such a way that it always works, but at the price of losing the natural interpretation of its central concepts. A psychologically meaningful interpretation of true scores and random error, as reflecting a stable characteristic and unsystematic variation respectively, is philosophically untenable. A philosophically acceptable interpretation of these concepts, as products of the imagination which refer to recurrent dispositions in a counterfactual state of affairs, is psychologically unattractive. Classical test theory systematically falls between these two stools.

It is my understanding that few, if any, researchers in psychology conceive of psychological constructs in a way that would justify the use of classical test the-
ory as an appropriate measurement model. Why, then, is the classical test theory model so immensely successful? Why is it that virtually every empirical study in psychology reports values of Cronbach's $\alpha$ as the main justification for test use? I am afraid that the reason for this is entirely pragmatic, and has been given in section 2.2.2: The common use of classical test theory does not involve testing the model assumptions. The lower bound strategy always returns a value for the internal consistency coefficient. In fact, this value can be obtained through a mindless mouse-click. Inserting the lower bound into formulae for disattenuating correlations between test scores, as advocated by Schmidt & Hunter (1999), will further allow one to boost validity coefficients to whatever level is desired. All this will come at no additional costs, for it does not require any of the tedious work involved in latent variable models, which moreover have a tendency to prove many of the commonly held interpretations of test scores illusory. Applying classical test theory is easy, and a commonly accepted escape route to avoid notorious problems in psychological testing, such as constructing unidimensional tests. The model is, however, so enormously detached from common interpretations of psychological constructs, that the statistics based on it appear to have very little relevance for psychological measurement. Coupled with the unfortunate misinterpretations of the true score as the construct score, of random error as irrelevant variation, and of reliability as some kind of fixed characteristic of tests, instead of as a population dependent property of scores, it would seem that large parts of the psychological community are involved in self-deception. Wishful thinking, however, is not a particularly constructive scientific procedure, and mystifying test theoretical concepts is certain to obstruct, rather than stimulate, progress in psychology. I therefore hope that the analysis reported here has added to the understanding and demystification of classical test theory concepts, and has made clear that much more is needed for an adequate treatment of psychological test scores.