WINTER 2025, VOL. 38 NO. 1; ISSN 1051-0796

RMT

RASCH MEASUREMENT TRANSACTIONS

Transactions of the Rasch Measurement SIG
American Educational Research Association

Overview of The Issue

The Winter 2026 issue of Rasch Measurement Transactions (RMT) includes several articles and announcements that may be interesting to our community of Rasch measurement researchers.

The issue begins with three research notes, authored by Courtney Donovan, Agustin Tristan-Lopez, and Luigi Tesio.

Next, we present an announcement about the passing of Dr. Jim Sick, authored by Trevor G. Bond.

We end the issue with announcements and updates related to the
AERA Rasch Measurement Special Interest Group (SIG), including a call for nominations for the Benjamin Drake Wright Senior Scholar Award.

As always, we welcome your contributions to the next issue for RMT. We would appreciate receiving your research note, conference or workshop announcement, etc. by March 1, 2026. We respectfully request that you use APA 7 to format your references. Please contact Stefanie at swind@ua.edu or Leigh at leigh.williams@memphis.edu to submit something for inclusion.

Sincerely,
Stefanie A. Wind & Leigh Harrell-Williams

Rasch Measurement Transactions
www.rasch.org/rmt
Copyright © 2026 Rasch Measurement SIG, AERA
Permission to copy is granted.

Editors: Leigh M. Harrell-Williams
& Stefanie A. Wind
Email submissions to: swind@ua.edu
RMT Editors Emeritus: Richard M. Smith,
John M. Linacre, &
Ken Royal
Rasch SIG Chair: Stefanie A. Wind
Secretary: Kaiwen Man
Treasurer: Audrey Conway Roberts
Program Chairs: Kaiwen Man &
Audrey Conway Roberts

“What Measurement Ought to Be”: Integrating QuantCrit & Rasch Philosophies

Abstract

In the Rasch Measurement Transactions, Engelhard (1990) pushed the Rasch community to contemplate “the major measurement theories that have been proposed with consideration of what measurement ought to be.” I argue that we need to continue this conversation by examining QuantCrit alongside Rasch modeling. Rasch and QuantCrit philosophies can go hand in hand to support antiracist measurement. This paper provides context on QuantCrit and practical applications that can complement Rasch measurement philosophy and practices. QuantCrit is more than a simple reflection of our practices. It adds an explicit acknowledgement and commitment to creating equitable measurement tools, and ideally to us as *people* working to humanize and create a stronger measurement field.

Introduction

With the political dismantling of DEI efforts and so many of us in the Rasch SIG and measurement field conducting DEI-focused research, I’d like to push us to reflect on how Rasch and QuantCrit philosophies can align. In the Rasch Measurement Transactions, Engelhard (1990) pushed the Rasch community to contemplate “the major measurement theories that have been proposed with consideration of what measurement ought to be.” I encourage us to continue this conversation by integrating principles of Quantitative Critical (QuantCrit) Theory.

Mead (2008) notes the intent of measurement “is to make inferences based on the measures but analysis is a distinctly separate process from measurement. Measurement does not care if we simply collect and file the measures or use them to achieve world peace” (p.3). Thus, we need to think critically beyond the statistics to understand what is occurring and the frameworks behind how we are interpreting results and differences (Frisby, 2024). Integrating principles of QuantCrit alongside Rasch philosophy can support us to do this. The purpose of this paper is to introduce QuantCrit to the Rasch measurement community. The paper begins with a short reminder of Rasch as a philosophy followed by an introduction to QuantCrit philosophy. Then each of the tenets of QuantCrit are described then the paper concluding with examples on integrating those with common measurement practices.

Rasch Model Philosophy

The Rasch model is both a model and a philosophy. “It is a way of thinking about the world around us so that we can make better sense of the world, so that we can make better decisions about the world” (Linacre, 2007, 0:22). In Rasch philosophy, we believe in the model, not in the data itself. We set parameters and expectations with the underlying belief that the most parsimonious model and best understanding of a construct is the relationship between the difficulty of an item and the ability of a person. Thus, Rasch models “focus on the construct and not on playing with numbers" (Fisher, 1998, p.652).

QuantCrit Philosophy and Principles

QuantCrit research is a relatively new term used to highlight quantitative studies grounded in critical race theory (Ladson-Billings, 2013). Thus, it is built from Critical Race Theory and anti-racists movements, but has been expanded to include sexism, genderism, ageism, ableism, and other forms of discrimination against specific groups. Garcia, López, & Vélez (2018) provide a rich description of the historical lineage of methodologies and studies leading to the coining of the term QuantCrit. Put simply, QuantCrit is a pushback against the misconceived view that quantitative methods are neutral and therefore cannot be biased towards or against anyone. Gillborn et al. (2018) and so many others have opposed this long-held view, stating that "statistics are socially constructed in exactly the same way that interview data and survey returns are constructed, i.e. through a design process that includes… decisions about which issues should (and should not) be researched, what kinds of question should be asked, how information is to be analyzed, and which findings should be shared publicly" (p. 163). The components of QuantCrit therefore become a framework for researchers to reflect upon and actively use to guide our decisions, analysis, and interpretation of results.

While there are slight differences across early literature on QuantCrit, there are six central components which were first described by Gillborn, Warmington, & Demack (2018) and built upon by others (i.e. Van Dusen & Nissen, 2021).

