Original Article

Special Collection: Debunking Method Myths: Rethinking Conventional Wisdom About Methods in Psychological Measurement and Assessment, Other Areas of Psychology, and Related Fields

The Mythical Distinction Between Measurement and Prediction – Unifying These Approaches Through Estimation Theory

Steffen Zitzmann[1], Georg Krammer[2,3], Christoph Lindner[4], and Martin Hecht[3]

[1]Department of Psychology, Medical School Hamburg, Hamburg, Germany
[2]Institute of Business and Vocational Education, Johannes Kepler University Linz, Linz, Austria
[3]Department of Psychology, Helmut Schmidt University, Hamburg, Germany
[4]Department Clinical Translation & Research Hospital, Max Planck Institute of Psychiatry, Munich, Germany

Abstract: Over decades, the literature on educational and psychological assessment of latent attributes has proposed a number of approaches for obtaining and interpreting scores from tests and questionnaires, with measurement and prediction standing out. These approaches have traditionally been portrayed as fundamentally different, and it has been claimed that they would pursue different goals and have distinct interpretations. However, such views might reflect a myth – one perpetuated through repetition rather than theoretical necessity. With this article, we aim to challenge this long-standing belief by arguing that the distinction between measurement and prediction is not necessary. Specifically, we propose that both the measured score and the predicted score can be viewed as estimators of the same underlying quantity, the true score, thereby unifying them under the broader framework of estimation theory. This unification helps clarify conceptual relationships between assessment approaches and promotes greater coherence in how scores are interpreted across research and practice.

Keywords: assessment, latent attribute, measurement, prediction, estimation

Assessment is a cornerstone in all social sciences, particularly in education and psychology, and it is also deeply embedded in everyday life, where it serves as the basis for real-life decisions, for example, when evaluating a person's aptitude. In explaining the way latent attributes are quantified, many of us typically borrow from Stevens (1959) by stating that measurement in a broad sense is characterized by "the assignment of numerals to events or objects according to rule" (p. 25). We use measurement here consistent with Stevens's (1959) functional definition, while recognizing that contemporary literature places a much narrower, technical meaning on this term. Although Stevens's (1959) definition captures the essence of assessment (or measurement), it leaves open how different approaches put this idea into practice. Additionally, acknowledging the fallibility of each assessment, scholars have developed a variety of approaches for obtaining and interpreting scores from tests and questionnaires of which two have become particularly prominent in the psychometric literature: measurement in the narrow sense as grounded in classical test theory and prediction (e.g., Charter & Feldt, 2001). While the former emphasizes that the central result is a measure (typically in the form of the sum score), the latter defines the result as a prediction (typically in the form of the regression score; McDonald, 2011). To facilitate reading, we will hereafter consistently refer to these results as *measured score* and *predicted score*.

It is worthwhile to consider why the belief that these two types of score are fundamentally distinct may have emerged. One possible reason is that new ideas are often presented in ways that highlight their novelty and practical appeal. When a new method for assessing latent attributes

is introduced, it may therefore be tempting to emphasize an interpretation that appears especially intuitive or useful to researchers and practitioners, while downplaying the interpretive overlap with existing methods. A second reason may be the use of an implicit heuristic. Because the two scores are structurally distinct, that is, formally represented in different ways, this structural difference may easily be carried over into the conceptual domain, creating the impression that the scores must also differ at the level of interpretation. Finally, textbooks often discuss these scores in separate sections and note differences in their interpretation, sometimes even illustrating them with distinct examples (not necessarily using exactly our labels; e.g., Cohen & Swerdlik, 2005; Guilford, 1936; Gulliksen, 1950; Kelley, 1927; Lord & Novick, 1968; Nunnally & Bernstein, 1994; Sijtsma & van der Ark, 2021).

Recently, the debate has been fueled again. Stanley and Spence (2024) have made an important contribution by discussing different types of score. At the same time, in our view, their discussion reinforces the misconception by claiming that measured and predicted scores entail different models: the former being associated with the single-test-taker model and the latter with the many-test-takers model. Based on these differing emphases, the authors suggest that the appropriate score, whether measured or predicted, ultimately depends on the specific goal. When interest lies in the assessment of a single person, the measured score would be most adequate, whereas when one assesses a group of persons, for example, to select some of them (e.g., in hiring contexts), the predicted score should be preferred (but see Schmukle & Rohrer, 2025; Zitzmann, Lindner, et al., 2026, for divergent positions).

Through repeated presentation and reproduction of such claims, the impression of a fundamental distinction may gradually become entrenched. In Winnie-the-Pooh, two of the protagonists are unintentionally walking in circles, following their own tracks, which they mistakenly believe stem from an increasing number of creatures called woozles. The tracks make the protagonists believe that these woozles exist, and that they merely have to continue following the tracks to catch them. This story is typically used to illustrate how, through the widespread dissemination of a view, a group starts thinking that there is indeed truth in this view, thereby ultimately creating a myth. We argue that the belief that measured and predicted scores would be fundamentally different and entail different interpretations may be viewed as such a myth.

With this article, we aim to debunk this myth by critically reflecting on the often made distinction between measurement and prediction. Conceptual distinctions are valuable when they provide added value, whether by advancing a discipline through new insights or by promoting clarity and well-founded application. However, when such added value is absent, distinctions may become redundant. In some cases, they may even hinder progress by burdening the field with unnecessary categories and encouraging myths about distinct interpretations and applications. In the following, we will suggest that this can be avoided by recognizing that seemingly different concepts, such as the measured score and the predicted score, are conceptually of the same kind, even if they are not identical. Viewing both as estimators, and thus as instances of the same general principle of estimation, helps clarify their common conceptual basis: both aim to provide a judgment about a single person's latent attribute, that is, their true score. In other words, the core of our argument is the premise that both scores can be conceptualized as estimators, thereby unifying them under the umbrella of *estimation* theory. This unified estimation perspective contrasts with more traditional views that treat the scores as serving fundamentally different objectives, a view that we believe is ultimately more confusing than helpful.

