Why Has IRT Not Replaced CTT?*

Zhan Peida

2026-08-25

ChinaRxiv

Abstract

Compared with classical test theory (CTT), item response theory (IRT) can characterize the mechanistic relationship between individual traits and item responses at the item level, reduce the dependence of item parameter estimation on specific samples, and evaluate measurement precision at different levels of latent traits. However, IRT has not replaced CTT practices centered on total scores and traditional item analysis in general psychological research. This phenomenon cannot be attributed simply to the technical complexity of IRT or insufficient training in research methods; rather, it should be understood by distinguishing among three levels: theoretical gains, practical benefits, and practical adoption. The additional information provided by IRT can be transformed into practical benefits only when it is relevant to the research question and scientific inference. Its practical adoption is also constrained by practical costs such as learning and using the theory, as well as by path dependence in the disciplinary ecology, including existing training systems and research workflows. Accordingly, this article proposes a method-adoption framework of "theoretical gains–problem demands–practical costs–disciplinary ecology" and advocates a shift from "method substitution" to problem-driven measurement modeling, that is, selecting a measurement theory according to the measurement resolution required for scientific inference and the corresponding costs.

Full Text

Why Has IRT Not Replaced CTT?*

Zhan Peida

(School of Psychology, Zhejiang Normal University, Jinhua, 321004, E-mail: pdzhan@gmail.com)

Abstract Item response theory (IRT), compared with classical test theory (CTT), can characterize at the item level the mechanistic relationship between individual traits and item responses, reduce the dependence of item-parameter estimation on a particular sample, and evaluate measurement precision at different levels of the latent trait. However, in general psychological research, IRT has not replaced CTT practices that are centered on total scores and traditional item analysis. This phenomenon cannot simply be attributed to the technical complexity of IRT or to insufficient methodological training; rather, it should be understood at three levels: theoretical gain, practical benefit, and practical adoption. The additional information provided by IRT can be translated into practical benefit only when it is relevant to the research question and to scientific inference; its practical adoption is also constrained by practical costs such as learning and using the theory, as well as by path dependence in the disciplinary ecology, including existing training systems and research workflows. Accordingly, this article proposes an adoption framework of "theoretical gain–problem demand–practical cost–disciplinary ecology," and advocates a shift from

"method substitution" to problem-driven measurement modeling, that is, selecting a measurement theory according to the measurement resolution required by scientific theory and the corresponding costs.

Keywords: item response theory; classical test theory; psychological measurement; psychometric model

1 Introduction: Why Has Theoretical Progress Not Brought About Method Substitution?

A basic difficulty in psychological research is that many core research objects cannot be directly observed. Psychological constructs/traits such as intelligence, personality, and anxiety can only be inferred from individuals' overt responses to items, tasks, or situations. In other words, many variables used in psychological research are not given directly by the data, but are measurement results formed through measurement theory, scoring rules, and model assumptions. Therefore, psychological measurement is not merely assigning numbers to psychological phenomena; more importantly, it specifies under what conditions observed data can reflect the target psychological construct and support the corresponding scientific theory.

Classical test theory (CTT) has long constituted the foundational framework for psychological and educational measurement. By distinguishing observed scores, true scores, and measurement error, it has provided a concise and widely applicable theoretical language for evaluating the stability of measurement results (reliability estimation) and for item analysis. Item response theory (IRT), by contrast, takes item responses (responses to test items) as the unit of analysis and uses probabilistic models to describe the relationships among individuals' latent traits, item characteristics, and item responses. Under the condition that the model fits the data, IRT can reduce the dependence of item parameters on a particular sample and evaluate measurement precision at different levels of the latent trait, thereby supporting applications such as test equating, differential item functioning analysis, and computerized adaptive testing (Luo Zhaosheng et al., 2012; Chen et al., 2025; de Ayala, 2022; van der Linden, 2016). Thus, relative to conventional CTT practice, IRT can provide more detailed item- and person-level measurement information.

If one follows a linear logic of technological development, a new theory that can provide more measurement information would seem as though it ought gradually to replace the old theory. However, no such simple substitution relationship has formed between IRT and CTT. On the one hand, IRT has been widely applied in large-scale educational assessment, item-bank construction, and computerized adaptive testing; on the other hand, in general psychological research, it remains common to aggregate item scores into a total score and then, through reliability

* Corresponding author: Zhan Peida, pdzhan@gmail.com
This research was supported by the Zhejiang Provincial Philosophy and Social Sciences Planning Leading-Talent Cultivation Special Project (25QNYC010ZD).

After confirming measurement quality through traditional item analysis, directly entering the total score into subsequent substantive statistical analyses (such as correlation, regression, or difference tests) remains a common practice. Reviews across multiple fields have shown that applications of IRT exhibit clear domain differences, and that its theoretical development and practical application are not synchronized (Foster et al., 2017; Schroeders & Gnambs, 2025; ten Holt et al., 2010). This gives rise to the central question addressed in this article: if IRT can provide richer and more refined information than CTT on certain important problems in psychological measurement, why have these theoretical gains not been correspondingly translated into practical advantages in general psychological research?

One intuitive explanation is that "IRT is too difficult." Compared with CTT, IRT indeed requires researchers to further understand concepts such as latent variables, item response functions, model parameters, and model fit, and it often imposes more explicit requirements on research design and data conditions. However, explaining the limited application of IRT solely in terms of technical complexity is insufficient. In recent years, open-source software, analysis programs, and methodological tutorials have continuously lowered the operational threshold for IRT (Tu Dongbo et al., 2023; Zein & Akhtar, 2025), yet its use in psychological research remains limited. This suggests that whether a measurement theory enters research practice depends not only on whether researchers "can use it," but also on whether the additional information it provides is genuinely needed for the research question, and on how much cost must be paid to obtain such information.

Therefore, this article understands the practical adoption of IRT from four interrelated levels. The first is theoretical gain—that is, what additional measurement information and inferential capability one theory can provide relative to an existing theory; the second is problem demand—that is, whether this newly added information is relevant to specific research questions and final scientific inferences; the third is practical cost—that is, how much cost in learning, data, and model use is required to obtain, test, and interpret this information; and the fourth is disciplinary ecology—that is, the training system, research procedures, software environment, and publication norms within which theoretical choices are situated. Together, these four factors determine whether theoretical advantages can be transformed into practical advantages: theoretical gain produces actual benefit only when it matches research demand; whether actual benefit can further promote theory adoption also depends on its costs and on the degree to which the method fits the existing disciplinary ecology.