1. QuantCrit holds **the centrality of racism** at its core, "as a complex and deeply rooted aspect of society that is not readily amenable to quantification" (Gillborn et al., 2018, p. 158)
2. **Data and methods are not neutral.** Biases underscore all methods, but quantitative research is often presented in a manner to assume it is objective fact. This component reminds us to interrogate data and methods to minimize and openly discuss biases and assumptions.
3. **Data cannot 'speak for itself.'** Interpretation of analysis is vital. Numbers and statistics have no inherent value. We as researchers, statisticians, and psychometricians assign both the value and the meaning of results. Therefore, critical lens and marginalized voices must become part of this interpretation.
4. **Valuing narrative and counter-narrative.** Critical race studies value personal experiences captured in narratives (dominant voices) and counter-narratives (minoritized voices). While common in qualitative research, quantitative studies need to recognize and value the voices and experiences behind our numbers.
5. **Groups are neither natural nor inherent.** Here we recognize that

groups are created by people for specific purposes; that categorization is part of *human* nature, not of nature itself. Thus, we need to critically evaluate categories, historical context of categories, and groups we use in quantitative research.

6. **Taking an intersectional perspective.** Identity and experiences are multifaceted, so intersectionality is key to more accurate and meaningful understandings.

Integrating Rasch Philosophy, Measurement Practices, and QuantCrit

Anti-racist approaches to measurement are not new (Sablan, 2019; White, Bryd, & Malloy, 2025). Still, the explicit reflection and intentionality of design and interpretation of results from this stance is new to many of us. Keeping QuantCrit as a framework for Rasch measurement studies provides us with concrete strategies, reflection, and advice for interpretation of results. Here I share a few examples to get us thinking, not as an exhaustive list of practices to incorporate.

Intentionality of Design

Logically, we know data is not neutral; people make decisions on what and how it's collected and measured. We also know that agendas, political pressures, career pressures, etc. can influence what we study, what is funded, and what is published. QuantCrit pushes us to intentionally consider all aspects of our study designs (Frisby, 2024). This doesn't mean changing the Rasch model! Instead, it's a deep reflection on all aspects of our study design, starting with asking ourselves, what are we choosing to measure? As well as asking, are we unintentionally perpetuating certain stereotypes?

We must begin with deep considerations to ensure unbiased language in research questions and in sampling (Castillo & Gillborn, 2023). We throw the term "representative sample" around consistently in the measurement field, but representative of who? Who is the reference group and why were they chosen as such? How did you decide your sampling framework? How and where is your data collected and is that appropriate for your target population? What variables are you collecting (and omitting!) and why? What statistical choices are you making and do they account for minoritized groups which tend to have small sample sizes? Statistics used should value confidence intervals and effect sizes over p values due to sample size influences on p values (Van Dusen & Nissen, 2021). These are the questions we must be asking ourselves and our research teams from a design perspective.

Sampling is fundamental consideration because the groups and categorizations we use aren't natural. Meaning we as a society and you as a researcher, decide how groups are defined and measured. For example, which race categories are you using and why? Do you understand the history of how race has been defined and categorized in the US (Lee, 1993) or do you just default to using the census categories as 'the standard?' Are you including Hispanic, Latine, or LatinX and do you know the difference (Soto-Luna, 2023)? Are you including gender categories or sex at birth and acknowledge these may be different (Clarke, 2022)? If you are using gender, are you using two categories or more, and why?

Furthermore, measurement studies begin with a definition of the construct. With QuantCrit we must reflect on where this definition originates. Whose perspective, voice, and experience were used as a reference point for this definition? Were these people included in the measure development process and/or creating indicators of the construct? If you are working with a population that you are not a part of then it is critically important to ensure you have a colleague or expert review to include cultural representation on your research team. Blindly determining what to measure about people from differing cultures without input from them or insight into their culture is an extremely poor practice – that is still occurring in 2025! QuantCrit encourages us to stop and consider every decision we make.

Intersectionality & Invariance

The easiest place to see Rasch models and QuantCrit intersecting is through measure invariance (Morley et al., 2023). We strive for measures that are accurate and unbiased across groups, but this isn't always feasible and frankly shouldn't be expected in all cases. For example, a tool to capture training for novices is expected to be biased towards experts, much like a tool to measure play in children isn't likely to function the same between countries as philosophies and cultures vary in how play is viewed. This doesn't mean the tools don't work! Instead, we show where it does work and with who it works. Invariance allows us to consider intersectionality by examining group differences that we directly measure (i.e. demographic items) and/or indirectly measure by adding in qualitative and historical information to support and explain findings.

Intersectionality is more than simply combining two variables, such as gender and race. Instead, it is an acknowledgement that these social identities interact with the power structures in society that create privilege and oppression (Crenshaw, 2013; McCall, 2005). Therefore, the measurement bias we see could be more than an item simply 'not working as intended' for Black women, as an example. It could be that Black women experience both racism and sexism in a different manner than their counterparts (i.e. Black men, White women) and thus respond differently on that item which would not be captured if looking at differential item functioning by race and by gender separately. Buncher et al. (2025) provides a nice example of using Rasch models to investigate item bias under a QuantCrit framework for those wanting to see how this can look in a study.

