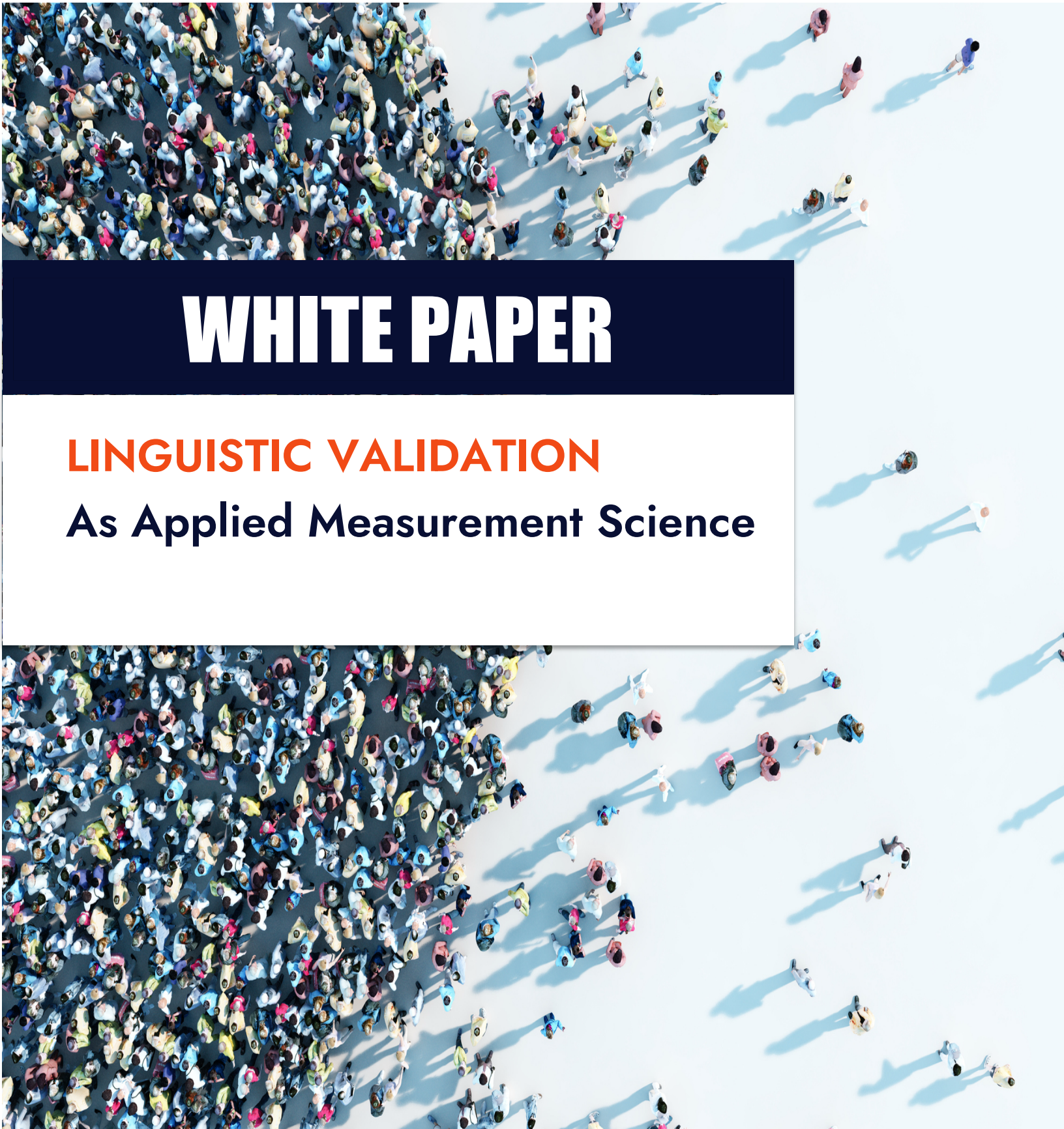
SANTIUM



WHITE PAPER

LINGUISTIC VALIDATION

As Applied Measurement Science

WHITE PAPER: TRANSLATING CLINICAL OUTCOME ASSESSMENTS

February 2026

LINGUISTIC VALIDATION AS APPLIED MEASUREMENT SCIENCE

SANTIUM

Table of Contents

INTRODUCTION

Linguistic validation is the process used to localize clinical outcome assessments for use across languages. It is an iterative translation process designed for high stakes clinical content. Clinical outcome measures are statistically sensitive instruments, and their wording carries measurement consequences. Every phrase is part of how the instrument functions.