Based on this line of thought, this article first compares the theoretical properties and experiential differences between CTT and IRT, and discusses the conditions and boundaries under which the proposition that "IRT is superior to

CTT" holds; it then explains the asymmetry between IRT' s theoretical development and practical application from four aspects: theoretical gain, problem demand, practical cost, and disciplinary ecology.

2 Comparison of the Theoretical Properties and Experiences of CTT and IRT

Although CTT and IRT are usually labeled classical psychometric theory and modern psychometric theory, respectively, they essentially address the same problem of measurement indirectness: researchers can infer the level of a specific psychological construct only from individuals' manifest responses to test items, experimental tasks, or other situations. The main differences between the two lie in how they use this observed information, how they describe the relationship between psychological constructs and manifest responses, and how they evaluate measurement error.

2.1 The Basic Logical Framework of Psychometric Theory

Psychometrics involves many aspects, including test construction, reliability and validity evaluation, score interpretation, and measurement models, but its core task can be summarized as follows: how to measure and infer psychological constructs that cannot be directly observed on the basis of observable behavior. Therefore, the core of psychometrics can be further summarized as psychological measurement modeling (Bollen, 2002; Sijtsma, 2012). Here, "modeling" does not mean that all psychological measurement must use complex statistical models; rather, it emphasizes that, from observed responses, psychological

the process of obtaining measurement results must always be based on certain theories, scoring rules, or model assumptions. Psychological measurement can therefore be summarized as a process of "target construct–manifest behavior–measurement model–quantitative inference" (Figure 1).

The true-score model in CTT assumes that an observed score consists of a true score and measurement error. Here, the true score is the expectation of the observed score obtained from infinitely repeated measurements of an individual under the same measurement conditions; it is a theoretical quantity in the statistical sense. CTT mainly develops around test total scores and their error properties, providing researchers with a simple theoretical framework for evaluating the stability of measurement results. IRT further implements this measurement relationship at the item level, using item response models to describe the probabilistic relationship among individuals' latent traits, item parameters, and observed responses. In other words, IRT attempts to explain the probability that an individual will make a certain response when the individual has a certain level of latent trait and the item has specific measurement characteristics.

It can thus be seen that CTT and IRT are not two mutually separate technologies. Both attempt to infer individuals' psychological construct levels from

Figure 1. Basic logical framework of psychometrics based on indirect measurement

Figure 1: Figure 1. Basic logical framework of psychometrics based on indirect measurement

manifest behavior; they differ only in the level of modeling and in the way parameters are specified. The former mainly models true scores and errors, whereas the latter describes, at the item level, the relationship between individual traits and item responses. Therefore, the important change represented by IRT relative to CTT can be understood as the ability to use more fine-grained information at the item and individual levels, and can also be summarized as an improvement in the resolution of psychometric modeling.

Figure 1 Basic logical framework of psychometrics based on indirect measurement

2.2 From test-level measurement modeling to item-level measurement modeling

The raw data of psychological tests are usually generated at the item level. For example, ability tests record whether an individual' s response to each item is correct or incorrect, and Likert scales record an individual' s graded choice for each item. In conventional CTT practice, these item responses are usually first summed or averaged to form a total score, and subsequent reliability evaluation and statistical analyses mainly center on this total score. IRT, by contrast, directly takes item responses as the unit of analysis and describes how individuals' latent traits and item characteristics jointly influence response outcomes.

Aggregating responses to multiple items into a total score is, in essence, a form of information compression. Information compression is not in itself problematic; the key is whether the discarded information is relevant to the research purpose. When items differ in difficulty, discrimination, and other aspects, the same total score may correspond to different response patterns, and such item differences have practical significance in certain measurement problems. However, it should not be assumed simply that "CTT only uses total scores, whereas IRT can necessarily distinguish different response patterns."For example, when the Rasch model satisfies the corresponding conditions, the raw total score can become

sufficient statistic for estimating an individual' s ability parameter; at this point, individuals with the same total score will receive the same ability estimate. The real distinctive feature of IRT is that it provides a set of probabilistic models at the item level, enabling researchers to specify clearly which item differences need to enter the measurement process and how these differences affect measurement results.

CTT and IRT also differ in how they describe item characteristics. In traditional CTT item analysis, item difficulty and item discrimination are usually empirical statistics calculated from a specific sample, and therefore vary with sample characteristics. IRT, by contrast, incorporates both individuals' latent traits and item parameters into the model. For example, in a two-parameter Logistic model, the difficulty parameter indicates the relative position of an item on the latent-trait scale, while the discrimination parameter reflects the item' s ability to differentiate individuals at different levels of the latent trait. This is also the basis for the relative invariance of item parameters and person parameters that IRT commonly emphasizes. However, this "invariance" is not unconditional. Only when the model fits the data well, the scale is properly determined, and the response mechanisms of different groups are basically consistent can item parameters remain relatively stable across different samples. If there is model misfit or differential item functioning, item parameters may still change. Therefore, rather than saying that IRT eliminates sample dependence, it is better to say that, under the conditions in which the model holds, it reduces the dependence of item parameters on a particular sample and makes whether the parameters have cross-group stability a testable question. Item-level calibration also allows different items to be organized onto a common scale, thereby supporting item-bank construction, equating and linking between different test versions, and computerized adaptive testing. In these applications, researchers need to know each item' s position in the entire measurement system and the information it can provide; relying only on total test scores usually makes it difficult to meet this requirement.

2.3. From Holistic Evaluation of Test Quality to Conditional and Personalized Evaluation of Measurement Precision

Another important difference between CTT and IRT lies in how measurement precision is evaluated. In conventional CTT practice, researchers usually use reliability or the standard error of measurement to evaluate the measurement quality of an entire test as a whole. This approach is simple and intuitive, and it also facilitates comparisons across different studies. However, the measurement precision of a test may not be the same for individuals at different trait levels. For example, an ability test composed mainly of medium-difficulty items may be able to distinguish individuals of medium ability fairly well, but may not further distinguish individuals with very high or very low ability. Therefore, "whether this test is reliable overall" and "how accurately this test measures individuals at a particular level" are not exactly the same question. In fact, because item difficulty in CTT depends on a particular sample, improving test reliability by adding medium-difficulty items essentially optimizes measurement precision only around the center of that sample; this implicitly reveals that the measurement precision of the same test differs for examinees at different ability levels. IRT directly describes this difference through the item informa-

tion function and the test information function. In IRT, different items can provide different amounts of measurement information at different locations on the latent-trait scale; correspondingly, the standard error of an individual' s latent-trait estimate also changes with the person' s latent-trait level. Thus, researchers can not only evaluate the overall performance of a test, but also further determine at which trait levels measurement is more precise and at which levels measurement error is larger.