Intentionality of Interpretation

Data does not "speak for itself;" humans interpret values and assign meaning. Yet there is still the perception that because you are a quantitative researcher that means your interpretations cannot be biased because statistics are not biased. Frisby (2024) cautions, “the outcomes of quantitative research may challenge the systems that marginalize individuals or perpetuate marginalization” (p.4). We must become cognitive, reflective, and accountable for our interpretations of data being cautious to not perpetuate stereotypes and historical, systematic biases. When we approach studies acknowledging the centrality of racism in society then our interpretations don’t speculate it might due to racism but instead address the impact of racism directly (Van Dusen & Nissen, 2021). This is a subtle but important distinction where we stop referring to group differences as a race or gender gap and instead name the cause, thus referring to the impacts of racism, sexism, and systematic bias. In this way we become more thoughtful researchers as we frame interpretations to avoid encouraging more harm to underrepresented groups we work with.

Finally, we must consider how we interpret findings when the model doesn’t work. We then become investigators to discover the ‘why’ and ‘what’ did not work as intended. Our interpretations should be informed by the experiences and insights of the communities and cultures we are working with. Here is when we get to go back to the qualitative roots of latent measurement to discover why an item or construct works well for Person A but not Person B. This takes us full circle back to the foundation of measurement where we can prioritize narrative and counter narrative stories. These interpretations should also lead to transparency in limitations with whom findings can be generalized to, measurement error, sampling limitations, etc. (Castillo & Gillborn, 2023).

Conclusion

I do recognize that not all studies need or require a QuantCrit framework, but for those where groups, categorizations, cultures, etc. are being considered, I hope that this provides a deeper way to actively consider bias and discrimination. Rasch and QuantCrit philosophies can go hand in hand to support antiracist measurement, but it is more than a simple reflection of our practices. Adding QuantCrit to frame measurement studies must become an explicit acknowledgement and commitment to a specific tool, study, and ideally to us as psychometricians.

Courtney Donovan, PhD
University of Colorado Denver
courtney.vidacovich@ucdenver.edu

References

Buncher, J. B., Nissen, J. M., Van Dusen, B., & Talbot, R. M. (2025). Is the fForce Concept Inventory biased across the intersections of gender and race? Physical Review Physics Education Research, 21(1). https://doi.org/10.1103/physrevphyseducres.21.010137

Castillo, W., & Gillborn, D. (2023). How to" QuantCrit:" Practices and questions for education data researchers and users. EdWorkingPaper No. 22-546. Annenberg Institute for School Reform at Brown University. Retrieved from:

https://scholar.archive.org/work/l6skmthtknfrfilgw3hc5v2xcq/access/wayback/https://edworkingpapers.com/sites/default/files/ai22-546.pdf

Clarke, J. A. (2022). Sex assigned at birth. Columbia Law Review, 122(7), 1821-1898.

Crenshaw, K. W. (2013). Mapping the margins: Intersectionality, identity politics, and violence against women of color. In The Public Nature of Private Violence (pp. 93-118). Routledge.

Engelhard, G. (1990). History and philosophy of measurement. Rasch Measurement Transactions, 4(3), 118.

Fisher W. P. (1998). Metaphysical Rasch. Rasch Measurement Transactions, 1998, 12:3 p. 652. https://www.rasch.org/rmt/rmt123e.htm

Frisby, M. B. (2024). Critical quantitative literacy: An educational foundation for critical quantitative research. AERA Open, 10, 23328584241228223.

Garcia, N.M., López, N. & Vélez, V. N. (2018) QuantCrit: rectifying quantitative methods through critical race theory. Race Ethnicity and Education, 21:2, 149-157, DOI: 10.1080/13613324.2017.1377675

Gillborn, D., Warmington, P., & Demack, S. (2023). QuantCrit: Education, policy, ‘Big Data’ and principles for a critical race theory of statistics. In QuantCrit. Routledge.

Ladson-Billings, G. (2013). Critical race theory—What it is not!. In Handbook of Critical Race Theory in Education (pp. 54-67). Routledge.

Lee, S. M. (1993). Racial classifications in the US Census: 1890–1990. Ethnic and Racial Studies, 16(1), 75-94.

Linacre, J. M. (2007). Rasch measurement: its philosophy. YouTube. https://www.youtube.com/watch?v=VlNt8jqcPZw

Mead, R.J. (2008) A Rasch primer: The measurement theory of Georg Rasch. Psychometrics Services Research Memorandum 2008–001. Maple Grove, MN: Data Recognition Corporation.

McCall, L. (2005). The complexity of intersectionality. Signs: Journal of women in culture and society, 30(3), 1771-1800.

Morley, A., Nissen, J. M., & Van Dusen, B. (2023). Measurement invariance across race and gender for the Force Concept Inventory. Physical Review Physics Education Research, 19(2), 020102.

Sablan, J. R. (2019). Can you really measure that? Combining critical race theory and quantitative methods. American Educational Research Journal, 56(1), 178-203.

Van Dusen, B., & Nissen, J. (2021). Tenets of QuantCrit. https://doi.org/10.48550/arXiv.2110.12871

White, A. M., Byrd, C. M., & Malloy, T. A. (2025). Reclaiming and recasting: An anti-racist approach to psychometric instrument development. Contemporary Educational Psychology, 102340.

