Brief Overview of Kane's Argument-Based Approach To Measurement Validation

Situated Within the Evaluation of Mixture Computerized Adaptive Test for Measuring Patient-reported Outcomes

BRUNO D. ZUMBO
University of British Columbia

RICHARD SAWATZKY
Trinity Western University

PAMELA A. RATNER
University of British Columbia

This report is part of the UBC Psychometric Research Initiative's knowledge mobilization and dissemination efforts to translate psychometric concepts for PROMs research. These efforts are funded in part by the Government of Canada's Canada Research Chairs Program in support of the Tier 1 Canada Research Chair in Psychometrics and Measurement, held by Professor Zumbo, and administered through the Social Sciences and Humanities Research Council.

Version March 8, 2026

CONTENT

CTRL+MOUSE CLICK TO FOLLOW THE LINK

Brief Overview of Kane's Argument-Based Approach To Test Validation

This brief note summarizes Kane's argument-based approach to validation and is situated within our work on developing a mixture computerized adaptive test (mixture-CAT) patient-reported outcome measure (PROM). While contextualize the overview with reference to the measurement of emotional well-being.

We also note a feature of the report's expository style. Because the report is intended to translate psychometric concepts for PROMs research, several foundational concepts recur throughout. Their meaning becomes clearer and more fully developed as related methodological and theoretical elements are introduced. As a result, key terms reappear across sections, allowing their significance to evolve with the broader framework.

Why Care About Measurement Validity? The "So What?" Question

Researchers often ask what they gain from the time and effort required for validation and why validity work relies on technical language that has developed over more than a century of theory and practice (Hubley & Zumbo, 1996; Zumbo, 2023). We address these questions by framing validity as the "Cinderella" of psychological and health measurement, PROMs research, and—at times—even psychometrics: treated as essential in principle but subordinated in practice.

Measurement Validity as the "Cinderella" of Psychological Assessment, PROMs, and Psychometrics

Calling measurement validity the **"Cinderella"** of a field is a **heuristic metaphor**: One is mapping features of the Cinderella story (central character, undervalued/ignored, later recognized) onto the status of validity work in research.

It bears repeating that, despite its centrality to measurement, researchers often treat measurement validation as an afterthought, **essential in principle but subordinated in practice**. This heuristic metaphor captures three recurring patterns.

- First, many texts and validation reports still define validity as the extent to which scores measure "what they are intended to measure" (Zumbo & Chan, 2014). In psychometric terms, this formulation risks a category error by reifying the operational scoring rule (i.e., the item set, scoring algorithm, and statistical model) as the construct itself, thereby conflating the mapping from observations to numbers with the substantive attribute those numbers are meant to represent; that is, the target construct. Absent an explicit account of the inferential links connecting observed responses, the scoring model, and the target construct, "intended-to-measure" can become circular

rather than evidentiary. Consistent with this concern, studies—particularly those using self-report measures of mental health, well-being, and pain—often foreground reliability, factor-analytic results, or correlations with convergent and discriminant measures while giving less attention to whether the proposed score interpretations and uses are warranted.

- Second, when investigators treat validation as a primary analytic task, they strengthen the credibility and utility of the claims and decisions drawn from PROMs and other assessment scores and reduce the risk of conclusions that the scores cannot support.
- Third, as interdisciplinary and applied PROMs research expands, validation is receiving renewed attention, creating an opportunity to align measurement practice with equity and person-centred aims.

Some widely used instruments, such as the Short Form–36 Health Survey (e.g., SF-36; Giraldo-Rodríguez & López-Ortega, 2024; Treanor & Donnelly, 2015), have accumulated substantial validity evidence, including evidence of cross-cultural, language, and population use. In contrast, many measures remain supported primarily by surface-level indices and limited justification of score meaning. Hubley et al. (2024) cautioned that "measures that are well-recognized or commonly used are often assumed to be of high quality, but it is sometimes surprising how little validity evidence exists to support the intended inferences" (p. 529). This concern underscores the need to prioritize validation in both theoretical and applied PROMs research.

A Critique and Further Unpacking of the Claim That a Test is Valid if it Measures What it Intends to Measure

Because of its prevalence in the research literature and textbooks, it is worthwhile to unpack the definition of validity as the extent to which an instrument measures "what it intends to measure." In short, the problem is that this definition invites a category error. However, what does that mean in the context of measurement validity? In short, it treats the scoring procedure as if it were the construct itself. In psychometric terms, this framing collapses the distinction between (a) the **measurement rule**—the item content, scoring algorithm, and statistical model that map observations to numbers—and (b) the construct's **ontological content**—the substantive attribute in the world that the score is meant to represent. When authors implicitly define the construct as whatever the scoring rule produces, the definition becomes circular (i.e., the instrument is "valid" because it yields the scores it was designed to yield) rather than evidentiary.