This line of thinking also affects item selection and test design. The quality of an item depends not only on its difficulty and discrimination in the overall sample, but also on whether it can provide sufficient information within the trait range that the study truly concerns. For example, if the research focus is identifying high-ability individuals, then more items that can provide information at relatively high ability levels are needed; if attention is paid to a certain clinical cutoff, then it is necessary to ensure that the test has sufficient measurement precision near that cutoff. Computerized adaptive testing is precisely an important application of this idea. Based on an individual' s current trait estimate, it selects subsequent items from an item bank that are most informative for that individual, so that different individuals can receive different items while at the same time

obtain comparable measurement results on a common scale.

It should be noted that one cannot therefore simply conclude that "CTT can only provide an overall reliability, whereas IRT can evaluate the measurement precision of different individuals." Conditional standard error of measurement methods within the CTT framework can also estimate measurement error at different score levels (Feldt et al., 1985). A more accurate distinction is that conventional CTT practice usually uses overall indices to summarize the measurement quality of an entire test, whereas IRT directly incorporates the fact that measurement precision varies with the latent trait level into the model, and develops corresponding analytic and applied methods around this feature.

Therefore, this advance of IRT relative to conventional CTT practice should not be simply summarized as a move from "low precision" to "high precision." Rather, it should be understood as a further development from primarily evaluating the overall measurement quality of a test to evaluating measurement precision at different latent trait levels. Such additional information may have important value in clinical cutoff judgment, identification of individuals with extreme trait levels, and individualized measurement; however, if a study mainly focuses on relatively stable group mean differences or relationships among variables, its practical gain may be relatively limited. This further shows that whether IRT' s theoretical advantages can be transformed into practical advantages depends on whether this additional measurement information is relevant to the final scientific inference.

3 Why "IRT Is Superior to CTT" Cannot Be an Unconditional Proposition?

3.1 Stronger Theoretical Expressiveness Comes at the Cost of a More Complex Model Structure

The reason IRT can estimate item parameters, individuals' latent traits, and measurement information at different trait levels is that it makes more specific assumptions than CTT about the data structure of item responses. For example, IRT models usually need to specify the dimensional structure of the latent trait, whether items satisfy local independence, and what form the item response function takes. Thus, IRT does not obtain more information under conditions completely identical to those of CTT; rather, it does so by introducing more model structure.

More model structure brings advantages, but it also increases the requirements for model use. The more complex the model, the more specific the questions that researchers can answer, but the greater the possibility of improper model specification. Even if the model can successfully estimate parameters, it is still necessary to further evaluate issues such as overall model fit, item fit, local dependence, differential item functioning, and parameter stability (Sinharay & Monroe, 2025). In other words, while IRT provides more measurement information, it also requires researchers to assume greater responsibility for judging whether the model assumptions are reasonable.

By comparison, the model structure of CTT is relatively simple. Although it can directly provide less item- and person-level information, it can also discuss measurement error and score reliability under fewer assumptions (Sijtsma et al., 2024a). Therefore, the relationship between IRT and CTT should not be understood as a contrast between a "precise model" and a "crude model." More accurately, the two embody a trade-off between model complexity and information yield.

3.2 Theoretical Differences Do Not Necessarily Lead to Substantive Changes in Research Conclusions

Even if IRT can provide more refined measurement information, this does not mean that research conclusions will necessarily change markedly after IRT is used. Studies by Fan (1998) and Yu Jiayuan (1989) both show that some item and person indices obtained from CTT and IRT can have relatively high consistency; when items are relatively homogeneous and the measurement structure is relatively simple, CTT test total scores and IRT latent-trait estimates may produce highly consistent individual rankings. This phenomenon does not negate the theoretical value of IRT, but rather indicates that theoretical differences between methods do not necessarily always carry through to the final scientific conclusions. For example, if a study is primarily concerned with a strong overall correlation between two variables, or with comparing relatively clear mean differences between two groups, and the individual rankings and statistical results

obtained by different scoring methods are almost identical, then the additional information obtained by adopting a more complex model may

The information may not change the substantive conclusion.

Therefore, when evaluating measurement methods, it is necessary to distinguish between two questions: how much information a method can provide, and how much information the research question actually requires. A measurement method is not better simply because it is more complex or more precise; a more reasonable criterion is whether the measurement information and precision it provides are sufficient to support the final scientific inference.

3.3 The Total-Score Debate Has Changed the Simple Opposition Between IRT and CTT

Recent discussions surrounding total scores have further shown that it is inaccurate to understand IRT and CTT simply as an opposition between "latent-variable scores" and "total scores." The main issue with total scores is not that they are "simple," but whether their construction and interpretation have sufficient measurement justification. On the one hand, researchers have pointed out that even when total scores are highly correlated with estimates of latent traits, the two may still differ in their psychometric properties; therefore, in the absence of argumentation, one should not by default simply add all items directly (McNeish, 2023). On the other hand, under an appropriate measurement structure, total scores can also become reasonable and valid measurement outcomes (Widaman & Revelle, 2023). Sijtsma et al. (2024a) further pointed out that, under a class of commonly used IRT models, total scores can support ordinal interpretations of latent traits, while CTT can provide valuable reliability information under fewer assumptions. McNeish (2024) emphasized that total scores are mathematically reasonable, but this does not mean that they are the best choice for all problems such as prediction, multidimensional constructs, or measurement invariance. Mislevy (2024) likewise pointed out that the key to evaluating whether a score is reasonable lies in what kind of inference the score needs to support. Sijtsma et al. (2024b), in their response, also acknowledged that in applications such as test equating and computerized adaptive testing, latent-variable models can provide information that total scores themselves can hardly replace.

Thus, what this debate has truly changed is not whether total scores are ultimately right or wrong, but rather the criteria for evaluating methods. Whether something is mathematically valid, whether it is statistically more refined, and whether it is most appropriate in practice are three different questions. Total scores do not automatically lose value because IRT exists; likewise, the fact that IRT can provide more fine-grained information does not imply that all psychological research should use IRT.

3.4 The Application of IRT Shows Clear Domain Differences

To discuss why IRT has not universally replaced CTT, it is also necessary to understand with caution the judgment that "IRT is insufficiently applied." At present, there is a lack of unified bibliometric evidence covering all fields of psychology, so it is inappropriate to make a sweeping claim that IRT is used at a very low rate in "all of psychology." A more prudent statement is that existing research consistently shows that the application of IRT has obvious domain clustering, and that its degree of adoption differs across branches of psychology.