Organizing Measurement Evidence for Rasch Analysis: A Macro–Meso–Micro Workflow

The Rasch model provides a coherent mathematical and substantive framework for transforming ordinal observations into interval measures. However, the volume and heterogeneity of Rasch outputs often lead to fragmented interpretations unless results are structured in a purposeful sequence, such as a workflow that distinguishes analysis conducted at three levels: macro (test–person system), meso (individual item or individual person), and micro (response category/threshold).

This workflow offers a heuristic and reporting structure that improves transparency, reproducibility, and interpretability of Rasch-based validation work. It does not introduce new statistical procedures; rather, it synthesizes the knowledge and accumulated experience of diverse authors to prescribe an ordered analytic workflow that aligns existing Rasch diagnostics with decision-making needs at successive stages of measurement evidence.

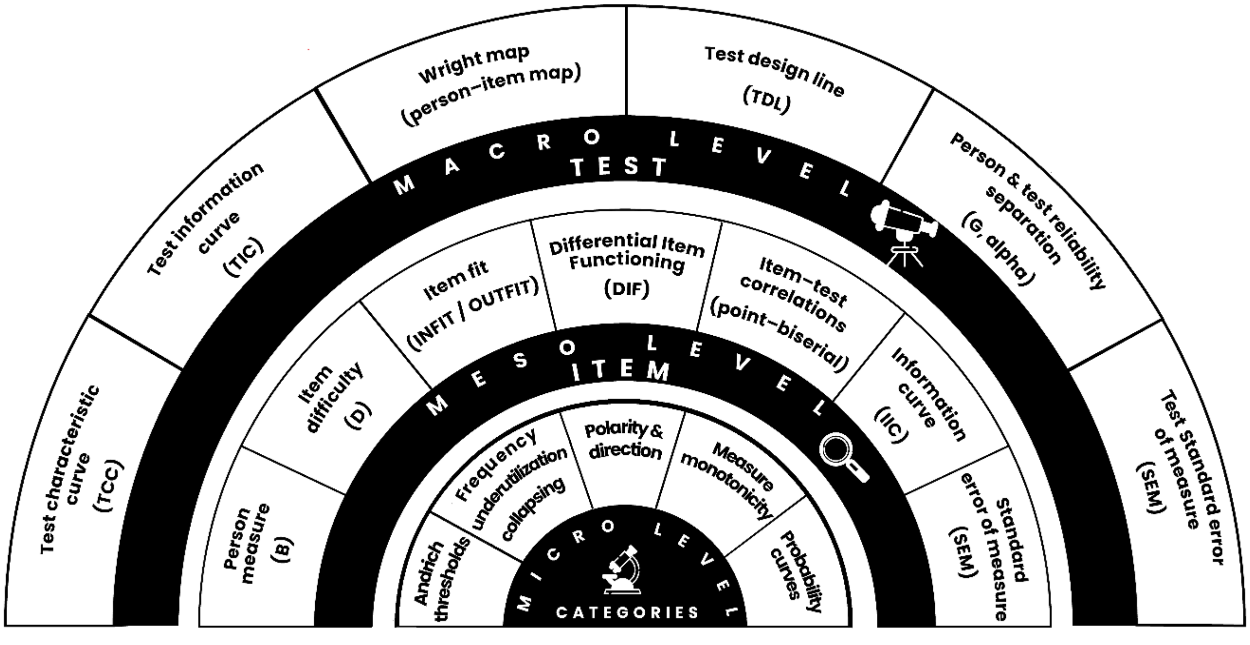
At the macro level, the workflow begins with the global functioning of the instrument across the sampled population. Core activities include assessing construct and scale validity, dimensionality, and evaluating person-ability and test-outcome reliability–separation indices. At this level, the analysis performs overall model-fit diagnostics aggregated across items and persons. Typical outputs include the Wright map (person–item map), test characteristic curves, the test design line, verification of cut-off points, and indices that summarize the instrument's targeting, expected measurement precision, and information across the latent continuum. Macro-level findings determine whether the instrument supports meaningful inferences before proceeding to more detailed diagnostics, whether further item-level refinement is warranted, or whether substantive reinterpretation of the construct is required before engaging in meso-level work (Bond, Yan, & Heene, 2021; Wright & Stone, 2004).

The meso level concentrates on the functioning of individual items, with attention to indicators that may reveal local dependence, content clustering, or unexpected differences across subgroups. Analyses include item calibration (difficulty estimates), item-fit statistics (INFIT and OUTFIT mean square), detection and quantification of differential item functioning (DIF) across relevant subgroups, interitem or item–test correlations, and identification of redundant or poorly discriminating items. The meso inspection is fundamentally diagnostic: it indicates which items conform to model expectations and which require modification or removal to preserve measurement invariance and fairness. Decisions at this level should always be informed by the measurement objectives determined at the macro level (Hambleton, Swaminathan, & Rogers, 1991; Wright & Stone, 1979). Importantly, misfit detected at the meso level may indicate a defective item, but it may also signal divergent response behavior requiring additional attention — for example, person-level anomalies associated with health conditions or contextual factors.