The Myth, and a First Approach to Unification

According to the traditional view, the concept of measurement is closely linked to classical test theory (Lord & Novick, 1968; Novick, 1966; Nunnally & Bernstein, 1994) and related frameworks (e.g., domain sampling, generalizability, or G theory). The central assumption of these approaches is that there exists a latent attribute represented by a quantity called true score T that is measured by a score X, the measured score. Every measured score is contaminated with measurement error E. The relation between T, X, and E is typically summarized by an equation stating that the measured score equals the true score plus a measurement error, with the additional statement that the mean ($\mathbb{E}$) across the errors is zero. Formally, for a person with fixed T:

$$X = T + E\,, \tag{1}$$

$$\mathbb{E}[E] = 0\,. \tag{2}$$

Notice also that the amount of scatter of the errors σ_E determines the precision of, and thus the confidence one can place in, a measured score. At the level of the population, it is further assumed that true scores and errors are orthogonal, meaning their covariance is zero.

On the other hand, there is the regression score $\widehat{T}$, which has long been understood as a prediction for the true score

(Crocker & Algina, 1986; Kelley, 1927, 1947; Lord & Novick, 1968; McDonald, 2011; Stanley & Spence, 2024; Zitzmann, 2023) and is consequently referred to as the predicted score. This score is obtained by inserting a value for X into the linear predictor function resulting from regressing the true scores of many persons on their measured scores (and possibly also on additional person characteristics). In the simplest case, the predicted score is calculated as:

$$\widehat{T} = \rho \cdot X + (1 - \rho) \cdot \mu_X \tag{3}$$

for a given value of X. ρ denotes the classical reliability and thus the amount of variability in the measured scores that is accounted for by true scores. Unlike other reliabilities, it does not vary as a function of the true score's value. μ_X is the mean of the population distribution of all measured scores.

To better understand how the predicted score deviates from the measured score, let us consider an example participant, Matilda, who has taken an aptitude test. Reliability is known to be $\rho = .8$ and thus reasonably high, and the metric for the true scores is the IQ metric, which has a mean of 100 IQ points and an *SD* of 15 IQ points. Suppose Matilda's measured score is X = 132 IQ points. This number exceeds the threshold of 130 IQ points for exceptionally gifted people, suggesting that Matilda might be among them. However, because it is very high compared to the majority of others in the population, it is plausibly assumed to be in part due to a positive measurement error. The predicted score $\widehat{T}$ adjusts Matilda's result downward in the direction of the population mean of 100 IQ points to an extent determined by the reliability. It is obtained as $\widehat{T} = .8 \cdot 132 + 20 \approx 126$ IQ points, which is closer to the mean and no longer suggests Matilda's exceptional giftedness. At the same time, this makes the predicted score appear more likely compared to the possibly too unrealistic measured score of 132 IQ points (see, e.g., Wainer, 2000; Zitzmann, Lindner, et al., 2026).

Traditionally, the relation between the measured score and the true score is framed as a predictive relationship ("X predicts T"; McDonald, 2011). While still the most popular notion of the predicted score, this notion has been criticized. The criticism stems either from emphasizing that alternative interpretations could be equally justified or, in a more confrontational manner, from pointing out that this predictive interpretation is incorrect. The perhaps most influential criticism of the latter kind comes from Stanley (1970), who paraphrased Glass (1968) by arguing that to satisfy the requirements of classical test theory, the predicted score would have to fulfill the definition of a measured score, namely being composed of the true score plus an error. The author explained that this would be true because one could express the predicted score also as:

$$\begin{aligned} \widehat{T} &= \rho \cdot X + (1 - \rho) \cdot \mu_X \\ &= \rho \cdot (T + E) + (1 - \rho) \cdot \mu_X \\ &= \underbrace{\rho \cdot T + (1 - \rho) \cdot \mu_X}_{T'} + \underbrace{\rho \cdot E}_{E'} \end{aligned} \tag{4}$$

showing that this score is made up of a new true score T', which results as a linearly transformed version of T, and a measurement error E'. If we were willing to accept the usefulness of this transformation, we could agree with Stanley (1970) that the predicted score is another measure of the latent attribute rather than a prediction.

Returning to our example of assessing Matilda's aptitude, we can more clearly see how to interpret her predicted score as another instance of measurement. To do so, we must first define a "new scale" by linearly transforming the IQ. This transformation retains the mean of the IQ metric but produces a lower *SD* of only 12 IQ points. On this transformed scale, Matilda's predicted score of $\widehat{T} \approx 126$ IQ points can be interpreted as a measured score that exceeds the new threshold of 124 IQ points, reinforcing the conclusion that she might be exceptionally gifted. It is interesting to note that this conclusion is inconsistent with the one drawn under the traditional view of the predicted score.

From a practical standpoint, however, one might question the usefulness of altering the scale, especially because neither Matilda nor the assessor is familiar with this transformed scale. Therefore, it may be more convenient for both to interpret the result using the familiar IQ metric. Unsurprisingly, when Matilda's predicted score is transformed back to the IQ scale, it again equals 132 IQ points. Since this transformation leaves the conclusion unchanged, one may ask whether the two measured scores are merely different representations of the same measure, differing only by linear transformation of scale. Indeed, the seemingly different measures are identical up to linear transformation. Given that the metric of the transformed scale is likely perceived as less intuitive than the IQ metric, the predicted score – when reinterpreted as a measure – offers little practical advantage.