This category error matters because the meaning of scores rarely follows automatically from an instrument's design. A measurement rule specifies *how* scores are generated. However, it does not, by itself, justify *what* those scores represent across contexts and populations. For example, a PROM item set might operationalize emotional well-being primarily in terms of calmness, optimism, control, and satisfaction. Yet the lived content of emotional well-being—particularly across cultures, traditions, and life experiences—may also involve relational obligations, spiritual harmony, moral emotions (e.g., shame or honor), collective stress, or culturally specific idioms of distress. If the construct is reduced to the instrument, researchers can mistakenly conclude that the PROM "validly measures emotional well-being" simply because it measures what its developers intended. The psychometric question, however, is whether the item pool and scoring model support defensible inferences about emotional well-being as experienced by respondents, for the intended uses and decisions.

> **Sidebar**: This definition of validity as whether an instrument measures "what it intends to measure" shares a great deal with the problematic use of operational definitions. The core

> mistake in operational definitions in psychology, for example, was treating the means of knowing a construct (the operation) as identical with the construct itself (the phenomenon). This confusion between epistemic access (how we measure) and ontological status (what exists) produced a century of conceptual flattening. For example, in PROMs research, the concept of *pain* would become "whatever pain scales measure." This was not a minor definitional slip; it was a category error: conflating a construct's measurement rule with its ontological content.
>
> Operationalism hardened into dogma because it was institutionally convenient: it enabled standardization (log file manuals, psychometric norms), it facilitated quantification (scores as data points), and it promised objectivity (procedural reproducibility). However, these advantages came at a massive theoretical cost: conceptual reification—treating task-dependent observables as the phenomena themselves.
>
> Defining validity as whether an instrument measures "what it intends to measure" invites a similar category error and hardening into dogma by treating a scoring procedure as if it were the construct itself. This collapses the distinction between the operational rule that generates scores and the substantive phenomenon those scores are meant to represent, making validity claims circular unless they are grounded in explicit theory and evidence.

A PROMs Example of How the "Measures What It Intends to Measure" Definition Invites a Category Error

The Strongest Warrant for Why We Need Contemporary Validity Theories Like The Argument-Based Approach

Early validity definitions treated validity as an inherent test property, exemplified by the 1920s assertion that "a test is valid if it measures what it is supposed to measure" (Hubley & Zumbo, 1996; Zumbo, 2023). This intuitive but operationally vague definition provided little guidance for empirical evaluation. Behaviorism's dominance in early 20th-century psychology influenced validity research through criterion-related approaches, which emphasized observable outcomes and predictive ability. However, researchers recognized that many constructs—particularly in emerging PROMs research—lacked reliable criterion measures.

Defining validity as whether an instrument measures "what it intends to measure" can make validity appear self-certifying. If instrument and survey developers intend an item set and scoring rule to define "emotional well-being," and then judge the measure as valid to the extent that it measures "emotional well-being," the claim becomes circular when the construct is implicitly treated as whatever the scoring procedure yields. In psychometric terms, this is a category error: it reifies the operational measurement rule (items, scoring algorithm, and model) as the construct itself, rather than treating the construct as a substantive attribute that the measurement rule is meant to represent.

A PROM illustrates the problem. An instrument might operationalize emotional well-being primarily in terms of feeling calm, optimistic, in control, or satisfied. Those item choices constitute a measurement rule and embed a particular content specification. However, the lived content of emotional well-being—especially across cultures and life experiences—may also include spiritual harmony, relational obligations, moral emotions (e.g., shame or honor), collective stress, or culturally specific idioms of distress. When researchers collapse the construct into the instrument, they risk concluding that the PROM "validly measures emotional well-being" because it measures what the developers intended the items to capture. The relevant validity question is instead whether the item pool and scoring model support defensible inferences about emotional well-being as experienced by respondents, for the populations and uses at issue.

Contemporary validity theories, including argument-based validation, address this risk by keeping constructs and measurement rules analytically distinct and by evaluating the inferential links between them. On this view, validity does not rest on fidelity to intention; it rests on the degree to which theory and evidence support the proposed interpretations and uses of scores, including the consequences of acting on those scores. As such, it is the **strongest warrant for** the need for contemporary validity theories, such as **the argument-based approach**.

A Central Yet Often Overlooked Psychometric Question

A central yet often overlooked question is: What does a scale score mean? Recall that a scale score is a summary score derived from responses to multiple survey items intended to measure a common psychological or health-related construct. Whether used to describe individuals, summarize group outcomes, or build statistical models, scale score interpretation underlies every claim made from measurement. Nevertheless, in practice, we often glide by this foundational issue. For example, Gadermann et al. (2023) used QOL-related measures to examine teachers' well-being during the pandemic. They reported that the education system's mental health supports primarily predicted workplace well-being.

In contrast, personal COVID-19 stressors most strongly predicted general mental health. These inferences about mental health and well-being may be plausible. However, without measurement validation evidence for each of these instruments in their context of use, the claims of the different predictive capacities remain vulnerable to alternative (plausible rival) explanations. This is particularly the case given that contemporary psychometric theory embraces the notion that scores based on responses to a survey instrument, aggregate survey respondent scores computed across a set of survey items, can reflect the intended construct as well as minor, but yet potentially influential, construct-irrelevant variance, including scoring features and social biases (e.g., Slocum-Gori et al., 2009; Gelin & Zumbo, 2003).