Adapting a clinical outcome assessment is inherently multidisciplinary. It draws on clinical knowledge, therapeutic-area expertise, linguistics, cross-cultural psychology, regulatory expectations, and measurement science, whether described as psychometrics or clinimetrics. No single professional background spans all these domains, yet successful adaptation requires understanding the construct being measured, how it is expressed clinically, how meaning is structured in the target language, how responses are shaped by cultural context, and how the resulting data will be evaluated.

Clinical outcome assessments are typically grouped into four categories: patient-reported outcomes (PROs), clinician-reported outcomes (ClinROs), observer-reported outcomes (ObsROs), and performance outcomes (PerfOs). Each operates in a different communicative register, but all share the same requirement: the language of the assessment must reliably elicit the clinically relevant information the instrument was designed to measure, whether symptoms, functioning, cognition, or quality of life. This is fundamentally a measurement task.

Producing a linguistically validated version therefore requires that everyone involved understands what each item is intended to measure and that this intent is communicated clearly from the sponsor and instrument developer through to the translation team. When that shared understanding is weak, the strain often becomes visible later, particularly at the site level where the translated instrument must work in real interactions between patients and clinicians.

WHAT COAs ARE DESIGNED TO MEASURE

A well-developed clinical outcome measure is not just a questionnaire with medical or statistical terminology. Each item is statistically and clinically designed to capture a specific aspect of a defined construct. Its wording intends to guide respondents toward a particular interpretation and respond to it. It carries statistical weight, tied to normative analyses of its target populations within a pre-defined clinical, functional, demographic, cultural and/or linguistic context. Because of this, the respondent's understanding is central to validity. In some settings, an item may require modification when a direct

rendering would not be culturally or contextually meaningful. These changes are carefully made to preserve the intent behind the question and ensure that the same concept is being measured at the same level of difficulty, even if the local expression or context differs.

Collectively, the items function as a measurement model, typically developed using Classical Test Theory or Item Response Theory frameworks. In some cases, Rasch modeling, a specific form of IRT, is applied to evaluate or refine the scale's measurement properties. The items in the instrument, their response options, their recall periods, basal and ceiling (or discontinue) rules, and the precise language connecting them have been selected and validated through a rigorous development process. And usually only in one language. It is a process intended to ensure that the instrument produces scores that are reliable, valid, and sensitive to (clinically) meaningful change within a specific target population.

When these instruments begin to move beyond their original settings, whether into another therapeutic area, disciplinary context, or language market, work is required to ensure that they continue to function as intended. Different disciplines approach this transition through their own lens. They tend to focus on the forms of equivalence most familiar to their field, without always seeing how and if those perspectives align with the others.

Linguistic validation sits at the intersection of these viewpoints. Its purpose is to ensure that the measurement established during instrument development remains intact as the assessment is implemented across languages and cultures. The following framework outlines how this objective is designed to be carried out in practice.

> **Confidence in the data generated by a clinical outcome measure in any language depends on the traceability of the process used to create that version.**
>
> — ***A collective agreement from regulatory body guidances***,
> FDA, EMEA, NICE, HAS and IGWiG

THE METHOD: FROM EDUCATION TO CLINICAL TRIALS

The methodological foundations of linguistic validation did not originate within pharmaceutical development. They emerged from educational and psychological testing, where cross-language equivalence has long been treated as a measurement problem rather than a linguistic one.

Since 1954, *The Standards for Educational and Psychological Testing* continue to emphasize that when an instrument is adapted for use in another language or culture, the goal is to preserve the construct being measured and support equivalent interpretation of scores, even if modifications to wording or format are required (AERA, APA, & NCME, 2014). Researchers examining intelligence and achievement testing across cultures recognized that direct translation was insufficient to preserve meaningful scores.

Over the following decades, formal guidance on test adaptation developed within psychometrics and cross-cultural psychology, culminating in structured frameworks for translation, documentation, and evidence of equivalence (Hambleton et al., 2005; International Test Commission, 2017). These frameworks emphasized conceptual equivalence, response behavior, and construct comparability across groups.

