ELSEVIER

Intelligence 30 (2002) 505–514

INTELLIGENCE

True scores, latent variables, and constructs: A comment on Schmidt and Hunter

Denny Borsboom*, Gideon J. Mellenbergh

Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

Received 27 April 2000; accepted 25 August 2001

Abstract

This paper comments on an article by Schmidt and Hunter [*Intelligence* 27 (1999) 183.], who argue that the correction for attenuation should be routinely used in theory testing. It is maintained that Schmidt and Hunter's arguments are based on mistaken assumptions. We discuss our critique of Schmidt and Hunter in terms of two arguments against a routine use of the correction for attenuation within the classical test theory framework: (1) corrected correlations do not, as Schmidt and Hunter claim, provide correlations between constructs, and (2) corrections for measurement error should be made using modern test theory models instead of the classical model. The arguments that Schmidt and Hunter advance in favor of the correction for attenuation can be traced to an implicit identification of true scores with construct scores. First, we show that this identification confounds issues of validity and issues of reliability. Second, it is pointed out that equating true scores with construct scores is logically inconsistent with the classical test theory model itself. Third, it is argued that the classical model is not suited for detecting the dimensionality of test scores, which severely limits the interpretation of the corrected correlation coefficients. It is concluded that most measurement problems in psychology concern issues of validity, and that the correction for attenuation within classical test theory does not help in solving them.

Keywords: correction for attenuation; classical test theory; modern test theory

* Corresponding author. Tel.: +31-20-525-6876; fax: +31-20-639-0026.
E-mail address: ml_borsboom.d@macmail.psy.uva.nl (D. Borsboom).

0160-2896/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S0160-2896(02)00082-X

1. Introduction

In a recent article in this journal, Schmidt and Hunter (1999) advocate the use of corrections for measurement error, especially the correction for attenuation. Their paper contains a number of arguments for a more widespread use of this correction, and, in addition, it addresses a number of objections against it. However, in our opinion, most of the presented arguments for the correction are either mistaken or misleading, and the most important objections against it are not discussed. Now it is true that "objections to the corrections can be found neither in the methodological literature [...] nor the substantive literature" and that "in this sense, these objections are part of the 'underground' or informal culture of psychological research." However, it is not true that these objections are "emotional and unthinking rather than rationally based" and that they "reflect an inability to think theoretically" (all quotations from Schmidt & Hunter, 1999, p. 189). There are, in fact, few good reasons for routinely applying such corrections within the classical test theory framework, and many good reasons for not doing so. The objective of this paper is to discuss the most important arguments against the routine use of the correction for attenuation as advocated by Schmidt and Hunter. Our discussion centers on two of the arguments against the routine use of the correction within classical test theory: (1) corrected correlations do not, as Schmidt and Hunter claim, provide correlations between constructs, and (2) corrections for measurement error should be made in a modern test theory model, instead of in the framework of classical test theory.

2. True scores are not construct scores

The Schmidt and Hunter paper is centered around the following argument. Theory testing means testing theoretical relations. These relations are between theoretical constructs and not between observables. Further, relations between observables are confounded with measurement error, since there is no such thing as errorless measurement. Therefore, we should estimate the theoretical relations instead of focussing on observed relations; and this means that we should correct for measurement error. Classical test theory provides such a correction in the form of the attenuation formulae. Therefore, the correction for attenuation should be routinely used in theory testing.

The above line of reasoning, however, rests on mistaken assumptions. Of particular interest here is the idea that true scores are construct scores, and that relations between true scores are relations between construct scores. This idea is central to Schmidt and Hunter's argument and is pervasive throughout their paper. This is most evident when they state that the corrected correlation coefficient between two measures x and y "is the estimated correlation between the construct underlying the measure x and the construct underlying the measure y" (p. 185), or when they speak of "the distinction between observed scores and construct scores (true scores)" (p. 189). True scores, however, are not construct scores, and neither do they necessarily reflect construct scores.

Although providing a definition of psychological constructs and construct scores is generally difficult, the true score has a clear definition in classical test theory. This allows

us to proceed from this definition to show that, upon any reasonable conceptualization of construct scores, these scores are not true scores. We first show that the identification of true scores with construct scores confounds issues of validity and issues of reliability. Second, we emphasize that such an identification is logically inconsistent with classical test theory itself.