Transitioning From Fragmented Validity Evidence to Contemporary Approaches

Growing dissatisfaction with criterion-based models from the 1930s to the late 1950s prompted the development of alternative approaches to validity. During this period, researchers proposed forms of validity such as factorial, face, intrinsic, and empirical validity, and often treated strong correlations as sufficient evidence. Guilford's (1946) assertion that "a test is valid for anything with which it correlates" reflected this correlation-based approach, which favored empirical association over theoretical coherence.

Cronbach and Meehl's (1955) formulation of construct validity marked a major shift toward theory-driven validation. They redirected attention from the instrument itself to the interpretation of scale scores. They emphasized the investigator's theoretical orientation rather than any single validation procedure. However, although their account moved validity theory in a more explanatory direction, it also blurred the distinction between validity as a concept and validation as a process. Their view that validation applies to score interpretations rather than to tests themselves has remained both influential and contested in subsequent theories of validity.

Cronbach (1971, p. 483) criticized researchers for treating validity as a disjointed concept. He noted that many validity studies reduced construct validity to a haphazard collection of correlations, rather than deliberately contrasting and comparing them, guided by meaningful theoretical interpretation. Recent evidence (e.g., Hubley et al., 2024; Zumbo & Chan, 2014) suggests that Cronbach's concerns remain valid, including that validation studies often lack explicit theoretical foundations to guide the interpretation of validity evidence.

As Shear and Zumbo (2014) observe, lacking an overarching theory of validity poses a greater challenge to interpreting and reporting measurement validity evidence than missing any single kind of validity concept. Without a unifying framework, researchers cannot confidently determine whether their validation efforts have achieved their intended goals. This theoretical vacuum makes it difficult to compare results across different validity studies, since each may operate under distinct, unspoken assumptions about what validity means. Such inconsistency undermines the statement in the *Standards* that validity is the most fundamental consideration in developing and evaluating survey instruments (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9) because it may not be clear what exactly a concern for validity entails in instrument and survey development and evaluation, leaving the term undefined and open to multiple interpretations.

To paraphrase Cronbach's observation, vague descriptions of validity likely lead to a haphazard assemblage of validity evidence. The argument-based approach provides a structured framework for organizing validation studies.

One Central Consideration is to Ensure We are Making Sound and Justifiable Conclusions and Decisions Based on the Instruments We Use

Validation addresses this consideration by evaluating whether evidence and theory support the proposed interpretations and uses of scores. When the evidence supports those interpretations and uses, researchers may treat them as valid inferences in line with the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Kane's Argument-Based Framework Emerges as a Dominant Viewpoint

As described by Zumbo (2023), more than a century of work on measurement validation – centered on the process of collecting and interpreting evidence from measurement practices to argue for the validity of our judgments and decisions about survey respondents – has led to an evolution in how we understand and undertake measurement validation. Moving away from focusing on types of validity, the field has increasingly focused on sources of evidence to support construct validity as a single type of validity within a unified framework. Mike Kane's argument-based validation framework has emerged as a dominant approach. In essence, all of the argument-based approaches, including Kane's approach, promote a two-step process: (i) beginning with statements or arguments of what we want to argue in the inferences we make from assessment scores, and (ii) then evaluating the defensibility of these arguments by interpreting relevant types of evidence that can support or refute each of the arguments.

The argument-based validation framework has received strong support in PROMs research, as it applies to both quantitative and qualitative assessment tools and to assessment programmes that utilise multiple assessment data points and forms.

Kane's argument-based approach to validation emerged as an influential development in validity theory, responding to complexity arising from the greatly expanded view of validity and validation practice (Cronbach, 1988; Kane, 1992, 2006, 2013; Shepard, 1993). Kane (1992) introduced this influential argument-based approach, providing a disciplined methodology for establishing validation plans and interpreting various types of validity evidence in validation studies.

Rather than defining validity, this approach serves as a "methodology or technology for validation" (Kane, 2004, p. 136) that supports different validity definitions. The approach centers on validating inferences and uses rather than survey instruments. Kane distinguishes between an interpretive argument (stating assumptions and inferences from observations to interpretations) and a validity argument

(evaluating the plausibility of proposed inferences). This framework guides researchers in allocating their effort and gauging the progress of validation.

What Kane Means by "Validation"

In this approach, **validity is not a property of the instrument or measure itself**. Instead, validity is about whether we can **justify the proposed interpretations and uses of scores**—especially the decisions that follow from them. Validation is therefore the process of **collecting and interpreting evidence** to support (or refute) that justification.

Two Linked Arguments: IUA and Validity Argument

Kane distinguishes between:

Interpretation/Use Argument (IUA)**:** a *structured statement* of what you want to claim about score meaning and score use—i.e., your proposed chain of reasoning from performance → score → interpretation → decision.

