Multitech Journal of Science and Technology (MJST)
Vol.3, No.8, 2026:897-916

The Evolution of Psychometrics In the Digital Era: From Classical Measurement Theory to Adaptive Assessment and Artificial Intelligence

Loso Judijanto
IPOSS Jakarta
Corresponding Author: Loso Judijanto losojudijantobumn@gmail.com

A R T I C L E I N F O

Keywords: Psychometrics; Psychological Measurement; Validity; Reliability; Item Response Theory; Computerized Adaptive Testing; Process Data; Artificial Intelligence; Algorithmic Fairness; Digital Ethics

Received : 2 June
Revised : 20 July
Accepted: 20 August

A B S T R A C T

Psychometrics is undergoing a fundamental change as psychological measurement increasingly shifts from static tests toward adaptive digital systems that are data-driven and supported by artificial intelligence. This article aims to explain how this transformation is changing the theoretical, methodological, and practical foundations of psychometrics, as well as to identify the scientific and ethical challenges that affect its credibility. The study uses a qualitative literature review with a narrative-integrative approach to peer-reviewed journal articles published since 2020. The synthesis shows that classical test theory remains important as a foundation, but item response theory, computerized adaptive testing, factor and network analysis, response time modeling, and machine learning expand the unit of analysis from final scores to response processes and multimodal data. Digitalization increases efficiency, personalization, and prediction accuracy, but it doesn't automatically improve validity. The quality of inferences still depends on clear constructs, proper reliability, measurement invariance, algorithmic fairness, transparency, privacy protection, and cross-cultural suitability. This article proposes a direction for psychometrics-by-design that integrates validity, fairness, data security, interpretability, and human oversight right from the instrument design stage. The implication is that the future of psychometrics isn't just about using fancier technology, but building a measurement ecosystem that's adaptive, auditable, human-centered, and backed by ongoing evidence of validity

INTRODUCTION

Psychometrics grew out of the need to turn constructs that can't be directly observed—like cognitive abilities, personality, attitudes, motivation, and psychological conditions—into inferences that can be reliably supported by data. So, the core issue in psychometrics isn't just about generating scores, but about building a coherent relationship between construct definitions, instrument design, statistical models, reliability evidence, validity evidence, and the decisions made based on scores. Recent literature shows that the quality of psychological research can suffer when measurement decisions aren't reported transparently or when instruments are used outside their validation context. Therefore, careful measurement is a requirement for accumulating scientific knowledge, not just a technical step before statistical analysis (Flake and Fried, 2020; Hussey and Hughes, 2020; Zickar, 2020).

This urgency has increased in the digital era. Computer- and online-based test administration, computerized adaptive testing (CAT), log files, response times, sensor data, and digital behavior traces have expanded the types of evidence available to psychometrics. Information units no longer have to stop at right-wrong or total scores; click sequences, response durations, answer revisions, navigation patterns, even multimodal data can be used to assess the processes underlying performance. These changes open up opportunities to make assessments more efficient and diagnostic, but at the same time, they demand validity theories that can explain how process data relate to the constructs being measured (Goldhammer et al., 2021; Ulitzsch et al., 2021; Levy, 2020).

At the same time, machine learning and artificial intelligence (AI) are speeding up the shift in psychometrics from models that are mainly explanatory to a mix of construct modeling and prediction. Algorithms can help with item selection, scale reduction, nonlinear pattern recognition, automatic scoring, and modeling of large-scale data. However, increasing predictive accuracy does not automatically mean construct validity. Highly predictive models can still lead to unfair decisions if the training data is not representative, the decision mechanisms are not transparent, or group inequalities are hidden by aggregate performance metrics (Gonzalez, 2021; Fokkema et al., 2022; Trognon et al., 2022; Pellert et al., 2024).

Digital transformation also broadens the social consequences of measurement. Online assessments can determine access to education, jobs, psychological services, and various other opportunities, making issues like fairness, differential item functioning, measurement invariance, privacy, informed consent, and data security part of psychometric quality itself. Studies on algorithm-based selection show that perceptions of fairness and transparency affect user acceptance, while research on algorithmic evaluation shows that high technical performance doesn't automatically eliminate discrimination. Therefore, digital psychometrics should be understood as a sociotechnical system that connects measurement theory, technology design, usage context, and decision governance (Acikgoz et al., 2020; Köchling et al., 2021; Hilliard et al., 2022; Johnson et al., 2022).

Even though the literature is growing rapidly, discussions are often fragmented: some studies focus on reliability and validity, others on IRT and CAT, data processes, machine learning, or algorithmic ethics. This article offers a synthesis that puts these developments into one flow: continuity from classical measurement theory, expansion toward latent and adaptive models, leveraging process data, and integrating AI that ultimately needs to be tested through evidence of validity, invariance, fairness, and accountability. With this framework, developments in psychometrics aren't read as a total replacement of old paradigms by new technology, but as an expansion of data sources and inference complexity that demands stronger evidence standards (Cortina et al., 2020; Sijtsma and Pfadt, 2021; Ramminger and Jacobs, 2024).

Based on this background, this article is guided by two research questions.
RQ1. How does digital transformation change the theoretical foundations, data sources, analysis methods, and psychometric practices from a score-based approach to adaptive assessment, process data, and artificial intelligence?
RQ2. What validity, reliability, invariance, fairness, privacy, transparency, and standardization challenges are most critical for digital psychometrics to produce accountable inferences and decisions?

The aim of this article is (1) to synthesize the conceptual and methodological developments in contemporary psychometrics, particularly the relationship between test theory, CAT, process data, machine learning, and AI; (2) to identify the benefits and limitations of digital transformation for measurement quality; and (3) to formulate a psychometrics development agenda that places validity, fairness, transparency, data security, and human oversight as design principles. With these goals, the article is expected to serve as a conceptual map for researchers and practitioners who need an integrative understanding of the direction of psychometrics in the digital era.

LITERATURE REVIEW

1 Psychometrics as the Science of Measurement Inference

Conceptually, psychometrics is a science that links latent constructs to observed indicators through measurement rules and inference models. Scores are not the constructs themselves, but rather representations that only make sense if there are theoretical arguments and empirical evidence supporting their interpretation. The contemporary perspective therefore views construct definitions, indicator quality, model accuracy, and usage context as part of a chain of reasoning. Criticism of questionable measurement practices shows that ignoring construct definitions, modifying instruments without justification, or incomplete reporting can weaken all kinds of subsequent inferences, even when the statistical analysis is sophisticated (Flake and Fried, 2020; Hussey and Hughes, 2020; Jebb, Ng and Tay, 2021; Ramminger and Jacobs, 2024).