At the micro level, the workflow addresses the internal structure of response categories and the ordinal-to-interval transformation formalized by Rasch models. This involves examining category functioning, Andrich thresholds for polytomous items (for items and persons), category probability curves, and the suitability of rating-scale versus partial-credit formulations. Micro-level diagnostics are essential, particularly when instruments employ Likert-type or frequency response formats. In such cases, the analysis evaluates potential disordered thresholds, category underutilization, or nonmonotonic category measures—all of which compromise the validity of person measures and may require verifying the polarity and directionality of the scale, revising the response format, collapsing categories, or reparameterizing the model (Andrich, 1978; Wright & Masters, 1982).

The macro–meso–micro workflow (Figure 1) enhances efficiency and supports defensible decision rules. A robust macro-level result (e.g., adequate unidimensionality and targeting) justifies allocating analytic resources to meso-level DIF studies and item refinements; conversely, generalized misfit at the macro level signals that meso or micro efforts will be insufficient until the underlying construct representation is corrected. This workflow also clarifies reporting: reviewers and practitioners can follow a logical path from population-level evidence to item-level adjudication and, finally, to category-level verification. Applying this workflow in large-scale empirical studies yields practical benefits,

Figure 1. The Rasch model and the Macro-Meso-Micro Workflow


Wright map (person–item map)
Test design line (TDL)
Test information curve (TIC)
Person & test reliability separation (G, alpha)
Test characteristic curve (TCC)
Test standard error of measure (SEM)
MACRO LEVEL
TEST
Item fit (INFIT / OUTFIT)
Differential Item Functioning (DIF)
Item–test correlations (point–biserial)
Item difficulty (D)
Information curve (IIC)
Person measure (B)
Standard error of measure (SEM)
MESO LEVEL
ITEM
Frequency underutilization collapsing
Polarity & direction
Measure monotonicity
Andrich thresholds
Probability curves
MICRO LEVEL
CATEGORIES

supporting multiple downstream uses—score reporting, scale shortening, and adaptive test development—without sacrificing measurement rigor (Tristán & Vidal, 2007; Bond et al., 2021).

It is important to emphasize that the macro–meso–micro nomenclature is a reporting and procedural convention: the substantive Rasch theory underlying each level remains unchanged. The proposed

workflow simply organizes the various outputs and diagnostics into a coherent sequence that more clearly communicates the evidential basis for measurement decisions—particularly for new practitioners unfamiliar with Rasch terminology and procedures, who may otherwise understand the model only at one of the three levels. Authors adopting this convention should explicitly cite the statistical procedures used at each level and provide reproducible outputs (maps, fit tables, category probability plots) to ensure transparent decision-making for reviewers and end users.

Agustin Tristan-Lopez, PhD
Instituto de Evaluacion e Ingenieria Avanzada, Mexico &
Honorary Research Fellow, Imperial College, London, UK

References

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573. https://doi.org/10.1007/BF02293814

Bond, T. G., Yan, Z., & Heene, M. (2021). Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.). Routledge. ISBN 9780367141424.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage.

Tristán, A. L., & Vidal, U. R. (2007). Linear model to assess the scale’s validity of a test. Paper presented at the American Educational Research Association Annual Meeting, Chicago. ERIC ED501232. https://eric.ed.gov/?q=ED501232&id=ED496126

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. MESA Press. https://www.rasch.org/

Wright, B. D., & Stone, M. H. (1979). Best test design. MESA Press. https://www.rasch.org/

Wright, B. D., & Stone, M. H. (2004). Making measures. The Phaneron Press. https://www.rasch.org/

Cross-cultural validation of questionnaires: More than a measurement issue

Measurement as a fact: relative and objective

Measurement represents a quantitative difference, reflecting a relationship between objects. It's important to note that there is no such thing as an "absolute" zero measure, which may seem counterintuitive. For example, if a cube were the only object in an otherwise empty universe, and you were an observer outside this universe, you would not be able to determine whether the cube is "big" or "small" (an absolute value judgment). However, if there are two cubes, you can objectively compare their sizes—regardless of the measuring instrument, the objects, the observer, or any extraneous factor—according to fundamental measurement principles (Luce & Tukey, 1964).

Equating of questionnaires across classes of respondents: a thorny issue

Same questionnaire, different variables?

A Rasch analysis draws raw ordinal scores on cumulative questionnaires ("scales") into the same realm as physical measures. An ideal measure, including those derived from questionnaires, should be independent of the objects or individuals being measured, as long as the data fit the Rasch model. This means that the scale structure, i.e., the difficulty levels or "hierarchy" of the items or their categories in polytomous scales, is invariant, similar to the ticks of a metric ruler, across distinct classes of respondents. Unfortunately, each class may be influenced by specific sources of distortion of the items' hierarchy due to extraneous factors. As a result, the same questionnaire may address qualitatively different variables, despite the reassuring appearance of shared numeric measures.

A special case arises with questionnaires that have been translated for use in different linguistic and socio-cultural contexts, leading to two types of difficulties.

Difficulty #1: Semantic non-equivalence

A literal translation may only provide an approximate meaning. For example, the adjective "aching" is often used in various pain questionnaires. In many European languages, "ache" is considered synonymous with "pain." However, the phrase "aching pain" is meaningful to English speakers, while it appears redundant (like "painful pain") in other languages.