A Walk Through Argument-Based Validation in PROMs

PROMs Research

PROMs research examines how to measure patients' health experiences accurately and use those measurements to improve care, research, and decision-making. PROMs research matters because healthcare is not only about survival or lab values. It is also about whether people can sleep, work, think clearly, manage pain, and live well. In that sense, PROMs research helps make healthcare more patient-centered by treating the patient's experience as evidence rather than just anecdote.

Researchers in this field usually focus on a few main goals.

First, they **develop questionnaires or rating scales** to measure symptoms, physical functioning, emotional well-being, and social participation.

Second, they **test whether those questionnaires are trustworthy**. For example, they ask:

- Does the measure actually assess what it claims to assess?
- Are the scores consistent?
- Can the measure detect meaningful change over time?
- Do people from different groups interpret the items similarly?

Third, they study **how to use PROMs well in practice**. A questionnaire may work in theory, but researchers still need to know whether it is understandable, fair, useful in clinics, and sensitive enough to guide decisions.

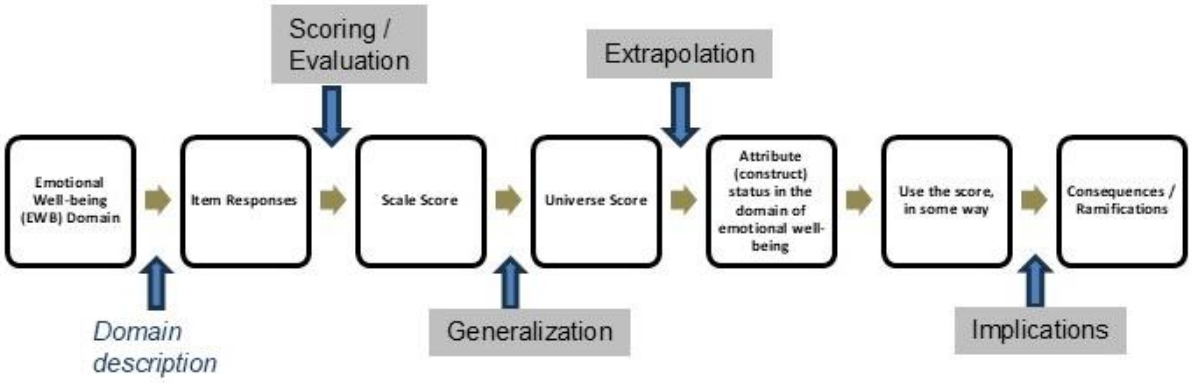
A Glance at Argument-based Measurement Validation

More recently, validation scholarship in PROMs has increasingly emphasized a two-step process: first, specifying the claims that underlie the inferences drawn from PROM scores, and second, evaluating the defensibility of those claims by considering the evidence that supports or refutes each inference. This argument-based validation framework is especially well-suited to PROMs research because it can be applied to both quantitative and qualitative measurement approaches, as well as to broader assessment programs that draw on multiple sources of data, repeated measurements, or different forms of assessment.

Let us describe the first step of the two-step process. Figure 1 presents Kane's original four inferences—scoring (or evaluation), generalization, extrapolation, and implications—along with domain description as an additional inference for PROMs. In applying argument-based validation to PROMs, we argue that researchers should routinely include domain descriptions as additional inferences. Specifically, domain description should precede the conventional four inferences (see the grey boxes in Figure 1).

In the following sections, we first describe the domain description and then the four conventional inferences used in PROMs. For each inference, we identify the relevant assumptions and typical sources of backing. We then present a holistic overview of the resulting framework in Table 1 later in this report, including the inferences, their associated claims, assumptions, and supporting evidence.

Figure 1. Inferences in Mixture-CAT PROM for Emotional Well-Being Validation (inferences in grey)



Mixture-CAT PROM for Emotional Well-Being

Domain Description. PROMs research focuses on the development, validation, and use of patient-reported outcome measures to assess health from the patient's perspective. Unlike clinical indicators or laboratory findings, PROMs capture aspects of health that only patients can directly report, such as symptoms, functional status, quality of life, and emotional well-being. Accordingly, this field examines whether PROMs accurately, consistently, and meaningfully measure these constructs across populations and contexts. PROMs research also addresses how such measures can be applied in clinical care, health services evaluation, and research to support more patient-centered decision-making. By incorporating patients' reports of their own experiences, PROMs research helps ensure that evaluations of health and treatment reflect not only biological outcomes but also their practical and personal consequences.

A separate domain description inference is warranted because researchers must first determine whether the selection, design, and administration of the PROM adequately represent the target domain—or, in PROMs terminology, the target construct—well enough to distinguish construct-relevant from construct-irrelevant variance.

Scoring/Evaluation. The scoring or evaluation inference claims that patients' responses are scored using procedures that produce observed scores with the intended properties. In the context of PROMs, this inference focuses on whether the methods used to translate item responses into scale scores reflect current best practices. This includes evaluating whether item functioning is acceptable, whether response

options and scoring models perform as intended, and whether the measure differentiates respondents appropriately across levels of the target construct for the intended purpose. The inference also addresses the uncertainty (measurement error and reliability) that accompanies the scale scores and the appropriateness of the administration procedures used to collect patient responses.