The historical journey from measuring individual differences to modern psychometrics can be seen as a process of making the connection between theory and data more formal. The legacy of classical test theory (CTT) emphasizes observed scores, true scores, and measurement error; later developments expanded this through factor models, IRT, response time models, psychometric

networks, and computational models. Even though technologies and techniques change, the key questions stay the same: what is being measured, how consistently is it measured, and how well is the score interpretation supported by evidence. So, methodological innovations should be judged based on their contribution to the quality of inference, not the complexity of the algorithm (Zickar, 2020; Kim et al., 2020; Epskamp, 2020).

2. Reliability, Validity, and Measurement Invariance

Reliability describes the level of consistency or precision of scores under certain measurement conditions, but it's not a single trait that permanently belongs to a test. Recent literature emphasizes that the reliability coefficient should be chosen based on the data structure and the purpose of interpretation. Automatically using Cronbach's alpha can be misleading when the tau-equivalence assumption isn't met; omega is often more informative, but choosing it still requires understanding the measurement model. So, good reliability reporting should explain the sources of variation, the population, the administration conditions, and the model behind the estimate (Hayes and Coutts, 2020; Cortina et al., 2020; Sijtsma and Pfadt, 2021; Ellis, 2021).

In contemporary psychometrics, validity is better understood as the quality of interpreting and using scores rather than just a label once given to an instrument. Evidence of validity needs to connect the content, internal structure, relationships with other variables, response processes, and the consequences of use. In the context of cross-group comparisons, measurement invariance and differential item functioning (DIF) become crucial because score comparisons only make sense if the relevant parameters function equivalently. Regularization approaches and structural-change tests show the development of methods for detecting parameter instability that might be hidden in aggregate analyses (Belzak and Bauer, 2020; Huth et al., 2022; Maassen et al., 2025; Stefana et al., 2025).

3. From CTT to IRT and Contemporary Latent Models

CTT is still important because it provides the basic language about scores, errors, and reliability, and it's easy to use when developing instruments. Its limitations show up when item characteristics and ability estimates depend too much on the sample and the test format. IRT shifts the focus to the probability of item responses as a function of latent ability and item parameters, allowing item information, difficulty, discrimination, and—on certain models—guessing to be analyzed explicitly. The variety of IRT models used in the literature shows that model selection should follow the response format, dimensionality, and inference goals, not just analytical habits (Kim et al., 2020; Cai and Houts, 2021; Tutz, 2022; Bürkner, 2022).

Latent models have also developed through factor analysis and network approaches. Exploratory and confirmatory factor analysis remain central for checking dimensionality, but decisions about factor retention are a major source of uncertainty. Simulations and machine learning have been used to improve factor retention decisions under certain conditions, while exploratory graph analysis offers an alternative way to identify dimensional structure. On the other hand, network models focus on relationships between components as the object of analysis and can be useful when the assumption of a single common latent

cause is insufficient. All of these approaches emphasize that the internal structure of instruments needs to be tested, compared, and theoretically justified (Goretzko and Bühner, 2020; Golino et al., 2020; Goretzko and Bühner, 2022; Epskamp, 2020).

4. Computerized Adaptive Testing and Adaptive Measurement

CAT combines a calibrated item bank with an item selection algorithm to present the most informative items for estimating ability at the moment. The theoretical advantage is efficiency: two participants don't have to get the same set of items to reach comparable precision. However, CAT requires proper item calibration, clear stopping rules, item exposure control, content management, and assessment of calibration errors. Recent literature also extends CAT to multidimensional and cognitive-diagnostic settings, so the adaptation not only estimates ability levels but can also identify more specific mastery patterns (van der Linden and Jiang, 2020; Lin et al., 2021; Bao et al., 2021; Sorrel et al., 2021; Huang and Cai, 2021).

Adaptive implementation should also be evaluated from the perspective of user experience and clinical or educational decisions. Shorter measurements can reduce respondent burden, but efficiency shouldn't come at the cost of construct coverage or comparability. In mental health services, for example, CAT is seen as potentially supporting measurement-based care because it can update estimates based on individual responses, but integrating it into practice still requires professional interpretation and infrastructure that ensures data quality. The concept of adaptation level also needs to be reported explicitly so that performance across algorithms can be compared fairly (Wyse and McBride, 2021; Carlo et al., 2021).

5. Process Data, Response Time, and Digital Behavior Footprints

Digitization produces process data that was previously hard to observe on paper tests. Log data records events, event attributes, the sequence of actions, and timestamps; this information can help explain strategies, disengagement, revisions, or problem-solving paths. However, process indicators don't automatically have psychological meaning. Validation has to start with a theory about the expected cognitive or behavioral processes, then move on to extracting indicators, modeling, and checking how those indicators relate to the target constructs. Without this step, high-resolution data can produce interesting correlations but poor interpretation (Provasnik, 2021; Goldhammer et al., 2021; Ulitzsch et al., 2021).

Response time has become one of the most mature forms of data processing in psychometrics. Combined accuracy-time models, censoring for not-reached items, semiparametric factor models, and diffusion-IRT approaches show that speed and accuracy can hold different information about the response process. Response time can also help identify heterogeneous strategies, but its interpretation needs to distinguish between cognitive speed, motivation, response strategy, technology access, and administration conditions. Therefore, response time should be treated as additional evidence that enriches the validity argument, not as a substitute for performance scores (Lu and Wang, 2020; Hong

et al., 2021; Henninger and Plieninger, 2021; Kang et al., 2022; Guo et al., 2022; Liu and Wang, 2022).

6. Machine Learning, AI, and Computational Psychometrics

Machine learning gives psychometrics a set of methods for prediction, pattern recognition, item reduction, and exploring nonlinear relationships. Using algorithms to shorten scales can reduce administrative burden, while supervised learning models can provide extra evidence about criterion-related or construct-related patterns. However, machine learning optimizes predictive functions that aren't necessarily identical to psychological constructs. That's why evaluation should compare out-of-sample performance with transparency, feature stability, cross-group generalization, and theoretical coherence (Gonzalez, 2021; Camilleri et al., 2021; Trognon et al., 2022; Fokkema et al., 2022).