Summing up, Stanley (1970) sought to integrate prediction into classical test theory by demonstrating that predicted scores can satisfy its assumptions if the metric for the true score is adjusted accordingly. He argued that each type of score – whether measured or predicted – requires its own corresponding true score. Through this attempt to unify scores, prediction effectively becomes measurement. Overall, we consider his approach to be conceptually valuable but not practically compelling, agreeing with Feldt and Brennan (1989), who concluded that it "might be worth considering if

the examiner routinely used the estimated true score [*predicted score* according to our notation] in the interpretation of test results (which no one does)" (p. 358; see also Charter & Feldt, 2001). However, it is worth noting that Stanley's (1970) proposal was subsequently adopted in works examining how measurement error should be correctly accounted for in assessments (Glutting et al., 1987), with the resulting interval later implemented in widely used instruments, such as the Wechsler Adult Intelligence Scale, Third Edition (Tulsky et al., 1997). While arguments have been raised against adopting this interpretative framework, we nevertheless acknowledge the attempt to unify measurement and prediction under the same theoretical umbrella.

In the next section, we will propose an alternative and, in our view, more compelling framework for unification – one that challenges the myth that predicted and measured scores differ fundamentally from one another and, unlike Stanley's (1970) approach, preserves the intuitive notion of a single true score.

Toward an Estimation Theory Approach to Unification

Many scholars have argued that the development of integrative perspectives and overarching frameworks is central to scientific progress and therefore a valuable endeavor (Friedman, 1974; Salmon, 2006). Indeed, there are many examples that support this view, such as Maxwell's unification of electricity and magnetism within a single theoretical framework. Although many prominent examples come from the natural sciences, we see no reason why efforts toward conceptual unification should not also be pursued in other disciplines.

We are convinced that the unification of seemingly different types of score may constitute a step forward in the development of a sound psychometric foundation. More specifically, we will argue that types of score often treated as unrelated can be understood as instances of a broader statistical concept, namely the estimator. At the heart of our argument is an estimation-theoretic perspective that allows the two scores to be viewed from a more comprehensive vantage point, yielding a simpler and more coherent understanding and thereby challenging the myth that they are fundamentally different. By reducing the number of competing interpretations surrounding these scores, this perspective helps move the field from a fragmented view toward a more integrated one. As scholars such as Gijsbers (2013) and Kitcher (1981, 1989) emphasized, unification can enhance understanding by revealing how apparently distinct concepts are connected. In the present case, recognizing that measured and predicted scores are conceptually more similar than often assumed may provide researchers and practitioners with a perspective that affects how these scores are interpreted and used in the future. In this sense, unification may contribute to progress by helping to overcome redundant or misleading interpretations that continue to appear in research and practice.

The Concept of Estimator

An estimator can be understood as a calculation rule $\widehat{\theta}(D)$ that produces a numerical value (the estimate) for a fixed parameter of interest θ (the estimand) based on data D. If the random experiment were repeated, the estimand would remain constant, but the data would differ. Because the estimator applies a deterministic rule to the data, the resulting estimate would also vary. Consequently, estimates differ from the estimand and can be expressed as the sum of the estimand and the error. Technically:

$$\widehat{\theta}(D) = \theta + \varepsilon(D)\,. \tag{5}$$

The notation $\varepsilon(D)$ highlights that the error depends on the specific data set. For simplicity, and to align with common notation, we will omit D in subsequent expressions. Because the error changes across repetitions of the experiment, it is treated as a random variable. Following standard convention, we refer to the *SD* of this error as the standard error of the estimator, denoted σ_ε (e.g., Hastie et al., 2009; Kelley, 1947; Kendall, 1963; Wasserman, 2004).

Estimators may have different properties that impose certain constraints on their estimates. One important property is unbiasedness, which requires that estimates obtained from many repetitions of the experiment scatter in such a way around the estimand that their mean equals the estimand. Equivalently, the mean difference between estimates and estimand is zero:

$$\mathbb{E}\left[\widehat{\theta} - \theta\right] = 0\,. \tag{6}$$

If this difference is nonzero, the estimator is said to be biased, and the size of the difference is referred to as bias. Another property relevant to our discussion is efficiency, which compares the extent to which two estimators' results scatter around the estimand. An estimator $\widehat{\theta}$ is considered more efficient than another estimator $\tilde{\theta}$ if its mean squared error is smaller:

$$\mathbb{E}\left[\widehat{\theta} - \theta\right]^2 < \mathbb{E}[\tilde{\theta} - \theta]^2\,. \tag{7}$$

An estimator that is biased but has a smaller standard error than an unbiased one can be more efficient (e.g., Cole et al., 2014; Greenland, 2000; Kelley, 1947;

Wasserman, 2004; Zitzmann et al., 2021). A greater efficiency means that an estimate has a higher likelihood to be close to the estimand.

The manner of defining an estimator described above closely aligns with the paradigmatic view of assessment according to which we assign numbers to objects, with each assessment being subject to error. Analogous to how estimation theory distinguishes among estimand, estimate, and estimator, we distinguish among the latent attribute, the assigned score, and the rule for obtaining the score based on observation. Accordingly, assessment can be redefined as the process of estimating a latent attribute – or, more precisely, as obtaining an estimate of the true score through an appropriate estimator.

When discussing estimation, it is essential to determine the metric of the estimand first. Once the metric has been fixed, it provides a reference point against which different estimators can be evaluated and compared. Using different metrics across estimators should be avoided, as this would undermine any meaningful comparison among them (e.g., Klopp & Klosner, 2021; Schuberth et al., 2023). To continue with our example of assessing Matilda's intelligence, we adopt the common IQ metric as our metric of choice. In aptitude assessment, this metric is generally regarded as the most commonly used one. Of course, alternative choices are possible.