Generalization. The generalization inference claims that the observed scores provide estimates of the expected scores across comparable items, forms, and administrations of the PROM. Backing for this inference examines whether parallel or alternate versions of the measure are equivalent, whether the PROM includes a sufficient number of well-functioning items to support stable score interpretation, and whether scores remain consistent across administrations and relevant measurement conditions. In other words, in this context, backing for the inference examines whether alternate versions of the PROM are equivalent, whether the number and quality of items are sufficient to support reliable generalization, and whether score patterns are consistent across administrations and relevant measurement conditions.

Extrapolation. The extrapolation inference concerns whether the construct operationalized by the PROM adequately represents the target patient-reported outcome domain. In contrast to domain description inference, which is typically conducted before instrument development, extrapolation inference is evaluated after the PROM becomes operational. At that stage, researchers can examine the extent to which PROM scores meaningfully correspond to and generalize across patients' lived experiences, functioning, health status, and health-related quality of life beyond the assessment itself.

Another Look at the "Chain of Inferences" Model (Classic 4-Inference Version)

Because of their nuance and the challenge of applying them in complex assessment or survey settings, there is value in re-visiting the description of Kane's four inferences.

The widely used practical entry point to his argument-based validation framework is Kane's four inferences. One can describe validation as (a) **articulating claims/assumptions** for these inferences and (b) **testing them**, organizing results into a coherent argument.

Inference 1 — Scoring

How do we move from an observation (an answer, a performance, a rater judgment) to a score?

Typical assumptions/evidence:

- scoring rules are appropriate and applied correctly
- raters are trained; rubrics work as intended
- response processes align with the construct (e.g., no construct-irrelevant shortcuts)

Scoring as the first inference: "translating an observation into one or more scores."

Inference 2 — Generalization

Do the observed scores generalize across the conditions we sampled (items, cases, raters, occasions) to a broader "universe" of similar situations? Generalization is using the score(s) as a reflection of item responses to a survey instrument in a survey response setting, wherever a questionnaire is completed.

Inference 3 — Extrapolation

Do scores derived from response to a survey instrument say what you claim they say about **real-world performance** (or the target domain outside the assessment setting)? Extrapolation focuses on "using the score(s) as a reflection of real-world performance."

Inference 4 — Implications

Are the **decisions/actions** you want to take based on scores appropriate, fair, and beneficial (or at least not harmful)? One needs to define implications as "applying the score(s) to inform a decision or action," and emphasize that evidence should be prioritized toward the *most questionable assumptions* in the chain.

Validity argument

The *evaluation* of that IUA—i.e., whether the inferences and assumptions are plausible and sufficiently supported by evidence.

Kane explicitly defends keeping the IUA as a separate scaffold (rather than collapsing everything into one "validity argument") because it helps prevent vague, overstated, or understated claims and keeps validators clear about *what exactly is being asserted and assessed*.

The validity argument guides the collection and interpretation of validity evidence. A useful analogy is that of an orchestral conductor who interprets, assembles, organizes, and presents the music to convey an intended artistic vision. No single instrument carries the performance on its own. Rather, the conductor coordinates multiple sections, each of which contributes a distinct but partial voice to the whole. The quality of the performance depends not only on the strength of the individual musicians, but also on how effectively their contributions are timed, balanced, and integrated. In a similar way, a validity argument does not rest on any single piece of evidence. Instead, it depends on the coordinated interpretation of multiple sources of evidence that, although individually incomplete, collectively support the proposed interpretation and use of scores.

Evidence is Organized Around Assumptions, Not a Checklist of "Types of Validity"

The measurement validation literature emphasizes a shift away from older "types of validity" toward testing the assumptions that underpin the interpretation/use argument and the validity argument.

- Kane's inference-based prioritization is distinct from earlier frameworks, which can feel like a list of evidence categories with no clear order.
- The modern consensus definition (APA, AERA, NCME *Test Standards*) focuses on validity as about interpretations and uses, and on validation as the building and evaluation of arguments for/against them; the Standards' five "sources of validity evidence" are presented as a practical structure for evidence-gathering.

In Kane's framework, those evidence "sources" are often selected because they **support or challenge a particular inference/assumption**—not because a checklist demands them.

A Set of Questions to Help Navigate Kane-style Validation

1. **Name the decision(s)** the assessment will support (low vs high stakes changes what needs evidence).
2. Draft an **IUA**: map the reasoning from observation → score → meaning → real-world claim → decision.
3. For each inference (scoring, generalization, extrapolation, implications), list the **key assumptions** that are most vulnerable or contestable.
4. Gather evidence targeted to those assumptions (often drawing on Standards-style sources of evidence, but guided by what your argument needs).

5. Assemble the **validity argument**: weigh the evidence and conclude how defensible the intended interpretation/use is, given context and consequences.