The development of generative AI expands that question. Studies on psychometric inventories in large language models show that psychological instruments can be used to test AI system response patterns, but the scores obtained can't just be equated with human psychological traits because the mechanism behind generating responses is different. More broadly, AI can help with item generation, scoring, adaptive feedback, and anomaly detection, but every use requires bias audits, model documentation, human oversight, and reevaluation when models or data change. So, computational psychometrics needs a closer link between measurement theory and data science (Pellert et al., 2024; Fokkema et al., 2022).

7. Fairness, Privacy, Transparency, and Ethics

Fairness in digital psychometrics has at least two layers: measurement fairness and decision fairness. The first layer asks whether items, parameters, or scores work equivalently across different groups; the second layer asks whether the algorithms using those scores produce errors and consequences that are justifiable. Research on automated recruitment shows that even accurate models can still be discriminatory, and applicants' perceptions are influenced by the type of information given about the automated process. The differential algorithmic functioning framework extends the logic of DIF to evaluate how algorithms behave across different groups (Acikgoz et al., 2020; Langer et al., 2021; Köchling et al., 2021; Johnson et al., 2022; Suk and Han, 2024).

Privacy and informed consent get more complicated when the data being analyzed isn't just explicit answers, but also metadata, logs, sensors, or behavior patterns. In the context of e-assessment and digital phenotyping, users might find it hard to understand the full flow of data and the consequences of how it's processed. That's why principles like data minimization, purpose limitation, secure storage, transparency, the right to understand decisions, and objection mechanisms need to be considered as part of assessment design. The literature on psychological AI ethics and digital health fairness emphasizes that human-sensitive technology requires governance that goes beyond a one-time consent at the start (Muravyeva et al., 2020; Fleming, 2021; Coghlan and D'Alfonso, 2021; Buda et al., 2022; Fulmer et al., 2021).

METHODOLOGY

This article uses a qualitative literature review with a narrative-integrative orientation, rather than a systematic literature review. This approach was chosen because the goal of the research is to build a conceptual and theoretical synthesis of digital psychometrics, connecting scattered literature on measurement theory, educational measurement, psychological assessment, computational methods, and AI ethics. The literature is treated as material to compare concepts, identify continuity and paradigm shifts, and construct an integrative argument; no meta-analysis, effect size calculation, or claims that all relevant publications have been covered are made. The flexible yet transparent nature of a narrative/integrative review is suitable for broad and interdisciplinary questions (Lubbe, ten Ham-Baloyi and Smit, 2020; Barry, Merkebu and Varpio, 2022; Sukhera, 2022).

The search was carried out purposively and iteratively, focusing on peer-reviewed journal articles published since 2020 that discuss psychometrics, measurement validity, reliability, measurement invariance, item response theory, computerized adaptive testing, process data, response time, machine learning, artificial intelligence, algorithmic fairness, online assessment, privacy, and digital phenotyping. Articles were retained if they were directly relevant to the research questions, had verifiable publication details, and had a DOI as well as consistent publisher/index pages. Sources that could not be verified or whose journal identity was doubtful were not used. Historical literature before 2020 was discussed only through the synthesis of contemporary articles so that the reference list meets the temporal limits set in this article.

The analysis was done through comparative reading and thematic coding across seven domains: (1) measurement foundations and validity; (2) reliability and invariance; (3) CTT, IRT, and latent models; (4) CAT and online assessments; (5) data processing and response time; (6) machine learning/AI; and (7) fairness, privacy, transparency, and governance. Findings across domains were then synthesized to answer RQ1 and RQ2 and to identify research agendas. Interpretation credibility was maintained through consistency between research questions, analysis categories, cited evidence, and limiting claims according to the nature of a qualitative literature review (Yadav, 2022; Barry, Merkebu and Varpio, 2022).

RESULTS AND DISCUSSION

1. Continuity and Shifts in Measurement Theory

The synthesis shows that digital psychometrics is more of a cumulative transformation than a revolution that breaks away from previous theories. CTT still provides basic concepts of error and score consistency, while IRT offers a more detailed way to link responses with item parameters and latent abilities. Expansions into multidimensional models, cognitive diagnosis, networks, and process modeling don't eliminate the need for reliability and validity; in fact, having more data sources increases the number of analytical decisions that can affect inferences. So, the development of methods should be seen as adding layers of evidence that need to be integrated, not as automatically legitimizing the use of more complex models (Cortina et al., 2020; Kim et al., 2020; Tutz, 2022; Zhang and Chen, 2022).

The most substantial change is the shift from scores as a single output to measurement models as dynamic systems. Total scores are still useful in many contexts, but digital data allows item parameters, response trajectories, timing, and administration context to be analyzed together. Models like diffusion-IRT and joint response-time modeling show how variations that were previously seen as noise can actually provide information about the process. The implication is that the definition of error also becomes more contextual: some variations might reflect relevant psychological processes, while others come from devices, connectivity, strategies, or administration conditions (Boehm et al., 2021; DeCarlo, 2021; Kang et al., 2022; Man, Harring and Zhan, 2022; Zhang et al., 2023).

2. Validity and Reliability as an Ongoing Process

Study results show that validity in the digital era needs to be treated as an ongoing process. When algorithms, item banks, interfaces, or user populations change, evidence that was sufficient for previous versions of an instrument might not be enough anymore. Model changes can shift score distributions, missingness patterns, or item functions; context changes can introduce new sources of variance. Because of this, the concept of validity-by-design should include documentation of construct definitions, the relationship between digital features and constructs, drift evaluation, regular invariance checks, and version change reporting. This principle aligns with the demands for measurement transparency and criticism of using instruments without context-specific evidence (Flake and Fried, 2020; Belzak and Bauer, 2020; Ramminger and Jacobs, 2024; Maassen et al., 2025).

Reliability also needs to shift from the habit of reporting a single coefficient toward evaluating precision that fits the design. In CAT, the standard error is individual and changes with responses; in multimodal data, sources of error can come from sensors, feature extraction algorithms, and prediction models. Therefore, digital psychometric reports need to explain what is being made reliable—scores, rankings, classifications, latent parameters, or decisions—as well as the conditions under which those estimates apply. The alpha versus omega debate matters not because one index is always right, but because it shows that reliability depends on the model and the intended inference (Hayes and Coutts, 2020; Sijtsma and Pfadt, 2021; Ellis, 2021).