The true score is one of the central concepts of classical test theory; actually, the derivation of the theory begins with a definition of the true score (Lord & Novick, 1968, p. 30). The true score of a subject is defined as the expected value of the observed scores, where the expectation is taken over an infinitely long run of independent repeated observations. So, for a person taking a psychological test, that person's true score is defined as the expected value of the observed scores over an infinitely long run of repeated independent administrations of that test. Such a long run of observations is unrealistic, of course, because human beings typically learn, fatigue, and change in many other ways during repeated administrations. As a result, repeated observations will not be statistically independent. Because the notion of independent replications is critical for the introduction of the probability model that Lord and Novick want to use, however, they introduce a thought experiment. In this thought experiment, the subject is brainwashed between each successive pair of measurements, so that the resulting observations may safely be considered independent. This allows Lord and Novick to define the true score as the hypothetical expectation over an infinite series of independent measurements. Of course, they readily admit that this definition is based on counterfactual premises and therefore has a limited interpretation. They use it primarily because it is mathematically convenient in defining some important concepts in classical test theory. For example, if observed scores are conceived of as composed of a true score plus random error (giving the classic equation $O_{\text{bserved}} = t_{\text{rue}} + E_{\text{rror}}$), it follows from Lord and Novick's definition of the true score that the expectation of the error scores is zero. This yields mathematically simple expressions for concepts such as reliability. The above definition of the true score, as the expected value over replications, allows Lord and Novick to define reliability mathematically as the ratio of true score variance to observed score variance across the population: Reliability is the proportion of true score variance that can be linearly predicted from the observed scores in a population of subjects (Mellenbergh, 1996).

Thus, in classical test theory, the true score is defined as the expectation of a hypothetical series of observed scores. Consequently, within the framework of classical test theory, the true score does not necessarily reflect a construct score, and it is certainly not identical with it—either by definition, by assumption, or by hypothesis. Classical test theory does, as a matter of fact, not assume that there is a construct underlying the measurements at all. 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," "$0.10/\sqrt{0.05 \times 0.05} = \ldots$," 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 total score on the test. This total 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 test–retest sense, because, in the general population, 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.

This argument shows that construct scores and true scores are different concepts. Now, reliability is exclusively about true scores, and so corrections for unreliability, such as the correction for attenuation, are also exclusively about true scores. Such corrections do not yield the correlation between construct scores, but instead provide a hypothetical correlation between hypothetical test scores; namely, the correlation between expected values on one test and expected values on another. To obtain the correlation between construct scores, on the other hand, one would need a correction for invalidity. This would depend on the construct definition and the conceptualization of a person's construct score, which are both difficult issues. Whether such a correction could exist is questionable; but that it is not the correction for attenuation is certain. Schmidt and Hunter's identification of true scores with construct scores obviously relies on a silent identification of issues of validity (which are about constructs) with issues of reliability (which are about true scores). That identification is false, as was recognized very clearly by Lord and Novick (1968), and has since been emphasized by various authors (see, for a good example, Lumsden, 1976). The adjective "true" in the term "true score" is a relic that stems from the times when reliability had not yet been clearly separated from validity. That was a mistake, and we should not make it again.

The distinction between construct scores and true scores is important from the perspective of validity, but the distinction actually goes much deeper. At this point, for example, the reader may be under the impression that *if* a test were valid (apart from the unavoidable random error), *then* the concepts of true score and construct score would coincide. Indeed, if such an identification of true scores with construct scores were compatible with classical test theory, one could at least characterize Schmidt and Hunter's position as logically consistent—be it under a substantial number of implicit assumptions. However, the identification of true scores with "valid" or "construct" scores is not, in general, compatible with classical test theory itself. It is, in fact, a widely recognized fallacy, known as the *platonic true score* interpretation (Lord & Novick, 1968; Lumsden, 1976; Sutcliffe, 1965). We can illustrate the fallacy with the following example, which is based on an example by Lord and 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 the construct score c on the nominal variable "having Alzheimer's." Thus, a patient who actually has Alzheimer's has construct score $c = 1$, and a patient who does not have Alzheimer's has construct score $c = 0$.

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 the

observed score X as follows. The value $X=1$ indicates the diagnosis "having Alzheimer's," and the value $X=0$ 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. What is the true score on the test? It is tempting to think that a patient's true score t on the diagnostic test is equal to the construct score (i.e., $t=c$). Specifically, the infelicitous use of the objective "true" suggests that a patient who actually has Alzheimer's, i.e., a patient with construct score $c=1$, also has a true score of $t=1$ on the test. For this indicates the diagnosis "having Alzheimer's," and it is, after all, true that the patient has that disease. Such an interpretation is called a platonic interpretation of the true score, because it conceives of the true score as the real score, i.e., as the construct score.