The Case of the Interpretation/Use Argument (IUA)

Below is a **Kane-style Interpretation/Use Argument (IUA) template** tailored to your **mixture-CAT PROM for emotional well-being** (for adults with chronic illness living at home), written so you can almost drop it into a protocol. It is being framed around the classic **scoring → Generalization → Extrapolation → Implications** chain, with **mixture-CAT–specific assumptions** (latent class membership, tailored item selection, fairness across cultures/life experiences).

In the current section of this note, we present a case of a Kane-style Interpretation/Use Argument (IUA) template tailored to the mixture-CAT PROM measurement system for emotional well-being (for adults with chronic illness living at home), written so you can almost drop it into a protocol.

It is being framed around the classic scoring → Generalization → Extrapolation → Implications chain, with mixture-CAT–specific assumptions (latent class membership, tailored item selection, fairness across cultures/life experiences).

Interpretation/Use Argument for a Mixture-CAT PROM System For Measuring Emotional Well-Being

Proposed interpretation and use (state up front)

Interpretation: A person's mixture-CAT score represents their *self-reported emotional well-being impact on daily life*, as experienced from their own frame of reference, and estimated using a tailored item set matched to their response pattern/background experience.

Use: Scores will be used to (a) support person-centred clinical conversations and care planning, and (b) evaluate outcomes and equity in services across diverse populations.

Target decisions (make concrete):

- Clinical: identify people with clinically meaningful emotional well-being burden; monitor change over time; prompt referral/support.
- System/research: compare outcomes across services or groups *without erasing* culturally different experiences; evaluate equity impacts of interventions.

Scoring Inference

Claim (C1): Observed responses to mixture-CAT items can be transformed into an accurate estimate of the person's emotional well-being impact score **and** (where applicable) their mixture component/class membership.

Warrant (W1): If item content is understandable and relevant, response processes align with intended meaning, the CAT algorithm functions correctly, and the mixture model is correctly specified, then scored estimates reflect the intended construct for that person.

Key assumptions (A1–A8) to test

- **A1 Response process fit:** People interpret each item as intended (including culturally/linguistically), and respond based on their lived experience (not social desirability, fear, or confusion).
- **A2 Content relevance across diversity:** Item pools (large item banks) include concepts that matter across backgrounds; culturally specific expressions of emotional well-being are not systematically omitted.

- **A3 Translation/adaptation integrity** (if multilingual): semantic, conceptual, and experiential equivalence.
- **A4 Algorithm correctness:** item selection, scoring, stopping rules, and exposure constraints are implemented correctly in the online tool.
- **A5 Mixture components are meaningful:** latent classes represent substantively interpretable response patterns (not artifacts).
- **A6 Classification accuracy:** assignment to mixture component is sufficiently accurate and stable for the intended use.
- **A7 Local independence/model fit:** within each component, item response model assumptions are reasonably met (or violations are understood/mitigated).
- **A8 Accessibility/usability:** interface works for varying literacy levels, disabilities, and tech access.

Evidence plan (examples)

- Cognitive interviewing / think-aloud with diverse participants; response process interviews after CAT sessions.
- Co-design workshops with patients/providers to vet meaning and missing concepts.
- Usability testing (mobile/desktop; low literacy; assistive tech).
- Model diagnostics: item fit, posterior predictive checks, and misfit sensitivity analyses.
- Classification checks: entropy/posterior probabilities; stability across repeated administrations.
- DIF and fairness checks *within and across* mixture components (cultural groups, language, life experience).

Generalization Inference

Claim (C2): The mixture-CAT estimate generalizes beyond the particular administered items to the person's standing on the intended emotional well-being construct **within the defined measurement universe** (i.e., the item bank and modeled populations/components).

Warrant (W2): If the item bank adequately samples the construct for each component, CAT precision is adequate, and scores are stable when the underlying state is stable, then the score generalizes.

Key assumptions

- **A9 Adequate precision where needed:** Standard errors are sufficiently small at decision points (screening thresholds, monitoring change).
- **A10 Bank coverage for each component:** each subgroup/component has enough informative items across the trait range.
- **A11 Reliability under CAT:** test or instrument information and conditional reliability are adequate across trait levels and across components.
- **A12 Temporal stability:** short-term test–retest stability when no real change is expected.
- **A13 Measurement invariance across administrations:** changes in item selection do not create artificial change.

Evidence plan

- Conditional SEM/Information curves by component and by key demographic/cultural strata.
- Test–retest study; mode effects (phone vs computer).
- Simulation studies comparing fixed-form vs mixture-CAT under known parameters.
- Evaluate whether stopping rules produce inequitable precision (e.g., more uncertainty for some groups).

Extrapolation inference

Claim (C3): Mixture-CAT scores correspond to **real-life emotional well-being impact on daily functioning**, as experienced by the person, and relate to external indicators in expected ways **without privileging majority-norm expressions**.

Warrant (W3): If the construct is well-represented, and the score captures lived experience rather than artifacts of culture/language/response style, then it should show coherent relationships with relevant external measures and outcomes.