3. CAT, Online Assessment, and Measurement Efficiency

CAT is a clear example of how psychometric and computational theories can produce practical benefits. Adaptive item selection can focus information on areas of the respondent's ability and reduce the number of uninformative items. However, studies show that the quality of CAT heavily depends on the item bank, calibration accuracy, selection constraint algorithms, and termination rules. Aggressive adaptation can boost efficiency but also increase item exposure, reduce content coverage, or heighten sensitivity to incorrect parameters. Therefore, efficiency should be treated as a secondary goal after the validity and security of the item bank (van der Linden and Jiang, 2020; Lin et al., 2021; Bao et al., 2021; Sorrel et al., 2021).

Online assessments expand access, but they also introduce varying testing environments. Devices, screen sizes, bandwidth, distractions, technical support, and supervision levels can all affect how people respond. Security issues are also becoming more important in high-stakes testing because remote administration can lead to risks like impersonation, unauthorized help, and item leaks. Proctoring policies need to balance test integrity with privacy and accessibility, especially since surveillance technology can disproportionately affect certain groups. Therefore, the equivalence between administration modes should be tested empirically, not just assumed (Langenfeld, 2020; Muravyeva et al., 2020; Carlo et al., 2021).

4. Process Data: From Answer Products to Process Evidence

Process data expands validation because it allows researchers to check whether responses are produced through processes that align with construct interpretation. Log files can show the sequence of actions, information exploration, revisions, and navigation patterns, while response times can help detect rapid responding or different strategies. The added value isn't in the amount of data, but in the ability to link process indicators with task theory. When indicators are created after looking at the data without a conceptual basis, the risk of interpretive overfitting increases; on the other hand, evidence-centered design can turn process data from a byproduct into a component designed from the start (Goldhammer et al., 2021; Ulitzsch et al., 2021; Levy, 2020).

Modeling response time also highlights the importance of missing and not-reached responses. Items left unanswered near the end of a test can reflect time limitations, disengagement, or different strategies, so treating all of them as wrong answers could mix up ability with speededness. Censoring models and robust response-time modeling offer more suitable alternatives. These findings are relevant for digital tests since timestamps are automatically available and can be used to separate different response mechanisms, as long as interpretations still take device and context factors into account (Lu and Wang, 2020; Hong, Rebouças and Cheng, 2021; Lozano and Revuelta, 2021; Röhner and Holden, 2022; Guo et al., 2022; Liu and Wang, 2022).

5. AI and Machine Learning: Prediction Opportunities, Validity Risks

Machine learning expands psychometrics' ability to handle nonlinear relationships and high-dimensional data. Studies show potential for scale length reduction, classification, and mapping of psychometric patterns that are hard to capture with simple linear methods. However, optimizing predictive accuracy can select features that actually act as proxies for social status, demographics, devices, or contexts that shouldn't be part of the construct. Models can also be stable in one sample but fail in another population. So, cross-validation should be complemented with external validation, subgroup analysis, and checks on the construct relevance of the features used (Gonzalez, 2021; Camilleri, Eickhoff and Weis, 2021; Trognon et al., 2022; Fokkema et al., 2022).

In generative AI, the risk of misinterpretation is higher because the model can produce human-like responses without having the same psychological mechanisms. Applying inventories to LLMs is useful for comparing output patterns, but it doesn't prove that the system has traits in the psychological sense. For using AI in human assessment, the human-in-the-loop principle is needed in the design, validation, and use of high-stakes decisions. AI can suggest items, flag unusual responses, or provide estimates, but the responsibility for interpretation and the consequences of decisions shouldn't be fully shifted to a system that can't be audited (Pellert et al., 2024; Fulmer et al., 2021).

6. Fairness, Invariance, and Algorithmic Justice

RQ2 directs attention to fairness as a bridge between psychometrics and decision ethics. Measurement invariance and DIF test whether the relationship between constructs and indicators is equal across groups; algorithmic fairness adds the question of whether decision functions treat groups fairly after scores are generated. The two layers are not interchangeable. Instruments that are invariant can still be used in unfair policies, while algorithms with fairness constraints can still be built on biased scores. Therefore, digital psychometric audits should check fairness end-to-end, from items, scoring, predictive models, decision thresholds, to outcomes (Belzak and Bauer, 2020; Johnson, Liu and McCaffrey, 2022; Suk and Han, 2024).

Studies on automated recruitment show that fairness also has a procedural side. Applicants look at not just the outcome, but also whether they understand the assessment process and have a fair chance to show their skills. Information about using algorithms can affect how users react, while automated video evaluations can keep biases even if the overall accuracy is high. This suggests that transparency, clear explanations, accessibility, and ways to appeal should be treated as part of assessment design, especially for high-stakes decisions (Acikgoz et al., 2020; Langer et al., 2021; Köchling et al., 2021; Hilliard et al., 2022; Leutner et al., 2023).

7. Privacy, Informed Consent, and Data Governance

The richer the digital data, the higher the risk that assessments collect more information than needed. Metadata, clicks, biometrics, and behavior traces can lead to secondary inferences that participants never realized when agreeing to the test. Literature reviews show that informed consent in e-assessments isn't enough if it's just a formal agreement; participants need to understand the types of data, how it will be used, storage, possible linkages, and consequences of

decisions. So, data minimization and purpose limitation need to be applied from the design stage, not after the database is collected (Muravyeva et al., 2020; Fleming, 2021; Coghlan and D'Alfonso, 2021).

Digital phenotyping clarifies this issue because everyday behavior data can be used to estimate psychological conditions. Fairness studies on phone-based and multimodal mental-health algorithms show the need to evaluate performance across groups and the potential for inequality due to different device usage patterns. Ethical challenges don't mean this method should be rejected; rather, they highlight the need for governance that includes audits, documentation of data provenance, security, access control, and procedures for stopping use when evidence of validity or fairness decreases (Davidson, 2022; Park et al., 2022; Buda et al., 2022; Jiang et al., 2024).