Having briefly addressed the important issue of fixing the metric, we will now turn to our central argument that the measured score and the predicted score are conceptually equivalent rather than fundamentally different. Specifically, drawing on estimation theory, we will argue that like the measured score in classical test theory, the predicted score can be viewed as a special case of an estimator.

Scores as Estimators

Let us first inspect the measured score, which – according to classical test theory – decomposes into the true score and a measurement error. This decomposition corresponds directly to our definition of an estimator. We need only consider a person's fixed true score as the estimand, and the difference between measured score and true score as the estimator's error:

$$\begin{aligned} X &= T + \{X - T\} \\ &= \underbrace{T}_{\theta} + \underbrace{E}_{\varepsilon}. \end{aligned} \tag{8}$$

In other words, the measured score satisfies the formal requirement of an estimator. Its bias and mean squared error are obtained as:

$$\mathbb{E}[X - T] = 0, \tag{9}$$

$$\mathbb{E}[X - T]^2 = \sigma_E^2. \tag{10}$$

This shows that the measured score is unbiased, and that its efficiency does not vary across different values of the true score as indicated by a mean squared error that is independent of the true score.

To see that the predicted score can likewise be interpreted as an estimator, yet as a different one, we continue to treat the true score as the estimand ($\theta = T$), not any linearly transformed version thereof. Computing the difference between predicted score and true score yields:

$$\begin{aligned} \widehat{T} &= T + \{\widehat{T} - T\} \\ &= T + \{\rho \cdot X + (1 - \rho) \cdot \mu_X - T\} \\ &= T + \{\rho \cdot (T + E) + (1 - \rho) \cdot \mu_X - T\} \\ &= \underbrace{T}_{\theta} \underbrace{-(1 - \rho) \cdot (T - \mu_X) + \rho \cdot E}_{\varepsilon'}. \end{aligned} \tag{11}$$

Whereas the error of the measured score corresponds directly to the measurement error and is thus unsystematic, the error of the predicted score (ε') has a more complex structure consisting of a systematic or deterministic component, $-(1 - \rho) \cdot (T - \mu_X)$, and a random component, $\rho \cdot E$.

To illustrate that interpreting the predicted score as an estimator can have practical implications, let us again return to our example in which Matilda's aptitude was assessed using both the measured and the predicted scores. Because we fixed the metric across estimators, both estimates can be interpreted relative to the same IQ metric. Specifically, Matilda's measured score of $X = 132$ IQ points remains higher than the vast majority of scores, suggesting that she may belong to the group commonly considered exceptionally gifted. However, this conclusion is not supported by the predicted score of $\widehat{T} \approx 126$ IQ points, which falls below the 130 IQ points threshold. Note that while this interpretation is consistent with the conclusion drawn under the traditional view, it contradicts Stanley's (1970) measurement perspective. According to Stanley's (1970) view, the predicted score results from a linear transformation of the scale, and such a transformation does not alter the conclusion. Having highlighted this important difference, we now turn to the properties that we ascribe to the predicted score from the perspective of estimation theory.

Regarding the first property, note that the presence of systematic error already explains why the predicted score is biased whenever reliability is less than perfect (except when the true score equals the population mean of measured scores). The bias can be expressed as:

$$\mathbb{E}\left[\widehat{T} - T\right] = -(1-\rho)\cdot(T-\mu_X)\,. \tag{12}$$

The direction and magnitude of this bias depend critically on the value of the true score. If the true score lies below the distribution's mean, the bias will be positive; if it lies above the mean, the bias is negative. Moreover, the farther the true score deviates from the mean, the larger the magnitude becomes. The mean squared error of the predicted score can be readily computed using a well-known decomposition (e.g., Wasserman, 2004; Zitzmann et al., 2015), yielding:

$$\mathbb{E}\left[\widehat{T} - T\right]^2 = (1-\rho)^2\cdot(T-\mu_X)^2 + \rho^2\cdot\sigma_E^2\,. \tag{13}$$

It is a function of the true score, indicating that it depends on the true score's value. Mean squared errors are larger for extreme true scores that deviate strongly from the population mean than for less extreme ones. When comparing the mean squared error of the predicted score with that of the measured score, we find that the former is smaller than the latter if and only if the true score – or, more precisely, its deviation from the population mean – satisfies the following condition:

$$(T-\mu_X)^2 < \frac{1+\rho}{1-\rho}\cdot\sigma_E^2\,. \tag{14}$$

This implies that despite being biased, the predicted score tends to be more efficient than the unbiased measured score for less extreme true scores, whereas the reverse holds for extreme true scores. Given the predicted score's inefficiency for extreme true scores, it is natural to ask whether it remains more efficient *on average* across all true scores. This question can best be addressed by integrating both mean squared errors over the distribution of true scores:

$$\int \mathbb{E}[X-T]^2\cdot p(T)dT = \sigma_E^2\,, \tag{15}$$

$$\int \mathbb{E}\left[\widehat{T} - T\right]^2\cdot p(T)dT = \rho\cdot\sigma_E^2\,. \tag{16}$$

Comparing the results, we find that the integrated mean squared error of the predicted score is smaller than that of the measured score if:

$$(1-\rho)\cdot\sigma_E^2 > 0 \tag{17}$$

which holds whenever reliability is less than perfect. We can therefore conclude that the predicted score is in general the more efficient estimator (see Lüdtke & Robitzsch, 2017; Wainer & Thissen, 2001; Zitzmann, 2025; Zitzmann et al., 2024).