Key assumptions

- **A14 Construct representation is equitable:** the tool captures diverse manifestations (e.g., idioms of distress, culturally shaped coping norms).
- **A15 Convergent relationships:** expected associations with related constructs (e.g., depression/anxiety scales, quality of life, fatigue, pain interference) hold across groups/components.
- **A16 Discriminant relationships:** weaker associations with unrelated constructs (e.g., unrelated physical measures) as predicted.
- **A17 Known-groups validity:** scores differentiate groups expected to differ (e.g., those reporting major stressors, recent hospitalization) *within* each cultural subgroup.
- **A18 Responsiveness:** scores change when meaningful change occurs (therapy initiation, pain flare resolution), similarly across groups.

Evidence plan

- Given the mixed-method triangulation: quantitative associations + qualitative: "Does this score reflect your life?"
- Multigroup/mixture structural models to test whether relationships differ by component or culture—and whether differences reflect real-world meaning vs bias.
- Anchoring vignettes or response-style analyses where relevant (acquiescence/extreme responding).

Implications inference

Claim (C4): Using mixture-CAT scores leads to **better, more equitable decisions** and outcomes than "one-size-fits-all" PROMs, with acceptable risks and clear communication of uncertainty.

Warrant (W4): If the score is interpretable, thresholds are justified, uncertainty is communicated, and the tool does not systematically disadvantage any group, then its use is appropriate and beneficial.

Key assumptions

- **A19 Interpretability:** clinicians/patients can understand what the score means; score reports are patient-centred and culturally respectful.
- **A20 Actionability:** the score meaningfully informs care planning and shared decision-making.
- **A21 Fairness in decisions:** the same score implies comparable lived impact across groups/components, or differences are explicitly modeled and communicated.
- **A22 Thresholds are justified:** if you use cut scores, they correspond to clinically meaningful states across groups (or are tailored/anchored appropriately).
- **A23 Consequences are monitored:** implementation does not increase inequities (e.g., fewer referrals for certain groups due to misinterpretation).
- **A24 Data governance/trust:** privacy, consent, and community accountability are addressed—especially important for marginalized communities.

Evidence plan

- Implementation pilots: workflow fit, shared decision-making outcomes, patient trust/acceptability.
- Decision studies: Do scores change clinical actions? Are actions appropriate?
- Fairness audits: referral rates, follow-up, outcomes by group/component before vs after implementation.
- Reporting design studies: evaluate comprehension and perceived respect for and fit of score reports.

What to include as "rebuttals" (the built-in "how could we be wrong?" list)

Write these explicitly in the IUA so the validity argument can address them:

- The mixture model "discovers" groups that reflect **response styles** rather than meaningful experience.
- Item pools (large item banks) still under-represent some communities → tailored CAT looks precise but is *precisely wrong*.
- DIF is "absorbed" into components (masking bias).
- Differential access/usability makes some groups more likely to drop out or answer differently.
- Clinical users misinterpret component membership as a label/stigma rather than a measurement aid.

A practical one-page summary structure (what you can put in a grant/protocol)

1. **Intended interpretation/use + decisions**
2. **Inference chain table** (4 rows: scoring/generalization/extrapolation/implications)
3. For each row: **Claim → assumptions → planned evidence**
4. **Equity commitments** (how will you involve communities; fairness auditing plan)
5. **Decision-risk statement** (what you will *not* use the score for until evidence supports it)

The Evidence Burden Differs by Intended Use: Monitoring Change, Shared Decision-Making, and Program Evaluation

Evidence burden differs by intended use, even when the underlying interpretation remains constant. For the mixture-CAT PROM of emotional well-being, all use cases require evidence that respondents across diverse backgrounds interpret the items as intended, that the item bank equitably represents relevant lived experience, and that the CAT algorithm functions accurately and accessibly. However, the strongest additional requirements vary by application. **Monitoring change** places the greatest weight on longitudinal properties—adequate conditional precision at the individual level, stability when no true change is expected, responsiveness when change occurs, and decision rules that incorporate measurement uncertainty to avoid overinterpreting noise. **Shared decision-making** prioritizes interpretability and acceptability: patients and clinicians must understand score reports and uncertainty, recognize the content as relevant and respectful, and use the results to support meaningful conversation without stigmatizing "labels" (including mixture-class information, if reported). **Program evaluation** imposes the highest burden for comparability and fairness: cross-site standardization, sufficient precision for subgroup and intersectional estimates, explicit handling of missingness and differential attrition, and rigorous evaluation of differential item functioning/invariance so that between-group differences do not reflect construct-irrelevant bias. **Across all three applications**, we will monitor consequences of use—including burden, access barriers, and potential perverse incentives—through equity-focused audits and stakeholder-informed governance to ensure that score use advances, rather than undermines, equitable care.

Table 1 presents the elements of an interpretation/use argument for the mixture-CAT PROM system, where the intended use of the resulting score is for program evaluation (quality improvement, outcomes, equity).