8. Synthesis to Answer the Research Questions

The answer to RQ1 is that digitalization is changing psychometrics on four levels at once. First, the data sources expand from item responses to process data and multimodal data. Second, models evolve from aggregate scores to latent-variable, adaptive, joint, and machine-learning models. Third, administration shifts from fixed formats to online and adaptive systems that can update estimates in real-time. Fourth, quality evaluation moves from one-time validation to continuous monitoring since models, item banks, devices, and populations can change. This transformation broadens diagnostic and personalization capabilities, but it still relies on a solid construct foundation and clear measurement evidence (Kim et al., 2020; Goldhammer et al., 2021; Wyse and McBride, 2021; Zhang et al., 2023).

The answer to RQ2 is that the biggest challenge isn't a lack of algorithms, but rather the gap between the speed of innovation and the strictness of inference evaluation. Validity, reliability, invariance, fairness, privacy, transparency, security, and standardization need to be treated as one system. A reliable model that isn't invariant can lead to biased comparisons; an accurate but opaque model can be hard to audit; a system that's valid in one population can fail after a distribution shift. So, the measures of success in digital psychometrics should include not just accuracy and efficiency, but also robustness, generalizability, explainability, fairness, and the ability to be accountable for the consequences of decisions (Flake and Fried, 2020; Maassen et al., 2025; Fokkema et al., 2022; Johnson, Liu and McCaffrey, 2022).

9. Psychometric Development Directions: Psychometrics-by-Design

Based on the synthesis, this article proposes psychometrics-by-design as a development orientation. The idea is to build measurement quality into the system from the start: constructs are defined before features are collected; data collected is limited to what's needed for inference; algorithms are evaluated on subgroups; uncertainty is reported; version changes are documented; and high-stakes decisions include human oversight and a review path. This approach combines validity-by-design, fairness-by-design, and privacy-by-design so that ethical aspects aren't just an afterthought at the end of system development (Fleming, 2021; Buda et al., 2022; Suk and Han, 2024).

This agenda also requires a culture of open measurement. Researchers need to provide construct descriptions, items or item specifications when possible, scoring procedures, evidence of reliability, the models being tested, as well as any changes to instruments transparently. Cross-context replication, drift audits, and out-of-sample testing should become routine practices in AI-based systems. In a fast-changing digital environment, psychometric credibility will be more determined by the community's ability to check and update the evidence than by claims that a single instrument is universally valid (Flake and Fried, 2020; Hussey and Hughes, 2020; Stefana et al., 2025).

CONCLUSIONS AND RECOMMENDATIONS

This article shows that the evolution of psychometrics in the digital era is a gradual expansion from the foundations of classical measurement toward a more adaptive, data-rich, and computational assessment ecosystem. CTT still provides an important framework regarding error and consistency, while IRT, factor and network models, CAT, process data, response-time modeling, and machine learning expand the ability to model individual differences and response processes. These changes shift the focus from just producing scores to building inference systems that can adapt to the characteristics of items, individuals, and context.

The main findings for RQ1 confirm that digitalization is changing data sources, methods, administration, and the psychometric validation cycle. Data that was previously missing—like response times, action sequences, and interaction traces—can now be used to understand how answers are generated. CAT can boost efficiency, while AI can help with item reduction, pattern recognition, and prediction. However, these benefits are only scientific if digital indicators are meaningfully related to constructs, models are tested out-of-sample, and system changes are treated as a reason to update measurement evidence.

For RQ2, the main challenge of digital psychometrics is keeping the connection between innovation and accountability. Reliability isn't enough without validity; validity isn't enough without invariance and fairness; predictive accuracy isn't enough without transparency, privacy, security, and interpretability. Since digital systems can change through updates to algorithms, item banks, devices, or populations, measurement quality needs to be monitored over time. An end-to-end approach is needed so that biases can be detected from the items and raw data all the way to decision thresholds and the consequences of use.

The practical takeaway is the need for psychometrics-by-design: constructs, validity evidence, fairness, data minimization, security, and human oversight should be built in from the start, not tacked on after implementation. Researchers and practitioners need to document instrument and model versions, check for drift, evaluate subgroup performance, and provide review mechanisms for high-stakes decisions. Following these principles, digitalization can boost access, efficiency, and accuracy without compromising user rights or replacing professional judgment with irresponsible automation.

Future research needs to test the longitudinal validity of adaptive systems and AI after deployment, develop fairness methods that integrate measurement invariance with algorithmic decision fairness, and build cross-language and cross-cultural benchmarks for multimodal data. Studies also need to compare the benefits of prediction with privacy costs and interpretation complexity, test human-AI decision systems, and strengthen reporting standards for continuously updated data processes and models. This agenda will determine whether digital psychometrics just evolves into a faster measurement technology or becomes a measurement science that is more transparent, fair, and meaningful for humans.

FURTHER STUDY

This study still has limitations, so further research on the topic 'The Evolution of Psychometrics in the Digital Era: From Classical Measurement Theory to Adaptive Assessment and Artificial Intelligence' is needed to improve the study and provide more insights for both the writer and the readers.

REFERENCES

Acikgoz, Y., Davison, K.H., Compagnone, M. and Laske, M. (2020) 'Justice perceptions of artificial intelligence in selection', International Journal of Selection and Assessment, 28, pp. 399–416. doi: 10.1111/ijsa.12306. Available at: https://doi.org/10.1111/ijsa.12306

Bao, Y., Shen, Y., Wang, S. and Bradshaw, L. (2021) 'Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously', Applied Psychological Measurement, 45(1), pp. 3–21. doi: 10.1177/0146621620965730. Available at: https://doi.org/10.1177/0146621620965730

Barry, E.S., Merkebu, J. and Varpio, L. (2022) 'State-of-the-art literature review methodology: A six-step approach for knowledge synthesis', Perspectives on Medical Education, 11(5), pp. 281–288. doi: 10.1007/s40037-022-00725-9. Available at: https://doi.org/10.1007/s40037-022-00725-9

Belzak, W.C.M. and Bauer, D.J. (2020) 'Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning', Psychological Methods, 25(6), pp. 673–690. doi: 10.1037/met0000253. Available at: https://doi.org/10.1037/met0000253

Boehm, U., Marsman, M., van der Maas, H.L.J. and Maris, G. (2021) 'An attention-based diffusion model for psychometric analyses', Psychometrika, 86, pp. 938–972. doi: 10.1007/s11336-021-09783-0. Available at: https://doi.org/10.1007/s11336-021-09783-0