The platonic interpretation of the true score is not, in general, consistent with classical test theory. For suppose that the sensitivity of the diagnostic test is 0.80. 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 $c=1$. This patient's true score is not $t=1$, because the true score of classical test theory is equal to the expectation of the observed score, which is $t=E(X|c=1)=0.80$. Suppose further that the sensitivity of the test is 0.70. 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 $c=0$), the true score is equal to $t=E(X|c=0)=0.30$. 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 example contains random 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. We conclude that, if measurements contain random error, the identification of true scores with construct scores is logically inconsistent with classical test theory in general. Since errorless measurement does not exist, as Schmidt and Hunter correctly assert, this means that true scores cannot, in general, be equated with construct scores. 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). Lord and Novick further show that the platonic interpretation of the true score leads to a nonzero expectation of the error score and to a correlation between true scores and error scores (see Lord & Novick, 1968, p. 39 ff. for the technical details; also see Lumsden, 1976, who discusses measurement levels other than the nominal level). Under these conditions, many of the fundamental theorems of classical test theory collapse. We add that these include theorems on which the correction for attenuation is based.

It is clear that the identification of construct scores with true scores is fundamentally incorrect. Not only does it confound issues of validity with issues of reliability, but it is logically inconsistent with classical test theory itself. Both problems can be traced to the following distinction. The true score is an entirely *syntactic* concept (i.e., it is defined in terms of the mathematical syntax of classical test theory). The construct score, however, is a *semantic* concept, because it depends on the meaning that we assign to the score in terms of

the construct (see Messick, 1989 on this point). In other words, the two concepts do not differ in degree, but in kind. The question of how syntactic concepts in test theory should be related to semantic concepts in substantive theory is one way to frame the question of validity. Within classical test theory, however, equating them is not an option. In view of these problems, it is misleading to suggest that the correction for attenuation yields an estimate of the correlation between construct scores, as Schmidt and Hunter repeatedly do.

3. Why modeling makes a difference

The identification of true scores with construct scores is not only conceptually weak and logically inconsistent. It also reflects an outdated philosophy of science, namely operationalism. The identification implies that, for example, a person's score on the construct "intelligence" is identical with the expected score on a hypothetical long run of independent administrations of, say, the Stanford–Binet. Such a conceptualization is operationalist in nature, because the true score is an operationalist concept: It is defined in terms of a series of operations, namely in terms of hypothetical repeated administrations of the test in question. Operationalism, a branch of logical positivism originally conceived of by Bridgman (1927), has since long been discarded, and for good reasons (an overview of which is given by Suppe, 1977). For example, if the construct score on intelligence is equated with the true score on the Stanford–Binet, it immediately follows that that construct score is not identical with the true score on the WAIS, the Raven, or any other intelligence test, for these true scores are defined through a different series of operations. Thus, we need a new construct for each new test that we devise. Apart from the fact that such a view leads to an undesirable multiplication of constructs, this is simply not the way we think about concepts like intelligence—at least, not anymore. Rather, we conceptualize intelligence as a trait that underlies, or determines, each of the true scores on the different tests, but is not identical to any of them.

Fortunately, the past century has witnessed major advances in the development of measurement models that can exactly convey this idea (Heinen, 1996). Such models are generically known as latent variable models, but when employed in a measurement context, they are commonly referred to as models of modern test theory. Most of the parametric modern test theory models are particular instances of the Generalized Linear Item Response Theory (GLIRT) model (Mellenbergh, 1994). Examples of models subsumed under the GLIRT framework are the Item Response Theory (IRT) models of Birnbaum (1968) and Rasch (1960) for dichotomous item responses, the models of Bock (1972) and Samejima (1969) for polytomous item responses, and the congeneric model of Jöreskog (1971) for continuous item responses. The congeneric model is the model with multiple indicators per latent variable, which is often used in structural equation modeling. What connects these models is the idea that one or more latent variables can be used to explain the pattern of observed item responses. This idea is embodied in the hypothesis that measurements of the same latent variable will be statistically independent, conditional on that latent variable. In other words, the association between measurements vanishes once the underlying latent variable is taken into account. This is called local independence.