As multinational clinical trials expanded in the 1990s and early 2000s, outcomes research adopted some of these principles. The now-familiar linguistic validation workflow, including independent forward translations, reconciliation, back translation, and cognitive debriefing, reflects this transfer of methodology into the health research context (Wild et al., 2005).

What emerged was an applied measurement framework designed to protect construct continuity as instruments crossed linguistic and cultural boundaries.

REGULATORY PERSPECTIVE

Regulatory agencies do not prescribe a specific adaptation methodology. Instead, they expect sponsors to demonstrate that an instrument used in multiple languages is reliable, valid, and appropriate for the populations studied.

The U.S. Food and Drug Administration guidance emphasizes evidence of content validity and documentation to support interpretation of clinical outcome assessment data, without mandating a particular translation process (U.S. Food and Drug Administration, 2009). European guidance similarly focuses on the reliability and interpretability of outcome measures in multinational research rather than specifying methods (European Medicines Agency, 2005). Regulatory expectations in Asia follow the same principle, requiring justification that the measure is meaningful in the target population (Pharmaceuticals and Medical Devices Agency, 2018). The procedures associated with linguistic validation are structured approaches used to help support this evidence in a consistent manner.

TRANSFERING MEASUREMENT TO A NEW LANGUAGE

The linguistic validation process typically includes forward translation by independent translators working from the source language, reconciliation of those translations, back translation into the source language by translators without access to the original, review of the back translation to explore

potential shifts in meaning, and cognitive debriefing through structured interviews with members of the target population to determine whether the adapted items are understood as intended (Wild et al., 2005).