Buda, T.S., Guerreiro, J., Omana Iglesias, J., Castillo, C., Smith, O. and Matic, A. (2022) 'Foundations for fairness in digital health apps', Frontiers in Digital Health, 4, 943514. doi: 10.3389/fdgth.2022.943514. Available at: https://doi.org/10.3389/fdgth.2022.943514

Bürkner, P.-C. (2022) 'On the information obtainable from comparative judgments', Psychometrika, 87, pp. 1439–1472. doi: 10.1007/s11336-022-09843-z. Available at: https://doi.org/10.1007/s11336-022-09843-z

Cai, L. and Houts, C.R. (2021) 'Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory', Psychometrika, 86, pp. 754–777. doi: 10.1007/s11336-021-09777-y. Available at: https://doi.org/10.1007/s11336-021-09777-y

Camilleri, J.A., Eickhoff, S.B. and Weis, S. (2021) 'A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan Executive Function System', Scientific Reports, 11, 16896. doi: 10.1038/s41598-021-96342-3. Available at: https://doi.org/10.1038/s41598-021-96342-3

Carlo, A.D., Barnett, B.S. and Cella, D. (2021) 'Computerized adaptive testing (CAT) and the future of measurement-based mental health care', Administration and Policy in Mental Health and Mental Health Services Research, 48(5), pp. 729–731. doi: 10.1007/s10488-021-01123-9. Available at: https://doi.org/10.1007/s10488-021-01123-9

Coghlan, S. and D'Alfonso, S. (2021) 'Digital phenotyping: An epistemic and methodological analysis', Philosophy & Technology, 34(4), pp. 1905–1928. doi: 10.1007/s13347-021-00492-1. Available at: https://doi.org/10.1007/s13347-021-00492-1

Cortina, J.M., Sheng, Z., Keener, S.K., Keeler, K.R., Grubb, L.K., Schmitt, N., Tonidandel, S., Summerville, K.M., Heggestad, E.D. and Banks, G.C. (2020) 'From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology', Journal of Applied Psychology, 105(12), pp. 1351–1381. doi: 10.1037/apl0000815. Available at: https://doi.org/10.1037/apl0000815

Davidson, B.I. (2022) 'The crossroads of digital phenotyping', General Hospital Psychiatry, 74, pp. 126–132. doi: 10.1016/j.genhosppsych.2020.11.009. Available at: https://doi.org/10.1016/j.genhosppsych.2020.11.009

DeCarlo, L.T. (2021) 'On joining a signal detection choice model with response time models', Journal of Educational Measurement, 58(4), pp. 438–464. doi: 10.1111/jedm.12300. Available at: https://doi.org/10.1111/jedm.12300

Ellis, J.L. (2021) 'A test can have multiple reliabilities', Psychometrika, 86, pp. 869–876. doi: 10.1007/s11336-021-09800-2. Available at: https://doi.org/10.1007/s11336-021-09800-2

Epskamp, S. (2020) 'Psychometric network models from time-series and panel data', Psychometrika, 85, pp. 206–231. doi: 10.1007/s11336-020-09697-3. Available at: https://doi.org/10.1007/s11336-020-09697-3

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

Fleming, M.N. (2021) 'Considerations for the ethical implementation of psychological assessment through social media via machine learning', Ethics & Behavior, 31(3), pp. 181–192. doi: 10.1080/10508422.2020.1817026. Available at: https://doi.org/10.1080/10508422.2020.1817026

Fokkema, M., Iliescu, D., Greiff, S. and Ziegler, M. (2022) 'Machine learning and prediction in psychological assessment: Some promises and pitfalls', European Journal of Psychological Assessment, 38(3), pp. 165–175. doi: 10.1027/1015-5759/a000714. Available at: https://doi.org/10.1027/1015-5759/a000714

Fulmer, R., Davis, T., Costello, C. and Joerin, A. (2021) 'The ethics of psychological artificial intelligence: Clinical considerations', Counseling and Values, 66, pp. 131–144. doi: 10.1002/cvj.12153. Available at: https://doi.org/10.1002/cvj.12153

Goldhammer, F., Hahnel, C., Kroehne, U. and Zehner, F. (2021) 'From byproduct to design factor: On validating the interpretation of process indicators based on log data', Large-scale Assessments in Education, 9, 20. doi: 10.1186/s40536-021-00113-5. Available at: https://doi.org/10.1186/s40536-021-00113-5

Golino, H., Shi, D., Christensen, A.P., Garrido, L.E., Nieto, M.D., Sadana, R., Thiyagarajan, J.A. and Martinez-Molina, A. (2020) 'Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial', Psychological Methods, 25(3), pp. 292–320. doi: 10.1037/met0000255. Available at: https://doi.org/10.1037/met0000255

Gonzalez, O. (2021) 'Psychometric and machine learning approaches to reduce the length of scales', Multivariate Behavioral Research, 56(6), pp. 903–919. doi: 10.1080/00273171.2020.1781585. Available at: https://doi.org/10.1080/00273171.2020.1781585

Goretzko, D. and Bühner, M. (2020) 'One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis', Psychological Methods, 25(6), pp. 776–786. doi: 10.1037/met0000262. Available at: https://doi.org/10.1037/met0000262

Goretzko, D. and Bühner, M. (2022) 'Factor retention using machine learning with ordinal data', Applied Psychological Measurement, 46, pp. 406–421. doi: 10.1177/01466216221089345. Available at: https://doi.org/10.1177/01466216221089345

Guo, J., Xu, X., Ying, Z. and Zhang, S. (2022) 'Modeling not-reached items in timed tests: A response time censoring approach', Psychometrika, 87(3), pp. 835–867. doi: 10.1007/s11336-021-09810-0. Available at: https://doi.org/10.1007/s11336-021-09810-0

Hayes, A.F. and Coutts, J.J. (2020) 'Use omega rather than Cronbach's alpha for estimating reliability. But…', Communication Methods and Measures, 14(1), pp. 1–24. doi: 10.1080/19312458.2020.1718629. Available at: https://doi.org/10.1080/19312458.2020.1718629

Henninger, M. and Plieninger, H. (2021) 'Different styles, different times: How response times can inform our knowledge about the response process in rating scale measurement', Assessment, 28(5), pp. 1301–1319. doi: 10.1177/1073191119900003. Available at: https://doi.org/10.1177/1073191119900003