This difference is relatively clear between ability testing and personality measurement. Reise and Henson (2003) pointed out that IRT had already been widely applied in large-scale ability, achievement, and qualification tests, whereas its application in personality measurement was relatively limited. Research on psychological scales also shows a similar phenomenon. ten Holt et al. (2010), in examining 41 empirical scale studies, found that 78% used only factor analysis, 15% used only IRT, and 7% used both methods. This result also reminds us that, in general psychometric research, IRT faces not only CTT; factor analysis and structural equation modeling likewise constitute mature traditions of latent-variable analysis in psychology. IRT is mathematically closely related to factor analysis for categorical data (Takane & de Leeuw, 1987), but the two have formed different methodological traditions in teaching, software, and practical use. The situation in industrial and organizational psychology is also representative. Foster et al. (2017) reviewed IRT applications in related areas since 2000 and found that, although IRT had received attention from some researchers and practitioners, there were still clear gaps in its knowledge dissemination and practical adoption. Until recent years,

Schroeders and Gnambs (2025) still point out that the psychometric advantages of IRT have not yet been fully utilized in general psychological research.

However, the situation is clearly different in large-scale standardized assessment. Test-bank calibration, equating and linking across different test versions, differential item functioning analysis, and computerized adaptive testing all require item-level information and a common scale; therefore, IRT can directly serve the core measurement tasks in these areas. In recent years, IRT has also been used in cognitive ability measurement such as matrix-reasoning item banks, in order to support item calibration and test-bank construction (Zorowitz et al., 2024). In these contexts, researchers choose IRT not in pursuit of a "more advanced" method, but because the information provided by IRT is exactly what the research task requires.

This domain difference provides an important clue: if technical complexity were the only reason limiting the application of IRT, then large-scale assessment should likewise tend to use simpler methods. Reality shows, however, that as long as the additional information from IRT directly serves the core measurement task, its relatively high methodological cost may be accepted. Thus, differences in the adoption of IRT across domains themselves indicate that whether

a method is widely used is not determined solely by its theoretical performance, but also by the degree of fit between the method and specific research questions.

4 Why Have the Theoretical Gains of IRT Not Been Translated into Practical Advantages?

The preceding discussion shows that, compared with conventional CTT practice, IRT can provide more fine-grained information at the item and person levels, but these theoretical gains have not made it naturally become the default method in general psychological research. To understand this phenomenon, it is not enough merely to compare how much information the two theories can provide; it is also necessary to further examine whether this information is truly needed by the research question, how much cost must be paid to obtain it, and in what kind of disciplinary environment methodological choices occur.

Accordingly, this article explains the practical adoption of IRT from four interrelated aspects: theoretical gains, problem demands, practical costs, and disciplinary ecology. Theoretical gains answer what additional information the method can provide; problem demands determine whether this information is relevant to the ultimate scientific inference; practical costs involve the learning, data, and model-analysis investments required to obtain and interpret this information; and disciplinary ecology involves external conditions such as methodological training, research processes, software tools, and publication norms. Theoretical gains form actual benefits only when they match problem demands, and whether actual benefits further promote practical adoption depends on whether they are sufficient to offset practical costs and whether they can be absorbed by the existing disciplinary ecology. Therefore, the practical competitiveness of a method is not determined by theoretical performance alone.

4.1 Theoretical Gains: What Additional Information Can IRT Provide?

The theoretical gains of IRT are mainly reflected in its ability to make use of item- and person-level information that is less directly retained in conventional CTT practice. By simultaneously considering individuals' latent traits and item characteristics through item response models, IRT can, under the conditions in which the model holds, reduce the dependence of item parameters on a particular sample, evaluate measurement precision at different latent-trait levels, and further support applications such as test equating and linking, differential item functioning analysis, test-bank construction, and computerized adaptive testing (Chen et al., 2025; de Ayala, 2022).

However, being able to provide certain information does not mean that a specific study needs that information. For example, in cross-group comparisons, if it is necessary to determine whether certain items have different measurement functions across groups, then differential item functioning analysis may directly affect whether the conclusion is valid; in clinical assessment, if a study focuses

on individuals near a certain cutoff value, the measurement precision in that interval may also have direct decision-making significance. Conversely, if a study mainly uses a composite score with sufficient validity evidence to test a relatively clear overall effect, further estimating each item' s

···parameters may not change the final conclusions.

Therefore, the fact that IRT can retain more information first of all implies a potential theoretical gain, rather than an actual advantage that exists in every study. When evaluating a measurement method, it is necessary to distinguish what information it can provide from what information the current study truly needs. Only additional information relevant to the research goal can further generate practical benefits.

4.2 Problem Demand: Does Additional Information Enter the Final Scientific Inference?

The value of a method depends first on what inferences the research ultimately hopes to make. Many issues in large-scale assessment, personnel selection, clinical decision-making, and individualized evaluation occur directly at the item or individual level. For example, score comparability must be maintained across different test versions; cross-group studies need to examine whether items have the same measurement function; computerized adaptive testing needs to select the most informative items on the basis of estimates of an individual' s current latent trait. In these situations, item parameters, measurement precision at different trait levels, and covariances are not auxiliary information, but part of the research task itself. Therefore, the model complexity added by IRT is relatively easy to translate into practical benefits.

By contrast, a large number of general psychological studies mainly focus on group mean differences or overall relations among variables. If a measurement instrument can already form a composite score with sufficient reliability and validity evidence, then more fine-grained estimates of individual latent traits may not necessarily substantially change group differences, correlational relationships, or other main conclusions. At this point, although IRT can still provide more measurement information, the marginal value of this information for the final scientific inference may be relatively small.

This distinction is related to Cronbach' s (1957) distinction between two paradigms of psychological research: statistical inference and critical science. The former, represented by classical experimental research, more often seeks general laws at the group level through manipulation and control; the latter, represented by the study of individual differences, attends more directly to the location of stable differences among persons and to the quality of their measurement. IRT is more likely to show advantages in the latter type of question, but it cannot therefore be simply concluded that "experimental research is suited to CTT, whereas individual-differences research is suited to IRT." What truly determines methodological demand is whether the final

scientific inference depends on measurement information at the item and individual levels.

Moreover, this demand also changes as questions in psychological research change. In recent years, cognitive and clinical psychology have increasingly inferred latent cognitive processes from observed data such as accuracy and reaction time through computational modeling (Liu Yikang, Hu Chuanpeng, 2024); psychometric research has also begun to jointly use multimodal data such as response outcomes, response times, behavioral sequences, and eye-movement information to describe individual and item differences more fully (Guo Xiaojun et al., 2024; Zhan Weida, 2022; Fu et al., 2024; Myers et al., 2022; Zhan et al., 2018). When the parameters of these models are further used to study stable individual differences, their reliability, validity, and stability across time again become issues of psychological measurement; in recent years, cognitive and experimental psychological research has also begun to refocus on measurement calibration and the quality of measurement at the individual level (Zhu Qianqian et al., 2025; Bach, 2024; Clayson, 2024; Rappaport et al., 2025).