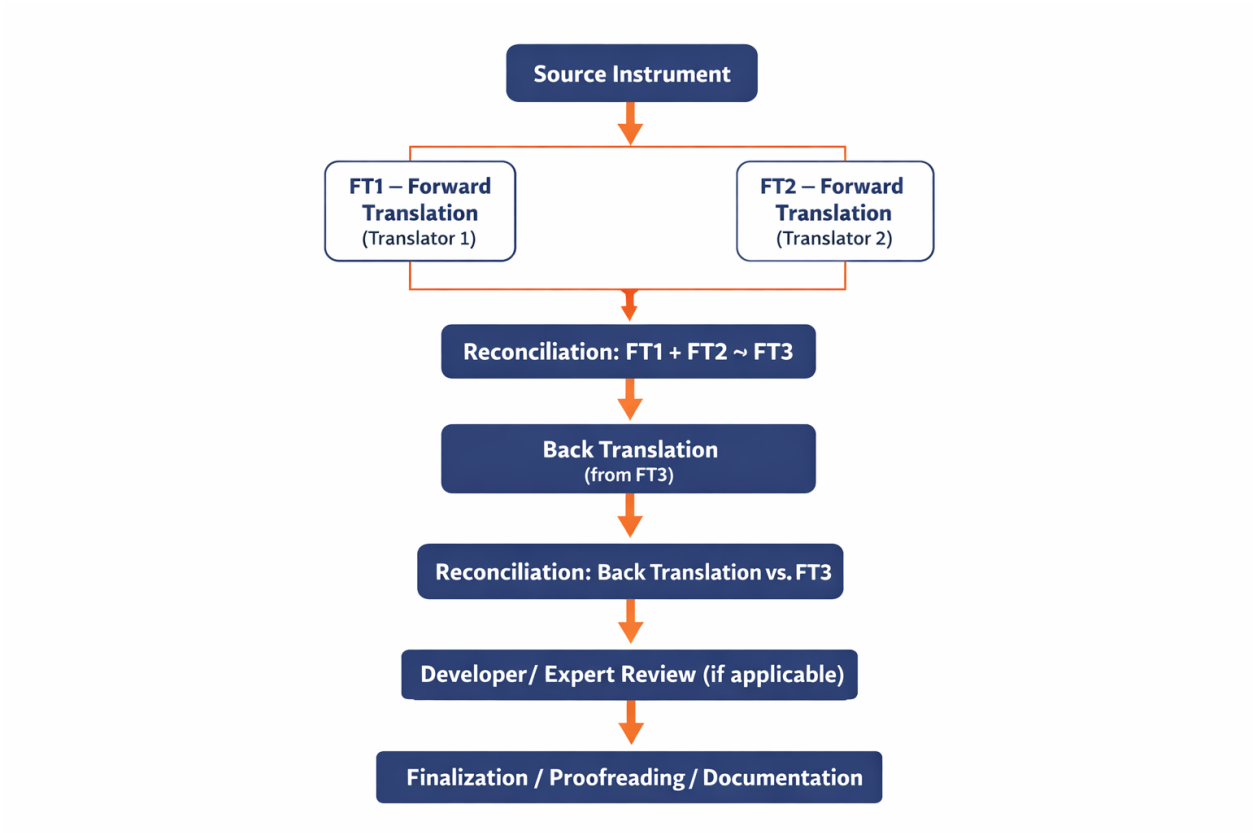


Figure 1. Standard Linguistic Validation Methodology

Each stage generates specific information about how the construct functions in the new context. Independent forward translations provide alternative representations of the same concept, allowing reconciliation to examine how meaning can be expressed within the linguistic system. Back translation serves as a diagnostic probe for divergences in interpretation. It is most useful without stylistic interference. Its value lies in revealing how the translated version resolves ambiguity and encodes meaning when rendered back into the source language. Cognitive debriefing examines whether respondents can access and interpret the experience or behavior the instrument is designed to measure within their own cultural and linguistic frame.

BRIDGING MEASUREMENT AND IMPLEMENTATION

A central distinction running through this work is that between *semantic equivalence* and *conceptual equivalence*. A translation may accurately reproduce the words of the source instrument but fail to function if the underlying construct does not operate in the same way in its new linguistic or cultural context (Herdman, Fox-Rushby, & Badia, 1998). Concepts such as fatigue, pain, or emotional functioning are expressed, interpreted, and socially contextualized differently across populations, and these differences can influence how respondents understand and answer assessment items (Patrick et al., 2011). Ensuring that the adapted instrument measures the same phenomenon, rather than simply using corresponding terminology, is therefore the central objective of COA adaptation.

Clinical Development Perspective

Procedural frameworks developed within clinical development and outcomes research have tended to emphasize consistency in how text is translated and documented. This emphasis has helped standardize multinational trial practices and improve traceability of decisions. At the same time, it can create an operational preference for maintaining close alignment with the source wording, even in situations where more substantive adaptation would better preserve the construct.

Psychometric Perspective

In contrast, the conventions of educational and clinical psychological testing have long recognized that item content may need to be modified when constructs are expressed differently across populations, provided that the adaptation is theoretically justified and supported by evidence of equivalence (Hambleton et al., 2005; AERA, APA, & NCME, 2014; International Test Commission, 2017). The primary concern in this literature is comparability of measurement rather than similarity of phrasing.

Translation Perspective

Translators are trained to value precision, and in many domains a high-quality translation is defined by linguistic equivalence, conveying the meaning of the source text accurately and naturally in the target language. This orientation is appropriate for most forms of content. In the context of clinical outcome assessments, however, the objective extends beyond linguistic accuracy to include measurement equivalence.

Considerations such as whether a phrasing change might alter how respondents interpret a question or select a response are not typically part of general translation training. In some research settings, introductory training on linguistic validation is provided to translators and project teams. This proprietary training tends to emphasize close alignment with the source text. While this reinforces

consistency and documentation, it may also limit confidence in making adaptations that would better preserve the construct in the target context.

In most workflows, translators engage with the instrument primarily as text, working from briefs, item definitions, or excerpts from technical manuals that do not fully convey the instrument's measurement logic. As a result, they are not typically positioned to evaluate construct-level considerations, which remain the responsibility of instrument developers, psychometricians, and the validation process (Wild et al., 2005; AERA, APA, & NCME, 2014).

Bringing the Perspectives Together

This separation of roles can create a gap between the measurement intent established during instrument development and the linguistic decisions made during adaptation, underscoring the need for structured collaboration across disciplines (Hambleton et al., 2005; International Test Commission, 2017).

Clinical development has historically emphasized procedural reproducibility, the testing literature construct comparability, and translation practice linguistic precision. Linguistic validation is most effective when these perspectives are brought together to support measurement continuity across languages and populations.

FUNCTIONAL LIMITS OF LINGUISTIC PRECISION

The implications of the gap in perspectives become most visible in how respondents interact with specific design features of the instrument. These features directly affect how the measure functions but are rarely examined explicitly during linguistic validation because at face value they appear to be correct, and equivalent.