Modern test theory models differ from the classical model in conceptual basis and in statistical formulation. In this context, it is surprising that Schmidt and Hunter partially justify the use of the attenuation formulae by an implicit reference to models of modern test theory. It is, indeed, the case that true scores bear a very strong relation to latent variables. In fact, latent variable models can be formulated as models that relate the true score to a latent variable. It is also true that structural equation models, as well as other latent variable models, employ a correction for attenuation in the estimation of correlations between latent variables. Schmidt and Hunter use this connection as follows. They state that the fact, that true scores are so strongly related to latent variables, ''. . . is very fortunate: if this were not the case, it would be necessary to use the more complicated and difficult IRT measurement model in research'' (p. 185). The argument then seems to be that, since latent variable models are a good thing, and since they incorporate a correction for attenuation, the use of the correction for attenuation without an accompanying model is the next to best thing. This assertion is then justified by the fact that true scores and latent variables are strongly related.

However, the modern test theory models are substantially different from the classical model. As we have seen in the previous paragraph, the classical model is unrestrictive: It can be applied whenever the observed score has an expectation in a hypothetical long run of observations. In practice, this means that classical test theory is always applicable. To paraphrase Lord and Novick (1968, p. 48), it is a tautology rather than a model. This is not the case with modern test theory models, because these models place restrictions on the structure of observed data.

The fact that modern test theory models are so much more restrictive than the classical model has an important consequence for the present discussion. Namely, the restrictions imposed provide the possibility to check whether the model fits the data. For instance, suppose one is interested in the correlation between two latent variables, one underlying the observed scores on test x, and one underlying the observed scores on test y. This correlation can only be interpreted if each of the tests conforms to a unidimensional measurement model. This means that, for each test, one latent variable sufficiently explains the pattern of observed test scores. Now if such a model fits, then one can devise a substantial interpretation of the correlation between the two latent variables underlying the test scores. To obtain such correlations within a modern test theory model, the researcher is well advised to use general latent variable software such as Mplus (Muthén & Muthén, 1998). Mplus can be used to specify IRT models, as well as factor analysis models, and routinely estimates the correlation between latent variables. Now, if each of the tests is unidimensional, the correlation between latent variables should not be substantially different from the correlation that would have been obtained if one had used the correction for attenuation in the classical framework. Thus, under the condition that more restrictive unidimensional models fit each of the tests, the corrected correlation will be a reasonable approximation to the correlation between the latent variables in question, and Schmidt and Hunter's argument applies.

The problem is, however, that the argument cannot be reversed. From the fact that the true score will provide a good approximation to the latent variable of interest if a unidimensional modern test theory model fits the test scores, it cannot be concluded that the true score will provide such an approximation in general. Schmidt and Hunter do not discuss this problem

but simply state that "the relation is usually close enough to linear" (p. 185). This is questionable. The approximation holds only under the condition that a unidimensional model underlies the test scores. Such models are very restrictive and will not "usually" fit the scores on psychological tests. Often, there will be a number of latent variables underlying the observed test scores. If this is the case, it is hard to see in which sense the true score on the test could be an approximation to anything whatsoever. It is even more difficult to devise a meaningful interpretation of the corrected correlation between two test scores.

The reason for this is the following. Again, suppose a researcher is interested in the relation between the latent variables underlying test *x* and test *y*. Now imagine that, for the observed scores on these tests, a unidimensional measurement model does not fit. For example, test *x* does not measure one, but two latent variables, and test *y* measures three latent variables. In such a case, it would be nonsensical to represent the correlation between the construct underlying the observed scores on test *x* and the construct underlying the observed scores on test *y* in a single correlation coefficient. For it would be plainly inconsistent to say that neither test *x* nor test *y* measures one latent variable, but that the correlation between these two latent variables is .50. The correct conclusion in such a case would be that we cannot capture the relation between the constructs in a single correlation coefficient.

Now, if it is found that a unidimensional model is too restrictive for a given psychological construct, this does not mean that it is impossible to measure the construct or that the construct should be abandoned. It means that the situation is too complex for a unidimensional measurement model. Again, if such a measurement model does not fit the test scores in the above example, we should not conceptualize the relation between the constructs of interest in a single coefficient. If either of the constructs is not unidimensional, we need more than one latent variable to represent that construct. As a consequence, we need more than one correlation coefficient to represent the theoretical relation between the constructs of interest. We note that this is only one of the possibilities that modern test theory models have to offer. For instance, it could also be the case that we should conceptualize one or both of the constructs as a higher-order factor in a confirmatory factor model. Many other conceptualizations of the relation between constructs can be investigated (and tested) by means of modern test theory models. It seems to us that finding that a construct is not unidimensional should not paralyze but stimulate research in the area.