These changes show that "group laws" and "individual differences" are not fixed methodological camps. As research further shifts from "what happens on average" to "why different individuals produce different performances," the demand for more fine-grained measurement information may also increase. Therefore, differences in the adoption of IRT across different fields do not indicate that the methods in some fields are more "advanced" ; rather, they reflect differences in the demand for measurement information across different scientific questions.

4.3 Practical Costs: What Exactly Does "IRT Is Too Difficult" Mean?

Even if IRT can provide the information that research truly needs, whether it is actually adopted still depends on how much cost must be paid to obtain that information.

It is insufficient to explain the underuse of IRT simply as "IRT is too difficult." A more accurate understanding is that, compared with simpler scoring and analytic methods, IRT usually entails higher costs of method use, including learning costs, data and research-design costs, model-evaluation costs, and result-interpretation costs.

First are learning costs. IRT requires researchers to understand concepts such as latent traits, item response functions, model identification, and local independence, and to select appropriate models according to item type and response process. Different models are not merely different options in software; they correspond to different assumptions about the item response mechanism (Yang Xiangdong, 2010). Therefore, being able to run a model does not mean being able to use the model reasonably.

Second are data and research-design costs. Whether an IRT model can be

estimated stably depends not only on sample size, but also on the number of items, model parameters, the distribution of latent traits, and the parameters of research interest. Therefore, it is difficult to apply a fixed sample-size standard to all IRT studies. A more reasonable approach is to plan sample size according to the specific model and research goals (Schroeders & Gnambs, 2025). This also means that researchers often need to consider the subsequent measurement model before data collection, rather than deciding what analytic method to use only after data collection is complete.

Third are model-evaluation costs. Successful estimation of a model does not mean that it fits the data adequately. Researchers also need to evaluate issues such as overall model fit, item fit, local dependence, item functioning differences, and parameter stability (Sinharay & Monroe, 2025). The more complex the model, the more information it can provide, but the corresponding model assumptions that must be tested and interpreted also increase.

In addition, there are costs of result interpretation and communication. Total scores usually have relatively intuitive meanings and are easy to use in research reports and practical decision-making; estimates of latent traits in IRT depend on the specific model and scale settings, and their numerical values usually need to be interpreted in conjunction with the corresponding measurement framework. If the results provided by a complex model are almost the same as simple scores, then the additional analytic and interpretive costs may reduce its practical appeal.

In recent years, the development of R packages such as *mirt* and various specialized software programs has substantially lowered the computational and operational barriers to IRT. However, software can hardly replace the researcher' s responsibility for model judgment: Why choose this model? Do the model assumptions correspond to the actual response process? Are the data sufficient to support parameter interpretation? What does poor model fit mean? Therefore, what is today called "IRT is too difficult" is, more importantly, not computational difficulty, but the requirement that researchers more actively propose, test, and interpret measurement models.

It can thus be seen that the meaning of practical costs depends on the needs of the problem. When the information provided by IRT is directly related to core scientific inference, higher methodological costs may be worth bearing; when simple methods can already answer the question sufficiently, the same costs may become unnecessary complexity. Therefore, not using IRT cannot simply be understood as insufficient methodological training; it may also be a reasonable choice made by researchers on the basis of actual benefits and methodological costs.

4.4 Disciplinary Ecology: Why Do Default Methods Have a Huge Path-Dependence Advantage?

Researchers do not choose methods in a completely neutral environment. Which methods are easy to learn and operate, which results are easy to communicate and publish, and which analyses require additional explanation are all influenced by existing training systems, research traditions, and publication norms. Therefore, considering only the cost–benefit relationship of a single study is still insufficient to explain why IRT has not become the default method in general psychological research.

Psychology has already developed a relatively mature research workflow of "measurement instrument–scoring–statistical analysis." Researchers usually select an existing scale, perform necessary reverse scoring of items and then sum or average the scores, evaluate the reliability of the scores through indicators such as internal consistency, and then use the resulting variables for correlation, regression, analysis of variance, or structural equation modeling. McNeish (2022) summarized similar practices as the "sum-and-alpha approach." This research workflow is simple, efficient, and convenient for teaching and communication. Of course, the problem does not lie in the use of total scores or internal-consistency indices per se. As noted above, under appropriate conditions, total scores can be entirely reasonable evidence for psychological measurement. What deserves more attention is whether convenient scoring methods lead researchers to ask fewer follow-up questions: Why are these items able to reflect the target construct? Do scores have the same meaning across different groups? Do measurement errors affect the final scientific conclusions? Studies by Flake and Fried (2020) and Flake et al. (2022) both show that, in psychology, decisions concerning the selection of measurement instruments, construct validity, and scoring methods often do not receive the same degree of attention as subsequent statistical analyses. Similar problems also appear in psychopathology research, where the definition and measurement of psychological constructs such as depression, as well as their validity evidence, are still regarded as important issues that constrain theoretical interpretation (Fried et al., 2022; Hayden, 2022).

These phenomena have not emerged only recently. Meier (2023) pointed out that some measurement problems in psychology show obvious persistence, and that some long-standing issues have not naturally disappeared with the development of statistical methods. In this sense, psychology to some extent exhibits a tendency toward "heavy statistical inference and light measurement modeling." This is not to say that psychological researchers do not value measurement; rather, it means that researchers often invest more effort in answering "how variables should be analyzed after they have been obtained," while relatively less discussion is devoted to "how these variables are formed from observed behavior." The measurement-modeling approach represented by IRT precisely requires researchers to move these questions forward. For example, before forming a total score, one needs to consider why items can reflect the target latent trait, whether different items have the same measurement function, whether

the model is consistent with the data, and at which individual or trait levels the measurement is more precise. These issues often need to be considered as early as the stages of measurement-instrument selection, item design, and data collection, and cannot be fully remedied at the data-analysis stage. At the same time, default methods also have clear advantages in terms of path dependence. Total scores, α coefficients, and common statistical models have already become a shared language familiar to textbooks, software, researchers, and reviewers; using these methods usually does not require extensive explanation of "why they are adopted." By contrast, when using non-default methods such as IRT, researchers often need to provide additional explanations of model choice, parameter meaning, model fit, and scoring methods. This creates a practical asymmetry: default methods usually do not need to prove their own necessity, whereas non-default methods often first need to explain why they are worth using.