Response Scales as Measurement Units

Response scale anchors are one of the most fragile elements in translating clinical outcome measures. In a Likert-type response format, anchors such as "not at all," "slightly," "moderately," "quite a bit," and "extremely" operationalize the response continuum from which scores are derived. How respondents perceive the distances between adjacent anchors can affect response distributions and the interpretability of statistical analyses (Likert, 1932; Streiner, Norman, & Cairney, 2015).

Anchor translation is not only a lexical task because it also requires identifying terms in the target language that function at comparable points along the same intensity or frequency continuum. Languages are more specific or categorical in the density and distribution of such expressions, and differences in response category functioning across cultural groups have been well documented in cross-cultural measurement research (Hambleton, Merenda, & Spielberger, 2005; van de Vijver & Leung, 1997). Selecting among alternative anchor terms can influence the effective measurement

range of the instrument, potentially compressing or expanding responses or shifting thresholds between categories without producing any obvious textual error.

Timeframes: Recall Periods and Test–Retest Logic

Test–retest reliability is established during instrument development using a specific interval between administrations chosen to balance two risks: respondents remembering their earlier answers and the underlying construct genuinely changing. The selected timeframe reflects assumptions about how stable the measured concept is expected to be over time and is part of the evidence supporting the instrument's reliability (AERA, APA, & NCME, 2014; Streiner, Norman, & Cairney, 2015).

In applied research settings, assessment schedules are sometimes adjusted to align with study visits or operational constraints. When these intervals differ from those used during validation, it becomes important to consider whether the measure is sufficiently stable or sensitive for the new timeframe. If the interval is too short, responses may reflect recall rather than true status. If it is too long, real clinical change may occur, making it difficult to distinguish measurement consistency from change in the construct.

These changes interact directly with translation because time is encoded differently in other languages. Recall-period phrasing, frequency expressions, and grammatical marking of duration or completion vary across languages. An interval that functions acceptably in one linguistic version may be interpreted differently in another if the temporal framing no longer aligns with how respondents naturally conceptualize that timeframe.

For this reason, adjustments to assessment timing should be considered alongside the translated recall language and response framing to ensure that respondents across languages are referencing comparable periods of experience. Without this alignment, differences in responses may reflect how time is expressed and understood rather than differences in the underlying construct (Hambleton et al., 2005).

Recall periods are another place where small wording choices can affect how the instrument works. When an outcome measure asks respondents to report on their experience "today," "over the past week," or "during the past month," that time frame is not just part of the phrasing. It is a design feature chosen to balance how accurately people can remember their experiences with how much those experiences naturally vary over time. These recall periods are tested during instrument development and are part of what makes the measure valid.

Languages differ in how they express time. Differences in tense, aspect, and available time expressions can make a direct translation sound unnatural, unclear, or slightly misleading. The adapted version must preserve the same reference period and perspective while still sounding natural to respondents completing the assessment. Context is critical.

Item Structure and Construct Sampling

Item structure also carries measurement meaning, not just its wording. The way a question is constructed, including clause order, use of negatives, and specificity of the behavior being asked about, is part of how respondents understand what is being measured. During instrument development, these structural features are tested to ensure that respondents interpret the item consistently and that responses relate to the intended construct (AERA, APA, & NCME, 2014; Willis, 2005).

Changes to item structure during translation can unintentionally alter the respondent's operational interpretation. For example, negatively worded items are often included to control for response patterns such as acquiescence, but research shows that they are cognitively more complex and more prone to misunderstanding if their structure is modified (Swain, Weathers, & Niedrich, 2008; Streiner et al., 2015). Similarly, simplifying or splitting a complex item may appear to improve readability while unknowingly changing how respondents map their experience onto the response scale.

It is also common for instruments to include items that appear to ask similar things in slightly different ways. This redundancy is intentional. Multiple items are used to capture different facets of the same construct and to improve reliability by allowing responses to be evaluated as a pattern rather than in isolation. Altering the structure or emphasis of one of these items during translation can change how it functions within that pattern, even if the wording still appears accurate (Streiner et al., 2015).

Many languages offer more than one grammatically correct way to express the same idea, with differences in emphasis, perspective, or how the timing of the action is framed. Translators are not trained in measurement; hence they tend to regress to focus on stylistic preferences instead of conceptual equivalence, which are aligned with their training. A stylistic change in structure can guide respondents toward slightly different interpretations and deviate from the psychometric purpose of the item.