In my view, "aching" describes the pervasive nature of pain: "I feel bad or ill because of this pain" (as in toothache, headache, or bellyache), which is a different experience from simply feeling pain in a specific body part (such as knee pain, back pain, or eye pain). A sophisticated Italian psychometric study translated the adjective "aching" as "dà sofferenza" ("it makes you suffer.") (Maiani & Sanavio, 1985). When I attempted to use this item to develop a new Rasch-consistent back pain scale, it did not fit the model appropriately. The term "suffering" includes psychological experiences that are not directly related to physical pain. This may explain why some respondents rated "dà sofferenza" very low, even though they had an overall severe condition. Other respondents provided erratic answers, interpreting "suffering" as synonymous with "pain" (Tesio et al., 1997)

Difficulty #2: Metric non-equivalence
Assuming that a literal-semantic translation is possible, the meaning of "how difficult" in different contexts remains to be determined. This may occur for two distinct reasons.

2a - Differential item functioning
Consider a different example from my experience with the Functional Independence Measure (FIM; Tesio et al., 2002), an international standard. On a scale measuring independence in activities of daily living (ADL), like the FIM, "Eating" may be accurately translated from English to Japanese. However, for any patient with upper limb disability, using a spoon or fork is often easier than eating with chopsticks. While "Eating" is usually the simplest item on ADL scales, it may become more challenging than tasks like Grooming or Upper-body dressing within communities that primarily use chopsticks. This illustrates a concept known as Differential Item Functioning (DIF). In other words, passing the "Eating" item indicates a higher independence in ADL in Eastern patients compared to their Western counterparts. Despite appearing numerically similar, the two versions of the scale do not measure the same variable, which is a subtle yet crucial distinction. Across European countries, too, the stability of item hierarchy in disability scales is not guaranteed. Rasch analysis effectively addresses, in part, this type of DIF through an "equating" procedure known as the split-item technique, which results in a shared hierarchy (Tennant et al., 2004; Tennant et al., 2024; Wright & Stone, 2004). This elegant solution, however, does not address another issue that, to my knowledge, remains overlooked in the Rasch literature: the equivalence of items' *values* across different respondent classes.

2 b - The same relative difficulty levels of measures may not mean the same absolute ability levels for persons.
Suppose you have shown that the item hierarchies are consistent across two groups of respondents, such as citizens of Country A and citizens of Country B. In that case, you can conclude that your questionnaire is measuring the same variable. However, you cannot assert that an "ability" measurement of, for example, two logits has the same absolute meaning for individuals in Country A as it does for those in Country B. By "meaning", I refer to the value that a score (which is directly related to the measure in Rasch-modeled questionnaires) holds in different Countries.

The term value here relates to the expected level of performance in a given item, in a specific context. For instance, the capacity to perform tasks such as "doing a heavy job," "walking around a block," or "engaging in social interaction" varies based on how one defines a "heavy" job, the size of the urban "blocks," and the complexity and formality of social interactions. Essentially, the scores assigned (like 0, 1, 2, etc.) are dependent on the specific context: a subject scoring "1" in one context may receive a higher score of "2", in the same item, when placed in another setting. However, it can be said that within each context the values are "absolute" — a term derived from the Latin adjective meaning "loose" or "untied" — as they are predetermined with respect to the scores on the questionnaire.

To clarify this concept, an example may be helpful. Consider a scenario where an ADL scale is accurately translated from the language spoken in Country A to that of Country B. Additionally, assume that Rasch analysis finds no evidence of DIF by Country, meaning the same hierarchical structure holds across both contexts.

Now, imagine that people in Country A have the same motor capacities of people in Country B. However, houses in Country A are consistently smaller, less comfortable, and less barrier-free compared to those in Country B, for many reasons—be they climate, the geography of the area, available building materials, traditions, people's income, etc. As a result, people in Country A will likely score items related to ADLs lower, indicating that more assistance is required, compared to Country B. However, Rasch analysis would assign the same "zero" measure to the average item difficulty level in both countries.

An intervention designed to enhance patients' independence, like physiotherapy, may result in similar changes in persons' measurements in both Countries. However, this change won't necessarily lead to the same overall level of independence: for instance, one allowing a 50% reduction in the caregiver's burden, or allowing the prediction of a safe return home (Stineman et al., 1997; Kushner et al., 2023). The information that differences in independence stem from housing rather than individuals remains obscured.

The example above clarifies that a *value* judgement is not determined solely by measurement, which indicates the relative position of an item (or person) along a "less-to-more" gradient. Instead, *value* is influenced by one or more context-dependent external criteria that interact in a complex way.

Value judgements are not decisions.

Value judgments, along with measurements, play a significant role in decision-making processes. For example, for disabled inpatients, a given ADL score is associated with a given likelihood of being discharged home. The housing context affects how valuable that score is. However, the final clinical decision will also depend on the availability of caregivers, nursing services, and other relevant resources.

Equating values, not only measures

In summary, only “more/less” statements are related to measurement, whereas terms like good/bad, sufficient/insufficient, correct/wrong, and even normal/pathologic (10) reflect value judgments. This creates an apparent paradox: measurements are relative yet still objective, while value judgments are “absolute”, yet they are subjective or conventional.