Table 1. Interpretation/Use argument for mixture-CAT PROM: Program evaluation (quality improvement, outcomes, equity)

Inference	*Claim*	*Assumptions (to be tested)*	*Evidence/Analyses*	*Risks/Mitigations*
Scoring	Computed scores comparably across sites, time, and populations, enabling aggregation for program-level analyses.	Standardized administration and scoring across settings; minimal site-specific artifacts; mixture components behave similarly across sites, or differences are modeled; data quality is adequate (missingness, mode effects).	Cross-site implementation QA; audit administration conditions; software/algorithm version control; missing-data pattern analysis; mode effects testing; mixture invariance checks across sites.	**Risk:** site differences reflect administration/mode rather than program impact. **Mitigation:** harmonize administration; adjust for mode; include site random effects; pre-specify QA thresholds and exclusion rules.
Generalization	Group-level estimates (means, distributions, change) are reliable and not driven by differential precision or sampling differences across subgroups.	Sufficient precision for subgroup estimates; measurement error and differential precision are accounted for; sample sizes support subgroup and intersectional analyses; attrition/missingness does not bias group comparisons.	Evaluate conditional precision by subgroup/component; weighting or measurement-error–aware models; power analyses for equity strata; sensitivity analyses for missing-not-at-random; assess differential attrition by group and baseline score.	**Risk:** equity conclusions are distorted by small-N or differential missingness. **Mitigation:** planned oversampling; combine qualitative equity monitoring; transparent uncertainty intervals; pre-registered analysis plans; interpret with cautious claims when precision is limited.

Inference	*Claim*	*Assumptions (to be tested)*	*Evidence/Analyses*	*Risks/Mitigations*
Extrapolation	Program-level differences and changes in scores reflect meaningful differences in lived emotional well-being impact, not construct-irrelevant bias.	The construct meaning is stable enough across groups for the intended comparisons; any group differences reflect true experience rather than bias; relationships with external indicators are consistent across groups/sites.	Multigroup/mixture models testing invariance and differential item functioning; external validity checks with related outcomes (service use, functioning, patient narratives); triangulate with qualitative program feedback and patient experience data.	**Risk:** mixture model "absorbs" bias into components, masking inequity. **Mitigation:** explicit DIF testing alongside mixture modeling; fairness-focused diagnostics; report both overall and group-/component-specific results.
Implications	Using scores for program evaluation improves quality and equity without stigmatizing communities or creating perverse incentives.	Stakeholders interpret results appropriately; reporting does not penalize programs serving higher-need groups; governance ensures respectful and safe use; results lead to actionable improvements.	Stakeholder interpretation studies; equity-focused reporting standards (stratified reporting + context); monitoring for unintended consequences (gaming, avoidance); participatory governance with community oversight; evaluation of QI actions taken and outcomes.	**Risk:** ranking/benchmarking unfairly disadvantages programs serving marginalized groups. **Mitigation:** emphasize improvement over ranking; risk adjustment where appropriate; contextualized dashboards; guardrails for use; co-developed dissemination plans with communities.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52(4)*, 281–302. https://doi.org/10.1037/h0040957

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum Associates.

Gadermann, A. M., Petteni, M. G., Molyneux, T. M., Warren, M. T., Thomson, K. C., Schonert-Reichl, K. A., Guhn, M., & Oberle, E. (2023). Teacher mental health and workplace well-being in a global crisis: Learning from the challenges and supports identified by teachers one year into the COVID-19 pandemic in British Columbia, Canada. *PLOS ONE, 18*(8), e0290230. https://doi.org/10.1371/journal.pone.0290230

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement, 63*(1), 65–74. https://doi.org/10.1177/0013164402239317

Giraldo-Rodríguez, L., & López-Ortega, M. (2024). Validation of the Short-Form 36 Health Survey (SF-36) for use in Mexican older persons. *Applied Research in Quality of Life, 19*(1), 269–292. https://doi.org/10.1007/s11482-023-10240-6

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6(4)*, 427-438. https://doi.org/10.1177/001316446600600401

Hubley, A. M., Ruddell, R. J., & Ma Zhu, S. (2024). Cracks, gaps, and holes in validation practice as evidenced from a validation synthesis of the English version of the Rosenberg Self-Esteem Scale. *European Journal of Psychological Assessment, 40*(6), 529–547. https://doi.org/10.1027/1015-5759/a000877

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123(3)*, 207–215. https://doi.org/10.1080/00221309.1996.9921273

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135–170. https://doi.org/10.1207/s15366359mea0203_1

Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in *Educational and Psychological Measurement*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 91–111). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_6

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405–450. https://doi.org/10.3102/0091732X019001405

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research, 92(3)*, 489–496. https://doi.org/10.1007/s11205-008-9303-y

Treanor, C., & Donnelly, M. (2015). A methodological review of the Short Form Health Survey 36 (SF-36) and its derivatives among breast cancer survivors. *Quality of Life Research, 24*(2), 339–362. https://doi.org/10.1007/s11136-014-0785-6

Zumbo, B. D. (2023). A dialectic on validity: Explanation-focused and the many ways of being human. *International Journal of Assessment Tools in Education, 10*(Special Issue), 1–96. https://doi.org/10.21449/ijate.1406304

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9