At the same time, source instruments themselves are not always written in fully formal or unambiguous English; COA items often use conversational or condensed phrasing to match how patients naturally speak, sometimes with regional expressions. These features must be interpreted functionally rather than normalized during translation so that the adapted version preserves how the original item was intended to work (Hambleton et al., 2005). For this reason, preserving functional structure, not just meaning, is necessary to maintain comparability across language versions.

Measurement concerns around scale anchors, recall periods, test–retest assumptions, and item structure can be practically operationalized within existing linguistic validation workflows. By incorporating targeted probes that explore how respondents interpret timeframes, use response options, and process item structure, cognitive interviews can generate evidence not only of comprehension but of functional alignment with the instrument's measurement intent. This does not require new validation studies. It requires connecting respondent feedback explicitly to the construct and design assumptions established during development.

At the same time, these operational considerations intersect with a broader question: how far adaptation can and should extend when constructs are embedded within different cultural contexts. Linguistic alignment alone does not resolve differences in how experiences are conceptualized, expressed, or socially situated. The boundaries between preserving measurement continuity and accommodating cultural variation therefore require careful definition.

COGNITIVE DEBRIEFING: EVALUATING FUNCTIONAL EQUIVALENCE

Translation quality is often evaluated in terms of wording accuracy, while cognitive debriefing confirms that respondents understand items, instructions, and response options. Yet the interviews already reveal more than comprehension. When respondents explain how they interpreted a question, what timeframe they considered, or how they selected a response, they are demonstrating how they operationalize the construct in practice.

Viewed this way, cognitive debriefing needs to be more than a check on clarity. It is an observable test of whether the instrument is functioning as intended in the new linguistic and cultural context. The information needed to assess functional equivalence is therefore already present within the standard linguistic validation workflow, even if it is not always analyzed through a measurement lens.

Cognitive debriefing is therefore well positioned to contribute to evaluating functional equivalence, even though it is not typically structured for that purpose.

CULTURAL ADAPTATION AND ITS LIMITS

Cultural adaptation involves modifying item content when concepts, behaviors, or references in the source instrument are not accessible or appropriate in the target setting. It is a recognized component of localizing clinical outcome assessments and implemented through narrower processes like transadaptation and transcreation. Adaptation differs from translation in that it may require a deliberate departure from the source wording to preserve the construct being measured rather than the literal phrasing (Hambleton et al., 2005; International Test Commission, 2017).

The scope of such changes, however, is constrained by measurement and evidentiary requirements. Once an item is substantively altered, it cannot automatically be assumed to function identically to the source version. When adaptations affect central aspects of the construct rather than peripheral examples or references, additional evidence may be needed to support comparability of interpretation and scores (AERA, APA, & NCME, 2014). Although this level of evaluation has traditionally been viewed as occurring outside linguistic validation, aspects of it can be addressed

within the LV process itself if the focus shifts from confirming comprehension and linguistic interpretation, to examining how respondents understand and operationalize the adapted item. This shift places greater emphasis on how cognitive interviewing is designed and facilitated, so that they elicit information about interpretation, decision processes, and use of response options. For this reason, adaptation decisions in COA development are expected to be documented and supported by empirical observations, typically from cognitive debriefing or related qualitative work (Wild et al., 2005), and moving away from linguistic judgment alone.

These limits become especially important in therapeutic areas where cultural context shapes how symptoms are experienced and expressed. In psychiatry, neurology and neuropsychology, for example, differences across cultures influence not only the language used to describe distress but also how experiences are categorized, which behaviors are recognized as clinically meaningful, and the norms governing whether and how such experiences are reported. Instruments assessing depression, anxiety, psychosis, or cognitive functioning therefore operate within culturally mediated frameworks of meaning. Adapting such measures for new settings requires engagement with clinical and measurement literature combined with linguistic expertise.

DOCUMENTATION: EVIDENCE OF MEASUREMENT INTEGRITY

Regulatory submission requirements continuously evolve, along with technologies and processes implemented within the industry. Submissions that rely on clinical outcome endpoints are expected to include clear documentation of how localized versions were produced and evaluated. FDA guidance on clinical outcome assessments and the Patient-Focused Drug Development initiative emphasize the need for evidence supporting content validity and appropriate interpretation within the target population (U.S. Food and Drug Administration, 2009; U.S. Food and Drug Administration, 2020). European guidance on patient-reported outcomes reflects similar expectations for transparency in development and adaptation processes (European Medicines Agency, 2005). Health technology assessment bodies such as NICE, HAS, and IQWiG likewise consider the quality and interpretability of COA evidence when reviewing outcomes used in reimbursement and access decisions (NICE, 2013; IQWiG, 2020).