Is it possible to equate the *value* of measures, rather than just the measures themselves? First, you need to be sure that "values" are comparable.

a-zero value, regardless of the scores
For instance, independence in ADL may be considered a "zero" value in contexts where receiving significant assistance from family members or service providers is perceived as a status symbol, even when individuals are in good health.

b- infinite value, regardless of the score
As an opposite example, consider the variable "Health-related quality of life-HrQoL" to which many questionnaires are dedicated. Some believe HrQoL (and even QoL in general) can be quantified more objectively by virtually trading "years in bad health" with "years in good health," or years with disability with years without disability. In this model, a higher ratio indicates a lower quality of life. This econometric approach may justify the rationing of health care resources to classes of individuals based on numerical indexes such as Quality-Adjusted Life Years (QUALY; Whitehead & Shehzad, 2010) and Disability-Adjusted Life Years (DALY) (Hay et al., 2017), raising severe ethical concerns. However, comparison of measures across two cultures is not possible if in one or both of them the "quality" of human life is assigned a unique value, perhaps an infinite one, so that "years" cannot be traded, regardless of the QoL scores (Oliver, 2004; Tesio, 2009).

Next, you must determine which item in one linguistic version corresponds to the difficulty of a given item in the other version. Just one item is sufficient due to the "sufficiency" property of Rasch measures. For example, in an ADL scale, you might decide that the difficulty of "Eating" in Country A corresponds to the difficulty of "Dressing lower body" in Country B. What subjective criteria influence this decision? There are no rules of thumb. One possibility is to base the decision on the average time caregivers spend assisting with these two activities (an external criterion-referenced decision) (Granger et al., 1990). Another possibility is comparing the percentage of people who can perform these different activities independently (a distribution-based criterion).

Once you "anchor" the two items, if the hierarchies of the item sets are equivalent, your process of equating is complete, allowing for fair comparisons of individuals' "abilities". However, the choice of "item equating" will inevitably remain a subject of debate.

Absolute statements imply subjective honesty.

According to the reliable GIGO law (garbage-in, garbage-out), if the measurement is correct, the ensuing decisions will likely be right; if the measurement is incorrect, the subsequent decisions will most likely be wrong. However, measurement is not enough.
The final lesson is that measurements and decisions do not tell the same story, so that

in doing research, one can never disentangle objective measurement from subjective value judgment (16). This is not a flaw, but rather the confusion is.

Luigi Tesio
Istituto Auxologico Italiano, IRCCS, and University of Milan, Italy
l.tesio@auxologico.it

References

Canguilhem, G. (1991). *The normal and the pathological.* (M. Foucault, Intro.). Zone Books.

GBD 2016 DALYs and HALE Collaborators. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: A systematic analysis. *The Lancet, 390*(10100), 1260–1344.

Granger, C., Cotter, A., Hamilton, B., Fiedler, H., & Hens, M. (1990). Functional assessment scales: A study of persons with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation, 71*(11), 870–875.

Hay, S. I., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., Abera, S. F., Aboyans, V., Abu-Raddad, L. J., Ackerman, I. N., Adedeji, I. A., Adetokunboh, O., Afshin, A., Aggarwal, R., Agrawal, S., Agrawal, A., … Murray, C. J. L. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet, 390*(10100), 1260–1344. https://doi.org/10.1016/S0140-6736(17)32130-X

Kushner, D. S., Johnson-Greene, D., Felix, E. R., Miller, C., Cordero, M. K., & Thomashaw, S. A. (2023). Predictors of discharge to home/community following inpatient rehabilitation in a U.S. national sample of Guillain–Barré syndrome patients. *PLOS ONE, 18*(5), Article e0285427.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1–27.

Maiani, G., & Sanavio, E. (1985). Semantics of pain in Italy: The Italian version of the McGill Pain Questionnaire. *Pain, 22*(4), 399–405.

Oliver, A. (2004). Prioritizing health care: Is “health” always an appropriate maximand? *Medical Decision Making, 24*(3), 272–280.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Harvard University Press.

Stineman, M. G., Goin, J. E., Granger, C. V., Fiedler, R., & Williams, S. V. (1997). Discharge Motor FIM–Function Related Groups. *Archives of Physical Medicine and Rehabilitation, 78*, 980–985.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. L., Slade, A., et al. (2004). Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR project. *Medical Care, 42*, 137–148.

Tesio, L. (2009). Quality of life measurement: One size fits all. Rehabilitation medicine makes no exception. *Journal of Medicine and the Person, 7*(1), 5–9.

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 2: Advanced model applications and the data–model fit assessment. *Disability and Rehabilitation, 46*(3), 604–617.

Tesio, L., Granger, C. V., & Fiedler, R. C. (1997). A unidimensional pain/disability measure for low-back pain syndromes. *Pain, 69*, 269–278.

Tesio, L., Granger, C. V., Perucca, L., Franchignoni, F. P., Battaglia, M. A., & Russell, C. F. (2002). The FIM™ instrument in the United States and Italy: A comparative study. *American Journal of Physical Medicine & Rehabilitation, 81*(3), 168–176.

Whitehead, S. J., & Shehzad, A. (2010). Health outcomes in economic evaluation: The QALY and utilities. *British Medical Bulletin, 96*(1), 5–21.

Wright, B., & Stone, M. (2004). *Making measures.* Phaneron Press.

In Memorium of Vale James Sick, Ed.D.