The classical model, in contrast, is blind to such important issues because it is untestable. Within the classical framework, the correction for attenuation will always return a single correlation coefficient for the relation between true scores—regardless of whether the constructs in question are unidimensional. In addition, the corrected correlation will always be higher than the observed correlation. However, the interpretation of the corrected correlation will be as limited as the classical model is unrestrictive. It is a truism of test theory that a stronger model yields stronger conclusions, and this truism applies equally to the interpretation of corrected correlations. The idea that the interpretation of the corrected correlation is independent of the modeling context can therefore only be based on wishful thinking. Within the classical framework, the corrected correlation will actually yield very little information about the relation between constructs, especially when compared to the amount of information that modern test theory models supply.

We conclude that the researcher, who exclusively relies on the classical model, may neglect the structure of psychological constructs. In effect, Schmidt and Hunter advise the researcher to assume that constructs are unidimensional, so that relations between constructs can be represented by a single correlation coefficient. This does not seem a constructive scientific strategy. With the presently available modeling techniques, it is not very difficult to check whether the required assumptions hold. If they do not, the researcher should either change the conceptualization of the construct in question, or revise the test. However, it would be a step back, and not a step forward, if the Schmidt and Hunter article would persuade researchers to use the correction for attenuation, instead of putting effort into making tests that can stand up to the demands of modern test theory and finding the appropriate model for their constructs.

4. Discussion

Psychology is a difficult discipline for many reasons. One of these reasons has to do with fundamental issues of measurement. The constructs that psychologists employ in their theories rarely allow for direct and accurate measurement. This is why it is generally difficult to assess the relations between constructs on the basis of empirical data. However, the main problem in psychological tests is not unreliability. Rather, it is the connection between our tests and our theoretical constructs that is problematic. In other words, the main issue in psychological testing is validity.

The correction for attenuation, however, is just a correction for unreliability and nothing more. As such, it can be useful in some situations. However, suggesting, as Schmidt and Hunter do, that the correction yields the correlation between constructs implies that it is a correction for invalidity. In our opinion, this is not only incorrect but also facile, because it suggests an easy way out of a difficult problem. Inferring relations between constructs on the basis of relations between observables is a difficult enterprise, which does not admit a simple solution. It is probably safe to assert that such a process of inference requires sophisticated arguments, which should combine theoretical considerations, empirical evidence, and psychometric knowledge (for a detailed discussion, see Messick, 1989). It is unlikely that a statistical correction formula could be a substitute for such an argument.

Furthermore, the correction for attenuation as employed in the classical framework is simply outdated. Modern test theory models are far more powerful than the classical model, both in determining the structure of constructs and in conceptualizing relations between them. Moreover, these models yield testable hypotheses, which is a virtue that cannot be overemphasized. This does not mean that modern test theory models buy one a ticket to theoretical heaven. Neither do modern test theory models solve the problem of validity—no psychometric model in itself can do that. They do allow, however, the researcher to see more clearly what is going on in the data and test hypotheses concerning the structure of psychological constructs. We therefore think that methodologists should encourage the use of such modeling techniques, instead of advocating a correction formula and conceptual framework that cannot perform any of these tasks.

Acknowledgements

The authors thank Conor Dolan, Gitta Lubke, and three reviewers for useful comments on earlier versions of this manuscript.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores reading* (pp. 397–423). Massachusetts: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bridgman, P. W. (1927). The logic of modern physics. New York: Macmillan.

Heinen, T. (1996). Latent class and discrete latent trait models: similarities and differences. Thousand Oaks: Sage.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, *27*, 251–280.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299.

Messick, S. (1989). Validity. In: R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Muthén, L. K., & Muthén, B. O. (1998). Mplus user's guide. Los Angeles, CA.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Paedagogiske Institut.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*, 183–198.

Suppe, F. (1977). The structure of scientific theories. Urbana: University of Illinois Press.

Sutcliffe, J. P. (1965). A probability model for errors of classification. *Psychometrika*, *30*, 73–96.