Thus far, we have argued that both the measured score and the predicted score can be interpreted as estimators, thereby challenging the myth that these two would be fundamentally different. From this perspective, the two scores can be directly compared in terms of the properties that arise from their shared nature as estimators. Although the predicted score yields biased estimates, these estimates scatter less around moderate true scores than those produced by the measured score. Consequently, the predicted score is particularly advantageous for estimating true scores in the middle range of the distribution. This advantage turns, however, into a disadvantage for extreme true scores at the lower or upper ends of the distribution. Nevertheless, as demonstrated, the predicted score proves to be more efficient when the mean squared error is averaged across the entire range of true scores, which may explain why psychometricians have largely accepted its use in assessment (Dudek, 1979; Lord & Novick, 1968; Nunnally & Bernstein, 1994; Schmukle & Rohrer, 2025). However, while efficiency is particularly important when the goal is to describe a person's latent attribute by assigning a value that is likely to be close to the true score, the efficiency of an estimator does not necessarily imply that its decision-theoretic properties are also optimal. Consider the case of a decision, for example, whether someone is admitted to a program, that depends on a fixed threshold. Figure 1 shows the probabilities that each score leads to correct and incorrect decisions as a function of the true score. These probabilities allow us to assess under which conditions the more efficient predicted score performs better than the measured score in decision-making situations. As can be seen in the figure, the predicted score is less likely to lead to a true-positive decision than the measured score (top left panel). This increases the risk of a false-negative decision, especially when the true score is close to the threshold of 130 IQ points (bottom left panel). At the same time, the predicted score is more likely than the measured score to yield a true-negative decision (top right panel) and thus reduces the risk of a false-positive decision (bottom right panel). This pattern suggests that the predicted score can be interpreted as the more conservative score when deciding whether a person should be classified as exceptionally gifted. However, it is important to note that the choice between the two scores depends on how the consequences of correct and

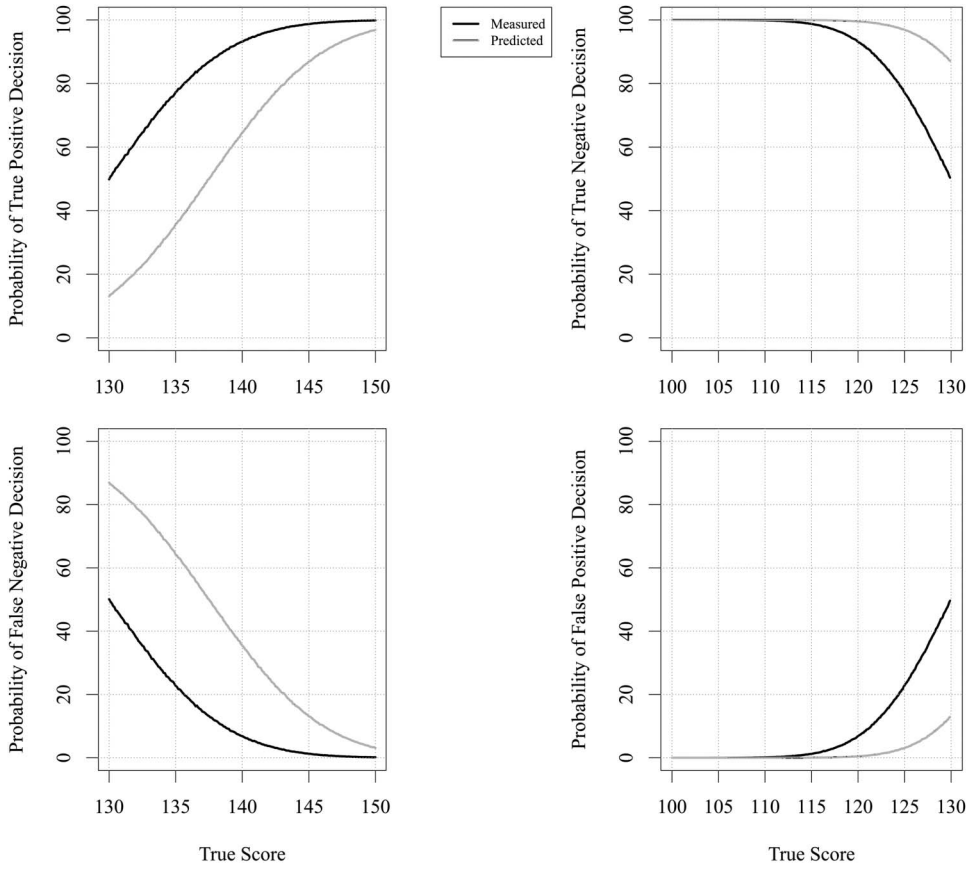

Figure 1. Probabilities of correct and incorrect decisions as a function of the true score.

incorrect decisions are weighted, which in turn is shaped by the specific context. For example, incorrectly classifying an exceptionally gifted student as not gifted, that is, making a false-negative decision, may deny that student access to advanced learning opportunities, which can lead to disengagement from learning and thus educational disadvantage. One could argue that this consequence is particularly serious and should be avoided, which would favor the use of the measured score.