It is my sad duty to report to you the death of our colleague and friend Jim Sick on January 6, 2026. Dr. Sick was a long-standing contributor to the Pacific Rim Objective Measurement Symposia focussing on the Rasch measurement of affective and cognitive individual differences amongst language learners, as well as the measurement of proficiency, achievement, and progress in language skills. Jim lived in Japan for many decades, but many of you will be surprised to learn that he arrived there from the US as a jazz guitarist—having studied music in Florida and Boston, then playing in San Francisco. Like many expats in Japan during that era, Jim gradually got into EFL teaching. He must have taken that choice rather more seriously than most because he went on to earn his Master's and Ed D degrees from Temple University Japan (TUJ), with particular interests in second-language assessment and statistical analysis. Jim's next logical step, of course, was expressed in his growing interest and expertise in Rasch measurement.

Dr Jim Sick became an active of the member of the Japan Association of Language Teaching (JALT) as well as its Testing and Evaluation SIG (TEVAL); he served as TEVAL President for four years. Jim taught English as a foreign language at high school, university, and graduate school levels, as well as teaching courses and workshops on language assessment, Rasch measurement, and technology assisted language learning. He ran PROMS workshops focussing on Facets analysis (MFRM) in particular, and was an early rep for Japan on the PROMS management board when I was the Chair. I always appreciated Jim's contributions to our business meetings as well as his substantive contributions in Rasch measurement presentations and workshops.

Dr Sick made his formal contributions to second language assessment as an Adjunct Professor, Temple University, Japan Campus, and Visiting Professor, Takushoku University Graduate School of English Education and served as a dedicated dissertation advisor to numerous doctoral students. He wrote on Rasch measurement in education for Shiken, the journal of the JALT TEVAL SIG. Jim's colleagues in Japan saw him as always great to work with, always very positive; and, of course, good to ask for advice about measurement. His scholarly contributions and leadership have left a lasting impact on the field of language assessment in Japan. Jim will be remembered by us as a friend.

Trevor G. Bond
PROMS Founder

Updates and Announcements from the Rasch Measurement Special Interest Group (SIG) of the American Educational Research Association (AERA)

The AERA Rasch SIG has several important updates to share with the Rasch community.

We appreciate your engagement with the SIG and look forward to connecting with you through SIG activities.

Call for Nominations: Benjamin Drake Wright Senior Scholar Award

We are currently accepting nominations for The ***Benjamin Drake Wright Senior Scholar Award***, which is an AERA-sanctioned award. This award is presented to an individual senior scholar for outstanding programmatic research and mentoring in Rasch measurement over the course of a career and who is still active in Rasch measurement research at the time the award is granted. **The award is open to scholars**

worldwide. Membership in AERA or Rasch Measurement SIG is not required of the nominee.

Eligibility Criteria: To be eligible for the award, nominees should meet the following criteria:

a. The nominee has designed and carried out programmatic research that originates in Rasch measurement and helps understand crucial phenomena in model definition, parameter estimation, fit assessment, construct specification, novel applications, the place of Rasch measurement in the history and philosophy of science, etc., as represented in a corpus of writings and research projects that have contributed to the theoretical development of the field as well as having been grounded empirically; AND
b. The nominee has developed the research capacity of the field, as attested to by the existence of a "school of thought" or intellectual heritage associated with the scholar's name, a heritage that includes other individuals whom the scholar has had a direct influence in encouraging and helping become productive in Rasch measurement research or an identifiable domain of Rasch measurement research within which the nominee's constructs and results are used regularly by other researchers.

The Rasch Measurement SIG recognizes also that other features of a person's work might add to the criteria above, strengthening a nomination. Among the criteria that could add to the basic ones is one or more of the following:

The nominee may also have made:

a. major contributions to broader fields of research in education, psychology, health care, or the social sciences, as represented by his or her participation (as author, speaker, or consultant) in research forums from fields other than Rasch measurement or by the recognition of his or her scholarship in other fields of inquiry (inclusive of all of educational research and the social sciences); OR
b. major impact on the practice of Rasch measurement, as represented by the existence of policy documents, curriculum materials, professional development programs, or a corpus of practitioner- or public-oriented literature to which the nominee has significantly contributed as an author.

The Award

The award includes a plaque and an invited address for the 2026 Rasch SIG business meeting at the annual AERA conference. An honorarium will be provided.

Nominations should include (and are restricted to) the following:

Individuals will be nominated via a letter of nomination emailed to the Rasch SIG secretary proposing the name of the nominee and describing the grounds on which the nominee meets the requirements for the award. Three criteria should be addressed in the letter:

- A brief (no more than 250-word) description of the program of research carried out by the nominee;
- A list of significant publications representing the contributions

described; and a list of scholars who have been significantly affected by the work of the nominee. The list of scholars may include, but need not be limited to, doctoral students who worked with the nominee. Current contact information for the list of scholars should also be included in the nomination.
- The nominee's CV.

Self-nominations will not be accepted.

The deadline for nominations is Friday March 13, 2026. Nominations should submitted by sending an email to the SIG Secretary, Kaiwen Man at kman@ua.edu.

Update on SIG Membership: Please Renew for 2026!

I am pleased to share that, after being placed on probationary status in 2025 due to declining membership, our membership roster exceeded AERA's minimum requirement by the end of the year. My sincerest thanks to those of you who renewed your memberships, joined the SIG, and encouraged your colleagues to do the same.

To avoid this situation again, please be sure to renew your SIG membership in 2026.

Sincerely,

Stefanie A. Wind
Chair, *Rasch Measurement SIG*