This difference is also related to the long-standing relative separation between psychology and psychometrics in methodological training. Borsboom (2006) pointed out that the distance between psychology and psychometrics is not only a technical issue, but is also related to the lack of sufficient integration among theory, training, and research practice. Wijsen et al. (2022) likewise noted that the high degree of technicalization of contemporary psychometrics, together with its distance from substantive psychological research, may make it difficult for psychometric methods to fully enter the core theoretical issues of psychology. Sijtsma (2012) also emphasized that the development of psychometrics needs to respond more directly to real scientific questions in psychology. General methodological training in psychology often places greater emphasis on how to analyze relationships among variables, while systematic training on how variables themselves are formed and what measurement properties they have is relatively limited. Therefore, the practical advantage of CTT-style simple scoring methods comes not only from technical simplicity, but also from their high compatibility with existing research workflows, methodological training, and publication norms. The so-called "heavy statistics, light measurement" is better understood as a relative allocation of methodological resources, rather than as a blanket criticism of psychological research practice.

Of course, this disciplinary ecology is not fixed. As psychology pays increasing attention to individual prediction, latent processes, and complex behavioral data, as computational modeling and other methods make the idea of "inferring latent processes from observed behavior" more widespread, or as research norms place greater emphasis on construct validity and

when measurement transparency is considered, the demands placed on measurement modeling may also increase accordingly. The practical value of IRT and other psychometric measurement models will likewise be reassessed as research questions, methodological costs, and disciplinary environments change.

Overall, the fact that IRT has not replaced CTT in general psychological research is not simply a matter of a more advanced method failing to become pop-

ular. The additional information provided by IRT can yield practical benefits only when it matches the research question; these benefits must also be weighed against the costs of method use, and ultimately occur within disciplinary ecologies that have particular training and research traditions. Therefore, theoretical gains can be stably transformed into practical advantages only when they meet the demands of the question, when the practical gains are sufficient to offset the practical costs, and when they can be absorbed by the disciplinary ecology. Thus, the question truly worth discussing is no longer "whether IRT should ultimately replace CTT," but whether researchers can, in light of specific scientific questions, choose measurement methods that provide sufficient measurement information without excessively increasing model complexity.

5 Discussion, Conclusion, and Prospects

5.1 From Method Replacement to Problem-Driven Measurement Modeling

The preceding sections show that understanding IRT and CTT as a competition in which "new methods replace old methods" easily obscures the actual demands that different research questions place on measurement information. Complex models are not inherently superior to simple models. A sum score that has been sufficiently justified by measurement theory and can support the target scientific inference may be more appropriate than latent trait estimates when the model is misspecified or poorly fitted; conversely, when research conclusions directly involve item differences, measurement error at different trait levels, cross-test comparisons, or individualized decisions, abandoning more refined measurement models merely because sum scores are convenient may likewise create a mismatch between method and research question.

Therefore, this article argues for shifting from the logic of "method replacement" to problem-driven measurement modeling. The core is not to require all psychological research to use IRT, but to determine, according to the final scientific inference, what kind of measurement information and what degree of model complexity are needed. Researchers should first clarify: What is the study attempting to measure? What inferences are intended based on the measurement results? On this basis, they should further judge what information will be lost if item responses are aggregated into a total score, and whether that information would affect the research conclusions. If the information abandoned is only weakly related to the final inference, a simple score may already be sufficient for the research need; if item differences, individual measurement error, or cross-group comparability are themselves part of the research question, then a measurement model capable of retaining this information is needed.

This problem-driven approach also means that measurement should not be treated merely as a scoring step before subsequent statistical analysis. The credibility of scientific inference depends not only on what statistical model is

applied after variables are obtained, but also on how observed behavior is transformed into measurement results with psychological meaning (Flake & Fried, 2020; Flake et al., 2022). Therefore, measurement reform cannot only add statistical and procedural norms; it also needs to pay further attention to the theoretical issues before data are formed, including construct definition and the measurement process (Paredes & Carré, 2024). If variables entering correlation, regression, or structural equation models themselves lack sufficient measurement basis, then increasing the complexity of subsequent statistical analyses cannot fully compensate for measurement problems at the front end.

In this sense, the more important implication of IRT for general psychological research may not be a particular model, but a measurement-modeling mindset: observed results such as item responses and task performance are not naturally existing psychological scores; researchers need to explain why these overt behaviors can reflect the target psychological construct, and according to what rules they are transformed into interpretable results.

the measurement results being interpreted. In other words, before using a variable for statistical analysis, one still needs to ask how that variable was formed and what kind of psychological interpretation it can support.

This line of thinking also helps us understand the relationship between psychological measurement and computational modeling in recent years. Computational models such as IRT and drift diffusion models do not have identical research aims: the former is mainly used for psychological measurement, whereas the latter is more often used to explain the cognitive processes behind behavior. But neither of them directly equates observed indicators such as accuracy, reaction time, or total item scores with theoretical objects. Instead, by means of formal models, they infer from externally manifested behavior the latent traits or psychological processes that cannot be directly observed. When the parameters of computational models are further used to study stable individual differences, whether these parameters themselves have sufficient reliability and validity also becomes a problem of psychological measurement. It can thus be seen that the concern of psychological measurement— "how observed behavior supports inference about latent attributes" —is not limited to traditional ability tests or self-report scales.

The development of artificial intelligence further highlights the distinction between prediction and measurement. Machine learning can improve predictive accuracy, but high predictive accuracy does not automatically explain what attribute a score represents, whether it can be compared across different tasks, or what theoretical interpretation it can support. Zhou et al. (2026) noted in research on artificial intelligence evaluation that aggregate performance on common benchmark tests is simultaneously affected by the capability of the artificial intelligence system and the specific task requirements; therefore, an average score cannot simply be interpreted as the system' s stable "ability." That study further distinguished between task requirements and system capabilities, and established a jointly interpretable scale. Its significance is not that

it proves IRT can directly solve the problem of artificial intelligence evaluation, but rather that, when researchers attempt to transform complex observed performance into interpretable and comparable attributes, they once again encounter the fundamental questions that psychological measurement has long faced: what exactly the observed performance reflects, and under what conditions different observed results can be placed on the same scale for comparison. Therefore, the development of artificial intelligence and computing technology has not made psychological measurement theory lose its significance. Larger data scales and more complex algorithms can reduce some analytic costs and create new research objects; but as long as researchers attempt to infer indirectly observable attributes from observable behavior, they still need to explain on what theory and model such inference is based. Of course, IRT is not the only solution to this problem; in the future, measurement models more suitable for new types of data and research questions may also emerge. What is truly of general significance is this principle: the formation of measurement results and the basis for their interpretation must be clearly justified.