Discussion

Researchers and practitioners employ various approaches to obtain and interpret test scores. Among these, two stand out in particular: measurement and prediction. Traditionally, these two have been portrayed as fundamentally different (e.g., Cohen & Swerdlik, 2005; Guilford, 1936; Gulliksen, 1950; Kelley, 1927; Lord & Novick, 1968; McDonald, 2011; Nunnally & Bernstein, 1994; Sijtsma & van der Ark, 2021) and as being associated with distinct goals and interpretations (Stanley & Spence, 2024). We suggested that this long-standing distinction might reflect a myth – a view perpetuated by tradition rather than theoretical necessity. Rather than treating them as distinct types of score and thereby reinforcing the myth, we argued that this distinction is unnecessary, and that both the measured and the predicted scores can be subsumed under the umbrella of a single overarching theoretical framework. Estimation theory provides such a framework, unifying the two types of score by treating them as qualitatively similar, yet quantitatively different special cases of estimators. The adoption of the estimation theory perspective naturally aligns with Lundberg et al.'s (2021) call to explicitly specify the core elements of estimation, namely the estimand, the estimate, and the estimator, thereby strengthening the link between theory and evidence. We clarify that the advantage of unification is primarily conceptual: contrary to traditional views and myths, the two scores are best understood as targeting the same underlying true score but differ in their statistical and

decision-theoretic properties. Both can be used to assess a person, although one may be preferable to the other depending on the context. When referring to a unification of the two scores, it is important to note that this does not imply that they are identical, isomorphic, or formally equivalent. Rather, our notion of unification involves reinterpreting both scores as estimators. The two seemingly separate scores are best understood as different special cases within a common conceptual framework, not as identical quantities. Accordingly, they cannot simply be substituted for one another, and any such reading would misrepresent our argument.[1]

Limitations

One limitation of our discussion so far is that we have not addressed the concept of reliability in detail, which plays an important role in assessment. This is why we believe that some reflections on implications for reliability are warranted. From the perspective of classical test theory, a reliability coefficient is expected to compare variability in true scores to the variability in measured scores via division of the former by the latter (Cho, 2016; Cronbach, 1951; Edelsbrunner et al., 2025; Lord & Novick, 1968; Zitzmann & Orona, 2025, 2026). The ratio is meaningful for the measured score because true score and measurement error are uncorrelated at the population level. This property does not change when the measured score is reinterpreted as an estimator. A different picture emerges for the predicted score, where this independence does not hold. True score and error are in fact correlated through the error's systematic part, which is a function of the true score. However, most researchers and practitioners might wish to use a reliability index as an indicator of the quality of their estimates. One possibility for defining reliability for predicted scores comes from a different strand of literature. Following scholars in the field of item response theory (e.g., Adams, 2005; Mislevy, 1992), we suggest defining reliability as the amount by which the predicted score reduces uncertainty in a person's latent attribute. Technically, this is achieved by dividing the variability in predicted scores by the variability in true scores, which seems to be the reciprocal of the classical definition. Interestingly, despite this apparent difference in definition, the results of both definitions converge to the same value (Lüdtke & Robitzsch, 2017; Stanley, 1970).

Practical Implications and Fairness Considerations

Beyond its conceptual contribution, the unification of both scores as estimators yields several concrete practical benefits that extend to the everyday work of researchers and practitioners. First, unification connects them to the broader statistical literature on shrinkage estimation, thereby providing the predicted score with an optimality justification that was formerly entirely invisible. Specifically, this score is a shrinkage estimator, which pulls the measured score toward the population mean, with the degree of shrinkage determined by the reliability. Shrinkage estimation has a distinguished history in statistics, with foundational results demonstrating that unbiased estimators can be inefficient, meaning that there exists another estimator with lower mean squared error. Under the traditional measurement-prediction dichotomy, this connection to shrinkage estimation and the concept of efficiency was completely obscured because the two scores were treated as belonging to different conceptual worlds. The unification of both scores makes it visible, situating the longstanding debate about measured versus predicted scores within one of the most productive traditions in twentieth-century statistics and giving researchers and practitioners a formal, mathematically grounded reason to prefer one score over another in assessment contexts. For example, when a person likely belongs to persons with true scores that are far from the population mean as in gifted assessment or any other context targeting the extremes of the distribution, their mean squared error grows large, making the measured score preferable. However, when the goal is to characterize any person from the population (i.e., no specific subgroup thereof), the predicted score is arguably the better choice because its integrated mean squared error is smaller than that of the measured score. This guidance is only possible because the unification of both scores as estimators provides the mean squared error as a common criterion.

Second, unification resolves a persistent source of confusion regarding the unit of analysis. Stanley and Spence (2024) argued that the choice between measured and predicted scores should depend on whether one is assessing a single person or a group. The unification renders this criterion irrelevant because both scores are estimators of the same true score. The unit of analysis is therefore not a relevant criterion for score selection. The relevant criterion is

[1] Note that estimating latent attributes can also be seen as a way of measuring – although not in the narrow sense in which classical test theory defines measurement. Yet, in a broader sense, consistent with Stevens's (1959) classic definition of measurement as the assignment of numerals to objects according to rules, any estimator performs a measurement function: It assigns persons values that reflect their standing on the latent attribute. Consequently, prediction is also a form of measuring a person, although with the term measurement understood outside the narrow framework of classical test theory.

purely statistical, implying that one need not first determine whether one is dealing with individuals or groups before selecting a score; instead, one can focus directly on the statistical properties that matter for one's specific application.

Third, unification renders the discrepancy between them directly interpretable as diagnostic information. Under the traditional view, the difference between the measured and predicted score was the difference between two conceptually distinct quantities and thus difficult to interpret in a unified way. Under a unified perspective, this discrepancy becomes simply the difference between two estimates of the same true score, and it is therefore directly interpretable. Returning to the example of Matilda, the 6 IQ-point difference between her measured score of 132 IQ points and her predicted score of 126 IQ points is a numerical statement about what the choice between the two estimators costs in this specific case. A large discrepancy serves as a warning signal that reliability is low, the observed score is far from the population mean, or both.