These expectations reflect a simple principle: confidence in the data generated by a clinical outcome measure in any language depends on the traceability of the process used to create that version. Cross-cultural measurement standards emphasize that adaptations must be documented sufficiently to demonstrate that score meaning is preserved across populations (AERA, APA, & NCME, 2014; International Test Commission, 2017). An adaptation carried out without systematic records of translation rationale, reconciliation decisions, identified discrepancies, and cognitive interview findings cannot be fully evaluated or defended if questions arise about interpretability or

comparability (Wild et al., 2005). The adaptation record therefore functions as methodological evidence rather than administrative output.

This places particular importance on the briefing provided at the start of a linguistic validation project. Work on COA instruments requires more contextual information than is typical for general medical translation. Those involved need a working understanding of the construct being measured, the population for which the instrument was developed, and the role of specific items within the scale so that adaptations maintain conceptual equivalence rather than surface similarity (Hambleton et al., 2005; Patrick et al., 2011). When this context is absent, decisions that appear linguistically sound may nonetheless alter how respondents understand or use the measure.

LINGUISTIC VALIDATION: AN APPLIED MEASUREMENT DISCIPLINE

Linguistic validation of clinical outcome assessments sits at the intersection of psychometrics, clinical measurement science, linguistics, cross-cultural psychology, regulatory governance, and therapeutic-area expertise. No single professional background encompasses all these domains. Adapting an assessment for use across languages and cultures requires simultaneous attention to the measurement model underlying the instrument, the clinical realities of the condition it evaluates, the linguistic resources of the target language, the cultural setting in which responses are given, and the evidentiary standards applied to the resulting data.

The multi-stage methodology that has developed in this field reflects this convergence of expertise. It is a structured way to apply different forms of knowledge at defined points in the adaptation process so that decisions are transparent, traceable, and open to evaluation. The process functions most effectively as a set of coordinated quality controls designed to identify potential sources of measurement distortion before they impact the workflow and before they enter the data. Within this framework, activities such as cognitive debriefing can continue to serve to confirm comprehension, but also to generate insight into how respondents interpret and operationalize items when those activities are explicitly linked to the instrument's measurement intent.

As clinical research continues to expand globally, the number of languages, settings, and modes of data collection involved in outcome assessment will continue to grow. Maintaining comparability across these contexts depends on collaboration among the disciplines that shape COA development and use, and on ensuring that the knowledge generated during instrument design, particularly its psychometric foundations, is shared with those responsible for its adaptation. The science of measuring clinical outcomes has advanced substantially in recent decades; the methods used to implement those measures across languages must continue to develop alongside it.

ARTIFICIAL INTELLIGENCE IN LINGUISTIC VALIDATION: OPPORTUNITIES AND BOUNDARIES

The rapid development of artificial intelligence has begun to influence nearly all areas of language services and clinical research operations. Translation engines, large language models, automated quality checks, and pattern-recognition systems are increasingly incorporated into workflows that were previously fully human-driven. Linguistic validation is not isolated from these developments. As AI tools become more accessible, questions arise regarding their appropriate role within the adaptation of clinical outcome assessments.

It is important to distinguish between operational efficiency and measurement responsibility. Linguistic validation is designed to preserve construct continuity and support evidentiary traceability. Any integration of AI into this process must therefore be evaluated against its potential impact on measurement integrity rather than solely on efficiency gains.

AI systems can process large volumes of text, identifying terminology inconsistencies, comparing language versions, summarizing qualitative data, and supporting documentation. These capabilities assist in administrative organization, transcript management, discrepancy tracking, and cross-language comparison. In structured contexts, AI can efficiently function as an auxiliary analytical tool.

However, AI systems operate by identifying statistical patterns in language. They do not possess an understanding of the clinical construct, the psychometric model, or the design assumptions embedded in the instrument. They cannot independently evaluate whether a proposed linguistic adjustment alters construct sampling, shifts response thresholds, or affects interpretability within the scale's measurement framework. Their output reflects learned language correlations rather than construct-level reasoning.