From this perspective, the framework proposed in this article—"theoretical gain-problem demand-practical cost-disciplinary ecology" —is also dynamic. With the development of software and computing technology, the cost of using complex models may decrease; as psychology pays more attention to individual differences, latent processes, and complex behavioral data, the demand for more refined measurement information may increase; changes in methods training, software tools, and publication norms may also change the degree of fit between a method and the disciplinary ecology. Therefore, whether a method becomes mainstream at present does not mean that it will retain the same practical position in the future.

Ultimately, problem-driven measurement modeling can be summarized as a simple principle: one should not first ask whether CTT or IRT ought to be used, but should first ask what the study is trying to measure, what one hopes to infer from it, and what degree of measurement information is sufficient to support that inference. When simple scores can already adequately support the research purpose, there is no need to add models merely for the sake of methodological complexity; when scientific conclusions depend on item or person information that total scores cannot retain, further modeling should not be halted merely because traditional methods are more familiar.

5.2 Conclusion and Prospects

Starting from the question "why has IRT not replaced CTT," this article discussed a more general methodological phenomenon: why can a kind of ···

Methods that can provide more measurement information do not necessarily become more universal research practices. This paper argues that IRT and CTT should not simply be understood as a replacement relation between an "advanced method" and a "backward method" ; rather, they should be regarded

as psychometric frameworks that can provide different degrees of measurement information and are applicable to different scientific inferences. The key to method selection lies not in the era in which a theory emerged or in the complexity of the model itself, but in whether the information it provides can meet the needs of the specific research question.

Accordingly, this paper proposes a framework for method adoption based on "theoretical gain–problem demand–practical cost–disciplinary ecology." Theoretical gain answers what additional information a method can provide; problem demand determines whether this additional information can be transformed into actual benefits; practical cost affects whether researchers are willing to invest more learning, data, and model-analysis resources for these benefits; and disciplinary ecology further influences whether a method can enter existing training systems, research workflows, and publication norms. Thus, the core judgment of this paper can be summarized as follows: only when theoretical gains meet problem demands, actual benefits are sufficient to offset practical costs, and the method can be absorbed by the disciplinary ecology, can those gains be stably transformed into practical advantages. Therefore, this paper does not advocate replacing CTT with IRT; rather, it advocates problem-driven measurement modeling that reduces inertia-driven method selection. If total scores already sufficiently support the target inference, there is no need to adopt IRT solely because IRT is more complex. However, if total scores obscure item differences, individual measurement error, or information about cross-group comparability that is directly relevant to scientific conclusions, more refined measurement models should not be rejected merely because traditional methods are more familiar. What truly needs to be demonstrated is whether the chosen measurement method can receive joint support from theory, data, and scientific inference.

This paper still has certain limitations. First, current evidence concerning the application of IRT mainly comes from different subfields of psychology and is not yet sufficient to accurately estimate the overall use of IRT across psychological research as a whole. Future work may use systematic bibliometric studies to compare changes in the use of CTT, IRT, and other measurement models across different psychological domains and time periods, and further examine the relationships among research aims, data conditions, method training, journal norms, and method adoption. Second, it is still necessary to test more directly under what conditions the theoretical gains of IRT affect substantive research conclusions. For example, systematic comparisons can be made between CTT total scores and IRT latent-trait estimates in terms of group differences, variable relations, individual classification, and judgments of individual change, thereby distinguishing which methodological differences mainly remain at the model level and which truly influence scientific inference. Some of the future research directions in the present manuscript already include these two main lines.

Finally, the most important significance of the debate between IRT and CTT is not to select a "winner"among psychometric methods, but to prompt psychology to rethink a more fundamental question: why can observational data become

measurement results of psychological constructs? More complex statistical analyses can examine already formed variables more fully, and machine learning can predict certain outcomes more accurately, but these methods themselves cannot automatically explain what variables represent, whether different measurement results are comparable, or what psychological theory they can support. As long as psychology still needs to infer psychological constructs that cannot be directly observed from observable behavior, measurement modeling remains an indispensable component of scientific theory.

Therefore, what truly may need to change is not CTT itself, but the research habit whereby, once several observed indicators are aggregated into a single number, the measurement work is assumed to be complete. What is more worthy of pursuit in the future of psychometric methods is not a label-based upgrade from "classical" to "modern," but a shift from inertia-driven scoring and method selection toward measurement modeling that is driven by scientific questions, has clear theoretical foundations, and can support target inferences.

Why Has IRT Not Replaced CTT?

Zhan, Peida

(School of Psychology, Zhejiang Normal University, Jinhua City, 321004, E-mail: pdzhan@gmail.com)

Abstract

Compared to classical test theory (CTT), item response theory (IRT) models the relationship between latent traits and item responses at the item level, reduces the sample dependence of item parameter estimation, and evaluates measurement precision across different levels of the latent trait continuum. Nevertheless, IRT has not superseded routine CTT-based practices, which primarily rely on sum scores and traditional item analysis, in general psychological research. This discrepancy cannot be simply attributed to the technical complexity of IRT or insufficient methodological training. Instead, it requires distinguishing among three distinct levels: theoretical gain, practical benefit, and method adoption. The additional information provided by IRT translates into practical benefits only when it is directly relevant to the research questions and scientific inferences at hand. Furthermore, its routine adoption is constrained by practical costs (e.g., learning and application) and the path-dependent constraints of the broader disciplinary ecology, including established training systems and research workflows. Accordingly, this article proposes a framework of "theoretical gain–problem demand–practical cost–disciplinary ecology" to explain method adoption. It advocates shifting the focus from a logic of "method replacement" to problem-driven measurement modeling, wherein the choice of a measurement theory is determined by the required measurement resolution for specific scientific inferences and the associated cost-benefit trade-offs.

Keywords: Item response theory; Classical test theory; Psychometrics; Psychological measurement models

References

Guo Xiaojun, Jiao Yuyue, Bai Xiaoyun, Luo Zhaosheng, & Li Hong. (2024). Joint modeling of psychological experimental data: Mixed effects of responses and response times. *Acta Psychologica Sinica, 56*(11), 1619–1633. https://doi.org/10.3724/SP.J.1041.2024.01619

Liu Yikang, & Hu Chuanpeng. (2024). Behavioral and cognitive neural evidence for the evidence accumulation model. *Chinese Science Bulletin, 69*(8), 1068–1081.

Luo Zhaosheng. (2012). *Fundamentals of item response theory.* Beijing Normal University Press.