Fourth, unification provides guidance for improving scores through the incorporation of additional information. Whenever collateral information is available, such as scores from prior assessments, related instruments, or theoretically motivated covariates (e.g., Zitzmann, Orona, et al., 2025), it can be integrated into the estimator to increase its efficiency. Under the traditional dichotomy, it was conceptually unclear whether incorporating such information constituted a shift away from pure assessment. The unification dissolves this: incorporating information simply yields a more efficient estimator.

Fifth, unification also opens a principled avenue for addressing fairness considerations in assessment – a dimension that was difficult to formalize under the traditional measurement-prediction dichotomy. To see why, it is instructive to consider how each score relates to distinct sources of unfairness. One such source is specific to the predicted score and arises from its biased nature. Because this score pulls all measured scores toward the overall population mean, it systematically underestimates true scores for members of groups whose mean exceeds this overall mean, and systematically overestimates true scores for members of groups whose mean falls below it. Due to unbiasedness, the measured score does not share this property. Importantly, the overall mean is not a neutral reference point because it may itself be the product of inequalities, such as unequal access to education. Therefore, biasing toward this mean risks perpetuating these inequalities, a consequence that becomes apparent once both scores are unified as estimators and the role of the mean as one determinant of bias is recognized. A natural solution is to replace the overall mean with group-specific means, yielding group-informed estimators. Another source of unfairness concerns asymmetric loss functions across groups. The mean squared error treats positive and negative deviations from the true score as equally costly. In practice, however, the consequences of false-positive and false-negative decisions are often asymmetric, and these asymmetries may vary across groups. For example, failing to identify a gifted child from a disadvantaged background may carry more severe consequences than the same incorrect decision for a child from a privileged background because the latter typically has access to more compensatory resources. The unification of both scores as estimators offers a solution to incorporate such asymmetric consequences into score selection. Standard mean squared error can be replaced by an asymmetric loss function, and the predicted score can be modified in such a way that it becomes the optimal estimator under this loss. In this way, fairness becomes a principle for estimator construction rather than an afterthought applied post hoc to scores.

To conclude, measured and predicted scores have traditionally been treated as fundamentally different, which has encouraged an overemphasis on the nuances between them and, in turn, distinct interpretations for their use in research and practice. We suggest that the lack of a unified framework may have been one factor that facilitated the emergence of these divergent interpretations. More generally, the absence of such a framework can leave room for ambiguity and thereby create space for competing views of what these scores fundamentally are. Unification does not eliminate all differences between the scores, but it can help place those differences on a clearer conceptual footing. Our estimation theory approach represents such a theoretical advancement, serving as a safeguard against confusion, misinterpretation, and the emergence of further myths. It holds promise not only for improving current research and practice but also for enriching the education of future researchers and practitioners. We hope that this article motivates readers to reflect on what we have referred to as a myth and to consider our proposed way of guarding against it. We also invite readers to critically examine its practical usefulness. Time will tell whether estimation theory will become a widely adopted framework of interpretation in assessment and stand alongside other appealing approaches, such as Bayesian inference (e.g., Levy, 2009; Levy & Mislevy, 2017; Mislevy, 1986; Zitzmann, Lindner, et al., 2026).

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2-3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, *19*(4), 350–364. https://doi.org/10.1177/073428290101900404

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment: An introduction to tests and measurement* (6th ed.). McGraw-Hill.

Cole, S. R., Chu, H., & Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, *179*(2), 252–260. https://doi.org/10.1093/aje/kwt245

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*(2), 335–337. https://doi.org/10.1037/0033-2909.86.2.335

Edelsbrunner, P. A., Simonsmeier, B. A., & Schneider, M. (2025). The Cronbach's alpha of domain-specific knowledge tests before and after learning: A meta-analysis of published studies. *Educational Psychology Review*, *37*(1), Article 4. https://doi.org/10.1007/s10648-024-09982-y

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. H. Linn (Ed.), *Educational measurement* (pp. 105–146). Macmillan.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5–19. https://doi.org/10.2307/2024924

Gijsbers, V. (2013). Understanding, explanation, and unification. *Studies in History and Philosophy of Science*, *44*(3), 516–522. https://doi.org/10.1016/j.shpsa.2012.12.003

Glass, G. V. (1968). Response to Traub's "note on the reliability of residual change scores". *Journal of Educational Measurement*, *5*(3), 265–267. https://doi.org/10.1111/j.1745-3984.1968.tb00638.x

Glutting, J. J., McDermott, P. A., & Stanley, J. C. (1987). Resolving differences among methods of establishing confidence limits of test scores. *Educational and Psychological Measurement*, *47*(3), 607–614. https://doi.org/10.1177/001316448704700307

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, *29*(1), 158–167. https://doi.org/10.1093/ije/29.1.158

Guilford, J. P. (1936). *Psychometric methods*. McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Kelley, T. L. (1947). *Fundamentals of statistics*. Harvard University Press.

Kendall, M. G. (1963). *The advanced theory of statistics* (2nd ed., Vol. 1). Charles Griffin.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531. https://doi.org/10.1086/289019

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. C. Salmon (Eds.), *Scientific explanation* (*Minnesota Studies in the Philosophy of Science*, Vol. 13, pp. 410–505). University of Minnesota Press.

Klopp, E., & Klosner, S. (2021). The impact of scaling methods on the properties and interpretation of parameter estimates in structural equation models with latent variables. *Structural Equation Modeling*, *28*(4), 182–206. https://doi.org/10.1080/10705511.2020.1796673

Levy, R., & Mislevy, R. J. (2017). *Bayesian psychometric modeling*. CRC Press.