In clinical outcome assessments, meaning is not limited to explicit wording. Subtle differences in intensity, emphasis, implied framing, or culturally embedded interpretation may influence how respondents understand and answer an item. AI systems are limited in their ability to detect what is implied rather than directly stated. They process observable linguistic signals but do not reliably interpret contextual nuance in the way that trained measurement professionals can.

These limitations become particularly salient in the context of cognitive debriefing.

Artificial Intelligence and Cognitive Debriefing

Cognitive debriefing involves structured qualitative interviews with members of the target population to explore how items are interpreted and how response options are used. It is important to clarify that cognitive debriefing participants are not clinical trial subjects. They are recruited specifically to evaluate the instrument prior to its use in a study. Their role is methodological rather than endpoint-generating.

AI tools reliably assist in transcription, thematic clustering, summarization of recurring comprehension issues, and structured reporting. In large multinational programs, such support may improve organizational efficiency and consistency in documentation.

Yet the core value of cognitive debriefing lies in interpretive analysis. Respondent explanations often reveal subtle construct shifts, indirect reasoning, hesitation, or culturally mediated reframing. A participant may verbally affirm understanding while simultaneously demonstrating conceptual misalignment. Detecting such patterns requires probing informed by construct theory and psychometric intent.

AI systems can summarize what was said. They do not independently evaluate whether the explanation reflects construct drift, altered response mapping, or culturally influenced reporting behavior. Nor can they replace the facilitation skills required to guide probing in real time.

Ethical and governance considerations must also be addressed. Although cognitive debriefing participants are not part of the clinical trial, interviews frequently involve discussion of health experiences and potentially sensitive information. The use of AI tools introduces considerations regarding data storage, cross-border processing, model training exposure, privacy compliance, and transparency. Clear policies must define how qualitative data are processed and protected.

In addition, AI models are trained on large-scale linguistic datasets that may not represent all cultural contexts equally. In therapeutic areas such as psychiatry or neurology, where symptom expression is culturally shaped, automated interpretation may oversimplify meaningful variation.

For these reasons, AI in linguistic validation is best understood as a supportive technology rather than a methodological substitute. It may enhance organization, comparison, and documentation. It does not replace clinical judgment, psychometric expertise, or culturally informed interpretation.

Linguistic validation remains an applied measurement discipline. Its primary responsibility is to preserve construct integrity before distortion enters clinical data. Any use of artificial intelligence within this process must be evaluated according to that standard.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

European Medicines Agency. (2005). *Reflection paper on the regulatory guidance for the use of health-related quality of life measures in clinical trials*.

Hambleton, R. K. (2001). *The next generation of the ITC test translation and adaptation guidelines*. European Journal of Psychological Assessment, 17(3), 164–172.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum.

Herdman, M., Fox-Rushby, J., & Badia, X. (1998). *A model of equivalence in the cultural adaptation of HRQoL instruments*. Quality of Life Research, 7, 323–335.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests (Second edition)*.

IQWiG. (2020). *General methods (Version 6.0)*.

Likert, R. (1932). *A technique for the measurement of attitudes*. Archives of Psychology, 140, 1–55.

NICE. (2013). *Guide to the methods of technology appraisal*.

Patrick, D. L., Burke, L. B., Gwaltney, C. J., et al. (2011). *Content validity—Establishing and reporting the evidence*. Value in Health, 14(8), 967–977.

Pharmaceuticals and Medical Devices Agency. (2018). *Guideline on patient-reported outcome measures*.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use (5th ed.)*. Oxford University Press.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). *Assessing three sources of misresponse to reversed Likert items*. Journal of Marketing Research, 45(1), 116–131.

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Sage.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage.

Wild, D., Grove, A., Martin, M., et al. (2005). *Principles of good practice for the translation and cultural adaptation process for PRO measures*. Value in Health, 8(2), 94–104.

U.S. Food and Drug Administration. (2009). *Guidance for industry: Patient-reported outcome measures*.

U.S. Food and Drug Administration. (2020). *Patient-focused drug development guidance series*.

Translating Clinical Outcome Assessments:
Linguistic Validation as Applied
Measurement Science
Author: Monika Vance
February 2026

Santium Media Corporation
13-3120 Rutherford Road
Suite 339
Toronto, ON L4K 0B2
Canada

Worldwide Inquiries:
Phone: +1.833.726.8486
Email: info@santium.com

santium.com