Tu Dongbo, Luo Fen, Wang Dali, & Cai Yan. (2023). flexCAT: A computerized adaptive testing development platform. *Educational Measurement and Evaluation, Bilingual Issue, 4*(1), 4. https://doi.org/10.59863/FUQC8311

Yang Xiangdong. (2010). Correspondence analysis between test item response mechanisms and assumptions of psychological measurement models. *Advances in Psychological Science, 18*(8), 1349–1358.

Yu Jiayuan. (1989). A comparative research report on classical measurement theory and item response theory. *Journal of Nanjing Normal University (Social Science Edition), 4*, 93–104.

Zhu Qianqian, Liu Zheng, Kang Chunhua, & Hu Chuanpeng. (2026). Measurement reliability of cognitive tasks: Progress and prospects. *Chinese Science Bulletin, 71*(11), 2472–2484.

Zhan Shida. (2022). Joint-crossed loading multimodal cognitive diagnosis modeling incorporating eye-movement fixation points. *Acta Psychologica Sinica, 54*(11), 1416–1438.

Bach, D. R. (2024). Psychometrics in experimental psychology: A case for calibration. *Psychonomic Bulletin & Review, 31*, 1461–1470. https://doi.org/10.3758/s13423-023-02421-z

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Chen, Y., Li, X., Liu, J., & Ying, Z. (2025). Item response theory—A statistical framework for educational and psychological measurement. *Statistical Science, 40*(2), 167–194. https://doi.org/10.1214/23-STS896

Clayson, P. E. (2024). The psychometric upgrade psychophysiology needs. *Psychophysiology, 61*(3), e14522. https://doi.org/10.1111/psyp.14522

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9*(4), 351–361. https://doi.org/10.1177/014662168500900402

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist, 77*(4), 576–588. https://doi.org/10.1037/amp0001006

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods, 20*(3), 465–486. https://doi.org/10.1177/1094428116689708

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology, 1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Fu, Y., Zhan, P., Chen, Q., & Jiao, H. (2024). Joint modeling of action sequences and action time in computer-based interactive tasks. *Behavior Research Methods, 56*, 4293–4310. https://doi.org/10.3758/s13428-023-02178-2

Hayden, E. P. (2022). A call for renewed attention to construct validity and measurement in psychopathology research. *Psychological Medicine, 52*(14), 2930–2936. https://doi.org/10.1017/S0033291722003221

McNeish, D. (2022). Limitations of the sum-and-alpha approach to measurement in behavioral research. *Policy Insights from the Behavioral and Brain Sciences, 9*(2), 196–203. https://doi.org/10.1177/23727322221117144

McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods, 55*(8), 4269–4290. https://doi.org/10.3758/s13428-022-02016-x

McNeish, D. (2024). Practical implications of sum scores being psychometrics' greatest accomplishment. *Psychometrika, 89*(4), 1148–1169. https://doi.org/10.1007/s11336-024-09988-z

Meier, S. T. (2023). Editorial: Persistence of measurement problems in psychological research. *Frontiers in Psychology, 14*, 1132185. https://doi.org/10.3389/fpsyg.2023.1132185

Mislevy, R. J. (2024). Are sum scores a great accomplishment of psychometrics or intuitive test theory? *Psychometrika, 89*(4), 1170–1174. https://doi.org/10.1007/s11336-024-10003-8

Myers, C. E., Interian, A., & Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology, 13*, 1039172. https://doi.org/10.3389/fpsyg.2022.1039172

Paredes, J., & Carré, D. (2024). Looking for a broader mindset in psychometrics: The case for more participatory measurement practices. *Frontiers in Psychology, 15*, 1389640. https://doi.org/10.3389/fpsyg.2024.1389640

Rappaport, B. I., Shankman, S. A., Glazer, J. E., Buchanan, S. N., Weinberg, A., & Letkiewicz, A. M. (2025). Psychometrics of drift-diffusion model parameters derived from the Eriksen flanker task: Reliability

and validity in two independent samples. *Cognitive, Affective, & Behavioral Neuroscience, 25*(2), 311–328. https://doi.org/10.3758/s13415-024-01222-8

Schurr, R., Reznik, D., Hillman, H., Bhui, R., ···Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. *Nature Human Behaviour, 8*, 917–931.

Schroeders, U., & Gnambs, T. (2025). Sample-size planning in item-response theory: A tutorial. *Advances in Methods and Practices in Psychological Science, 8*(1), Article 25152459251314798. https://doi.org/10.1177/25152459251314798

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*(6), 786–809. https://doi.org/10.1177/0959354312454353

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024a). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika, 89*(1), 84–117. https://doi.org/10.1007/s11336-024-09964-7

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024b). Rejoinder to McNeish and Mislevy: What does psychological measurement require? *Psychometrika, 89*(4), 1175–1185. https://doi.org/10.1007/s11336-024-10004-7

Sinharay, S., & Monroe, S. (2025). Assessment of fit of item response theory models: A critical review of the status quo and some future directions. *British Journal of Mathematical and Statistical Psychology, 78*(3), 711–733. https://doi.org/10.1111/bmsp.12378

ten Holt, J. C., van Duijn, M. A. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling, 52*(3), 272–297.

van der Linden, W. J. (2016). *Handbook of item response theory: Three volume set*. Chapman and Hall/CRC.

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods, 55*, 788–806. https://doi.org/10.3758/s13428-022-01849-w

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in psychometrics. *Perspectives on Psychological Science, 17*(3), 788–804. https://doi.org/10.1177/17456916211014183

Zein, R. A., & Akhtar, H. (2025). Getting started with the graded response model: An introduction and tutorial in R. *International Journal of Psychology, 60*(1), e13265. https://doi.org/10.1002/ijop.13265

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*, 262–286. https://doi.org/10.1111/bmsp.12114

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., Prunty, J. E., Li, Z., Sánchez-García, P., Jiang-Chen, K., Casares, P. A. M., Zu, J., Burden, J., Mehrbakhsh, B., Stillwell, D., Cebrian, M., ···Hernández-Orallo, J. (2026). General scales unlock AI evaluation with explanatory and predictive power. *Nature, 652*(8108), 58–67. https://doi.org/10.1038/s41586-026-10303-2

Zorowitz, S., Chierchia, G., Blakemore, S.-J., & Daw, N. D. (2024). An item response theory analysis of the matrix reasoning item bank (MaRs-IB). *Behavior Research Methods, 56*(3), 1104–1122. https://doi.org/10.3758/s13428-023-02067-8

Note: Figure translations are in progress. See original paper for figures.