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, *2009*, Article 537139. https://doi.org/10.1155/2009/537139

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

Lüdtke, O., & Robitzsch, A. (2017). Einführung in die Plausible-Values-Technik für die psychologische Forschung [An introduction to the plausible value technique for psychological research]. *Diagnostica*, *63*(3), 193–205. https://doi.org/10.1026/0012-1924/a000175

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*(4), 511–536. https://doi.org/10.1007/s11336-011-9223-7

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195. https://doi.org/10.1007/BF02293979

Mislevy, R. J. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh Press.

Schmukle, S. C., & Rohrer, J. M. (2025). Clarifying the choice of confidence intervals in psychological testing: A comment on Stanley and Spence (2024). *Advances in Methods and Practices in Psychological Science*, *8*(2), 1–3. https://doi.org/10.1177/25152459251328281

Schuberth, F., Schamberger, T., Rönkkö, M., Liu, Y., & Henseler, J. (2023). Premature conclusions about the signal-to-noise ratio in structural equation modeling research: A commentary on Yuan and Fang (2023). *The British Journal of Mathematical and Statistical Psychology*, *76*(3), 682–694. https://doi.org/10.1111/bmsp.12304

Sijtsma, K., & van der Ark, L. A. (2021). *Measurement models for psychological attributes*. CRC Press.

Stanley, J. C. (1970). Definition of true score appropriate for estimated true scores. *Educational and Psychological Measurement*, *30*(3), 525–531. https://doi.org/10.1177/001316447003000302

Stanley, D. J., & Spence, J. R. (2024). The comedy of measurement errors: Standard error of measurement and standard error of estimation. *Advances in Methods and Practices in Psychological Science*, *7*(4), 1–20. https://doi.org/10.1177/25152459241285885

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman, & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 18–63). Wiley.

Tulsky, D., Zhu, J., & Ledbetter, M. F. (1997). *WAIS-III WMS-III technical manual*. The Psychological Corporation.

Wainer, H. (2000). Visual revelations: Kelley's paradox. *Chance*, *13*(1), 47–48. https://doi.org/10.1080/09332480.2000.10542192

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 23–72). Lawrence Erlbaum.

Wasserman, L. (2004). *All of statistics*. Springer.

Zitzmann, S. (2023). A cautionary note regarding multilevel factor score estimates from lavaan. *Psych*, *5*(1), 38–49. https://doi.org/10.3390/psych5010004

Zitzmann, S. (2025). Einzelfallbezogene Veränderungsdiagnostik [Diagnostics of individual change]. In R. Dohrenbusch (Ed.), *Psychologische Begutachtung: Rechtliche Grundlagen – Leitlinien – Empfehlungen* (pp. 461–469). Springer.

Zitzmann, S., Bardach, L., Horstmann, K. T., Ziegler, M., & Hecht, M. (2024). Quantifying individual personality change more accurately by regression-based change scores. *Structural Equation Modeling*, *31*(5), 909–922. https://doi.org/10.1080/10705511.2023.2274800

Zitzmann, S., Lindner, C., Herzberg, P. Y., Hecht, M., & Krammer, G. (2026). *And the comedy goes on: Individual scores, standard errors, and confidence intervals from the perspectives of Bayesian*

inference and classical estimation. [Manuscript submitted for publication]. Department of Psychology, Medical School Hamburg.
Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, *50*(6), 688–705. https://doi.org/10.1080/00273171.2015.1090899
Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the performance of Bayesian approaches in small samples: A comment on Smid, McNeish, Miočević, and van de Schoot (2020). *Structural Equation Modeling*, *28*(1), 40–50. https://doi.org/10.1080/10705511.2020.1752216
Zitzmann, S., & Orona, G. A. (2025). Why we might still be concerned about low Cronbach's alphas in domain-specific knowledge tests. *Educational Psychology Review*, *37*(2), 1–18. https://doi.org/10.1007/s10648-025-10015-5
Zitzmann, S., & Orona, G. A. (2026). When Cronbach's alpha does (not) indicate the reliability of domain-specific knowledge tests and why. *Frontiers in Psychology*, *17*, Article 1796702. https://doi.org/10.3389/fpsyg.2026.1796702
Zitzmann, S., Orona, G. A., Konig, C., Lohmann, J. F., Bardach, L., & Hecht, M. (2025). Novick meets Bayes: Improving the assessment of individual students in educational practice and research by capitalizing on assessors' prior beliefs. *Educational and Psychological Measurement*, *85*(3), 483–506. https://doi.org/10.1177/00131644241296139

History
Received December 3, 2025
Revision received April 9, 2026
Accepted April 20, 2026
Published online June 11, 2026

Section: Methodological Topics in Assessment

Conflict of Interest
The authors declare that there is no potential conflict of interest – neither financial nor nonfinancial.

Publication Ethics
Steffen Zitzmann and Martin Hecht are guest editors of the journal Psychological Test Adaptation and Development. They were excluded from the peer review and editorial evaluation of the manuscript prior to acceptance. The editor-in-chief, René Proyer, handled the peer review of this paper.

Authorship
Steffen Zitzmann: conceptualization, methodology, formal analysis, writing – original draft. Georg Krammer: conceptualization, methodology, writing – review & editing. Christoph Lindner: conceptualization, writing – review & editing. Martin Hecht: conceptualization, formal analysis, writing – review & editing, supervision. All authors approved the final version of the article.

Funding
There is no funding information to disclose.

ORCID
Steffen Zitzmann
https://orcid.org/0000-0002-7595-4736
Georg Krammer
https://orcid.org/0000-0002-1259-0349
Christoph Lindner
https://orcid.org/0000-0001-5688-3146
Martin Hecht
https://orcid.org/0000-0002-5168-4911

Martin Hecht
Department of Psychology
Helmut Schmidt University
Holstenhofweg 85
22043 Hamburg
Germany
martin.hecht@hsu-hh.de