Hilliard, A., Guenole, N. and Leutner, F. (2022) 'Robots are judging me: Perceived fairness of algorithmic recruitment tools', Frontiers in Psychology, 13, 940456. doi: 10.3389/fpsyg.2022.940456. Available at: https://doi.org/10.3389/fpsyg.2022.940456

Hong, M., Rebouças, D.A. and Cheng, Y. (2021) 'Robust estimation for response time modeling', Journal of Educational Measurement, 58, pp. 262–280. doi: 10.1111/jedm.12286. Available at: https://doi.org/10.1111/jedm.12286

Huang, S. and Cai, L. (2021) 'Lord–Wingersky algorithm version 2.5 with applications', Psychometrika, 86, pp. 973–993. doi: 10.1007/s11336-021-09785-y. Available at: https://doi.org/10.1007/s11336-021-09785-y

Hussey, I. and Hughes, S. (2020) 'Hidden invalidity among 15 commonly used measures in social and personality psychology', Advances in Methods and Practices in Psychological Science, 3(2), pp. 166–184. doi: 10.1177/2515245919882903. Available at: https://doi.org/10.1177/2515245919882903

Huth, K.B.S., Waldorp, L.J., Luigjes, J., Goudriaan, A.E., van Holst, R.J. and Marsman, M. (2022) 'A note on the structural change test in highly parameterized psychometric models', Psychometrika, 87(3), pp. 1064–1080. doi: 10.1007/s11336-021-09834-6. Available at: https://doi.org/10.1007/s11336-021-09834-6

Jebb, A.T., Ng, V. and Tay, L. (2021) 'A review of key Likert scale development advances: 1995–2019', Frontiers in Psychology, 12, 637547. doi: 10.3389/fpsyg.2021.637547. Available at: https://doi.org/10.3389/fpsyg.2021.637547

Jiang, Z., Seyedi, S., Griner, E., Abbasi, A., Rad, A.B., Kwon, H., Cotes, R.O. and Clifford, G.D. (2024) 'Evaluating and mitigating unfairness in multimodal remote mental health assessments', PLOS Digital Health, 3(7), e0000413. doi: 10.1371/journal.pdig.0000413. Available at: https://doi.org/10.1371/journal.pdig.0000413

Johnson, M.S., Liu, X. and McCaffrey, D.F. (2022) 'Psychometric methods to evaluate measurement and algorithmic bias in automated scoring', Journal of Educational Measurement, 59, pp. 338–361. doi: 10.1111/jedm.12335. Available at: https://doi.org/10.1111/jedm.12335

Kang, I., De Boeck, P. and Ratcliff, R. (2022) 'Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model', Psychometrika, 87(2), pp. 725–748. doi: 10.1007/s11336-021-09819-5. Available at: https://doi.org/10.1007/s11336-021-09819-5

Kim, S.-H., Kwak, M., Bian, M., Feldberg, Z., Henry, T., Lee, J., Ölmez, İ.B., Shen, Y., Tan, Y., Tanaka, V., Wang, J., Xu, J. and Cohen, A.S. (2020) 'Item response models in Psychometrika and psychometric textbooks', Frontiers in Education, 5, 63. doi: 10.3389/feduc.2020.00063. Available at: https://doi.org/10.3389/feduc.2020.00063

Köchling, A., Riazy, S., Wehner, M.C. and Simbeck, K. (2021) 'Highly accurate, but still discriminatory: A fairness evaluation of algorithmic video analysis in the recruitment context', Business & Information Systems Engineering, 63, pp. 39–54. doi: 10.1007/s12599-020-00673-w. Available at: https://doi.org/10.1007/s12599-020-00673-w

Langenfeld, T. (2020) 'Internet-based proctored assessment: Security and fairness issues', Educational Measurement: Issues and Practice, 39(3), pp. 24–27. doi: 10.1111/emip.12359. Available at: https://doi.org/10.1111/emip.12359

Langer, M., Baum, K., König, C.J., Hähne, V., Oster, D. and Speith, T. (2021) 'Spare me the details: How the type of information about automated interviews influences applicant reactions', International Journal of Selection and Assessment, 29, pp. 154–169. doi: 10.1111/ijsa.12325. Available at: https://doi.org/10.1111/ijsa.12325

Leutner, F., Codreanu, S.-C., Brink, S. and Bitsakis, T. (2023) 'Game based assessments of cognitive ability in recruitment: Validity, fairness and test-taking experience', Frontiers in Psychology, 13, 942662. doi: 10.3389/fpsyg.2022.942662. Available at: https://doi.org/10.3389/fpsyg.2022.942662

Levy, R. (2020) 'Implications of considering response process data for greater and lesser psychometrics', Educational Assessment, 25(3), pp. 218–235. doi: 10.1080/10627197.2020.1804352. Available at: https://doi.org/10.1080/10627197.2020.1804352

Lin, Z., Chen, P. and Xin, T. (2021) 'The block item pocket method for reviewable multidimensional computerized adaptive testing', Applied Psychological Measurement, 45(1), pp. 22–36. doi: 10.1177/0146621620947177. Available at: https://doi.org/10.1177/0146621620947177

Liu, Y. and Wang, W. (2022) 'Semiparametric factor analysis for item-level response time data', Psychometrika, 87(2), pp. 666–692. doi: 10.1007/s11336-021-09832-8. Available at: https://doi.org/10.1007/s11336-021-09832-8

Lozano, J.H. and Revuelta, J. (2021) 'A Bayesian generalized explanatory item response model to account for learning during the test', Psychometrika, 86, pp. 994–1015. doi: 10.1007/s11336-021-09786-x. Available at: https://doi.org/10.1007/s11336-021-09786-x

Lu, J. and Wang, C. (2020) 'A response time process model for not-reached and omitted items', Journal of Educational Measurement, 57, pp. 584–620. doi: 10.1111/jedm.12270. Available at: https://doi.org/10.1111/jedm.12270

Lubbe, W., ten Ham-Baloyi, W. and Smit, K. (2020) 'The integrative literature review as a research method: A demonstration review of research on neurodevelopmental supportive care in preterm infants', Journal of Neonatal Nursing, 26(6), pp. 308–315. doi: 10.1016/j.jnn.2020.04.006. Available at: https://doi.org/10.1016/j.jnn.2020.04.006

Maassen, E., D'Urso, E.D., van Assen, M.A.L.M., Nuijten, M.B., De Roover, K. and Wicherts, J.M. (2025) 'The dire disregard of measurement invariance testing in psychological science', Psychological Methods, 30(5), pp. 966–979. doi: 10.1037/met0000624. Available at: https://doi.org/10.1037/met0000624

Man, K., Harring, J.R. and Zhan, P. (2022) 'Bridging models of biometric and psychometric assessment: A three-way joint modeling approach of item responses, response times, and gaze fixation counts', Applied Psychological Measurement, 46(5), pp. 361–381. doi: 10.1177/01466216221089344. Available at: https://doi.org/10.1177/01466216221089344

Muravyeva, E., Janssen, J., Specht, M. and Custers, B. (2020) 'Exploring solutions to the privacy paradox in the context of e-assessment: Informed consent revisited', Ethics and Information Technology, 22, pp. 223–238. doi: 10.1007/s10676-020-09531-5. Available at: https://doi.org/10.1007/s10676-020-09531-5

Park, J., Arunachalam, R., Silenzio, V. and Singh, V.K. (2022) 'Fairness in mobile phone-based mental health assessment algorithms: Exploratory study', JMIR Formative Research, 6(6), e34366. doi: 10.2196/34366. Available at: https://doi.org/10.2196/34366

Pellert, M., Lechner, C.M., Wagner, C., Rammstedt, B. and Strohmaier, M. (2024) 'AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories', Perspectives on Psychological Science, 19(5), pp. 808–826. doi: 10.1177/17456916231214460. Available at: https://doi.org/10.1177/17456916231214460

Provasnik, S. (2021) 'Process data, the new frontier for assessment development: Rich new soil or a quixotic quest?', Large-scale Assessments in Education, 9, 1. doi: 10.1186/s40536-020-00092-z. Available at: https://doi.org/10.1186/s40536-020-00092-z

Ramminger, J.J. and Jacobs, N. (2024) 'Primacy of theory? Exploring perspectives on validity in conceptual psychometrics', Frontiers in Psychology, 15, 1383622. doi: 10.3389/fpsyg.2024.1383622. Available at: https://doi.org/10.3389/fpsyg.2024.1383622

Röhner, J. and Holden, R.R. (2022) 'Challenging response latencies in faking detection: The case of few items and no warnings', Behavior Research Methods, 54, pp. 324–333. doi: 10.3758/s13428-021-01636-z. Available at: https://doi.org/10.3758/s13428-021-01636-z

Sijtsma, K. and Pfadt, J.M. (2021) 'Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors', Psychometrika, 86, pp. 843–860. doi: 10.1007/s11336-021-09789-8. Available at: https://doi.org/10.1007/s11336-021-09789-8

Sorrel, M.A., Abad, F.J. and Nájera, P. (2021) 'Improving accuracy and usage by correctly selecting: The effects of model selection in cognitive diagnosis computerized adaptive testing', Applied Psychological Measurement, 45(2), pp. 112–129. doi: 10.1177/0146621620977682. Available at: https://doi.org/10.1177/0146621620977682

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E.A. and Fusar-Poli, P. (2025) 'Psychological, psychiatric, and behavioral sciences measurement scales: Best practice guidelines for their development and validation', Frontiers in Psychology, 15, 1494261. doi: 10.3389/fpsyg.2024.1494261. Available at: https://doi.org/10.3389/fpsyg.2024.1494261

Suk, Y. and Han, K.T. (2024) 'A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning', Journal of Educational and Behavioral Statistics, 49(2), pp. 151–172. doi: 10.3102/10769986231171711. Available at: https://doi.org/10.3102/10769986231171711

Sukhera, J. (2022) 'Narrative reviews: Flexible, rigorous, and practical', Journal of Graduate Medical Education, 14(4), pp. 414–417. doi: 10.4300/JGME-D-22-00480.1. Available at: https://doi.org/10.4300/JGME-D-22-00480.1

Trognon, A., Cherifi, Y.I., Habibi, I., Prudent, C. and Demange, L. (2022) 'Using machine-learning strategies to solve psychometric problems', Scientific Reports, 12, 18922. doi: 10.1038/s41598-022-23678-9. Available at: https://doi.org/10.1038/s41598-022-23678-9

Tutz, G. (2022) 'Item response thresholds models: A general class of models for varying types of items', Psychometrika, 87, pp. 1238–1269. doi: 10.1007/s11336-022-09865-7. Available at: https://doi.org/10.1007/s11336-022-09865-7

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R. and Pohl, S. (2021) 'Combining clickstream analyses and graph-modeled data clustering for identifying common response processes', Psychometrika, 86(1), pp. 190–214. doi: 10.1007/s11336-020-09743-0. Available at: https://doi.org/10.1007/s11336-020-09743-0

van der Linden, W.J. and Jiang, B. (2020) 'A shadow-test approach to adaptive item calibration', Psychometrika, 85, pp. 301–321. doi: 10.1007/s11336-020-09703-8. Available at: https://doi.org/10.1007/s11336-020-09703-8

Wyse, A.E. and McBride, J.R. (2021) 'A framework for measuring the amount of adaptation of Rasch-based computerized adaptive tests', Journal of Educational Measurement, 58(1), pp. 83–103. doi: 10.1111/jedm.12267. Available at: https://doi.org/10.1111/jedm.12267

Yadav, D. (2022) 'Criteria for good qualitative research: A comprehensive review', The Asia-Pacific Education Researcher, 31, pp. 679–689. doi: 10.1007/s40299-021-00619-0. Available at: https://doi.org/10.1007/s40299-021-00619-0

Zhang, S. and Chen, Y. (2022) 'Computation for latent variable model estimation: A unified stochastic proximal framework', Psychometrika, 87, pp. 1473–1502. doi: 10.1007/s11336-022-09863-9. Available at: https://doi.org/10.1007/s11336-022-09863-9

Zhang, S., Wang, Z., Qi, J., Liu, J. and Ying, Z. (2023) 'Accurate assessment via process data', Psychometrika, 88(1), pp. 76–97. doi: 10.1007/s11336-022-09880-8. Available at: https://doi.org/10.1007/s11336-022-09880-8

Zickar, M.J. (2020) 'Measurement development and evaluation', Annual Review of Organizational Psychology and Organizational Behavior, 7, pp. 213–232. doi: 10.1146/annurev-orgpsych-012119-044957. Available at: https://doi.org/10.1146/annurev-orgpsych-012119-044957