The Measurement Reliability Crisis in Event-Related Potential Research: Evidence from Bilingual Language Control

Projnya Mojumdar[1] & Purba Mojumdar[2]

[1] Department of English, University of Creative Technology Chittagong (UCTC), Bangladesh

[2] East Delta University, Bangladesh

Corresponding Author: Projnya Mojumdar (projnya@uctc.edu.bd)

ORCID: 0009-0002-5988-3031

Preprint Date: November 1, 2025

Submitted to: Meta-Psychology

Materials: https://osf.io/q2dkj/

Abstract

Event-related potential (ERP) research exhibits a critical methodological blind spot: inter-condition correlations (ρ_{XY}), the parameter that mathematically determines difference wave reliability via the Lord-Novick formula, are systematically unmeasured. This creates fundamental interpretive ambiguity because Classical Test Theory demonstrates that difference wave reliability is inversely related to ρ_{XY}, with high correlations suppressing reliability even when constituent waveforms are psychometrically sound. Systematic examination of four foundational ERP studies in bilingual language switching (2001-2016, $N_{citations} \approx 1{,}575$) confirms this pattern: none report difference wave reliability or inter-condition correlations, despite these measures being the primary dependent variables isolating language control processes. Sensitivity analysis reveals that this vulnerability extends across the entire plausible parameter space: even when inter-condition correlations are moderate ($\rho_{XY} = .50$-$.60$), achieving adequate difference wave reliability ($\rho_{DD'} \geq .70$) requires constituent reliabilities approaching the ceiling of current ERP methodology ($\rho = .85$-$.88$). When experimental control produces high correlations ($\rho_{XY} \geq .70$, a plausible inference from structural analysis and cross-domain precedent in fMRI and other ERP components, though empirically unverified in language switching), the degradation becomes catastrophic: constituent reliabilities of .85 yield difference wave reliabilities near

zero. The conspicuous absence of empirical ρ_XY measurements and reliability verification means the field cannot determine whether two decades of N2 debate reflect genuine neurocognitive complexity, measurement instability, or both in unknown proportions. If inter-condition correlations prove high in tightly controlled paradigms (as rigorous experimental design would predict), inconsistent findings become expected outcomes of unreliable measurement rather than theoretical puzzles requiring reconciliation. However, problems persist even if correlations are moderate, indicating systematic vulnerability regardless of precise parameter values. Three reforms can address this gap: mandatory reporting of difference wave reliability and inter-condition correlations (ρ ≥ .70 for group comparisons, ρ ≥ .80 for individual differences research), adoption of Generalizability Theory frameworks that partition variance across multiple sources, and single-trial analytical methods that avoid categorical subtraction. These principles extend beyond bilingual research to all ERP domains using difference waves. Systematic psychometric verification should precede theoretical elaboration; the measurement revolution Parsons (2022) called for requires not novel techniques but rigorous application of established principles.

Keywords: event-related potentials, psychometric reliability, difference waves, bilingual language control, N2 component, measurement crisis, classical test theory, psychometrics

1. Introduction

The replication crisis in psychological and neuroscientific research has prompted intense scrutiny of methodological practices, statistical analyses, and reporting standards (Open Science Collaboration, 2015). However, a more fundamental issue has received insufficient attention: the psychometric properties of neural measurements themselves. As Parsons (2022, MP.2020.2577) argued in Meta-Psychology, "Measurement matters and I call on readers to help us move from what could be a measurement crisis towards a measurement revolution." Event-related potential (ERP) research, despite its precision in temporal resolution and widespread adoption in cognitive neuroscience, has proceeded for decades with minimal attention to a bedrock principle of measurement science: the systematic assessment and reporting of reliability.

This oversight is particularly consequential because ERP research relies heavily on difference waves, calculated by subtracting one averaged waveform from another to isolate neural activity specific to a cognitive process. While conceptually elegant, this method harbors a critical vulnerability: the reliability of a difference score is not simply inherited from its constituent parts. Classical Test Theory (CTT) demonstrates that difference wave reliability is inversely related to the correlation between conditions, a correlation that paradoxically increases when researchers implement rigorous experimental control (Lord & Novick, 1968, pp. 66-67). This creates a counterintuitive situation where exemplary experimental design can produce psychometrically unstable measures.

The two-decade debate surrounding the N2 component in bilingual language switching serves as a powerful case study for understanding how measurement issues can masquerade as theoretical problems. Foundational studies by Jackson et al. (2001) and Christoffels et al. (2007) reported differing N2 patterns, launching extensive theoretical efforts to reconcile these findings through models of inhibitory control, conflict monitoring, and adaptive language control (Green & Abutalebi, 2013). However, none of these foundational studies reported the psychometric reliability of their N2 difference waves, leaving open the possibility that the apparent inconsistencies reflect measurement instability rather than neural complexity.

This paper presents a rigorous theoretical analysis of the measurement reliability problem in ERP research, demonstrates its manifestation through systematic examination of representative studies from bilingual language control, and proposes a comprehensive framework for methodological reform. While the analysis focuses on the N2 component in language switching, the principles apply broadly across ERP research domains, suggesting that many unresolved contradictions in the literature may stem from unexamined psychometric foundations.

2. Theoretical Foundations: The Psychometric Imperative in ERP Research

2.1 The signal-to-noise problem and the centrality of reliability

Event-related potential research confronts a fundamental biophysical challenge: isolating minuscule time-locked neural signals from substantially larger ongoing electroencephalographic (EEG) background activity. As Luck (2014, p. 11) notes, "The ERPs are quite small, typically ranging from less than 1 μV to 10 μV, whereas the ongoing EEG activity typically ranges from -100 μV to +100 μV." This signal-to-noise challenge is compounded by residual variability even after averaging. The conventional solution involves trial averaging, which theoretically causes random, non-time-locked background activity to approach zero while preserving consistent, time-locked signals. Signal-to-noise ratio improves with the square root of trial count, providing clear rationale for high-trial-count experimental designs (Luck, 2014).

However, trial averaging does not eliminate all sources of uncontrolled variability. Residual noise in averaged waveforms introduces measurement error that propagates through statistical analyses, attenuating effect sizes and reducing statistical power. This reality elevates the importance of psychometric properties, particularly reliability (a measure's ability to consistently quantify the same construct across different data subsets). Throughout this paper, we use "reliability" to refer primarily to internal consistency (consistency across trials within a session, assessed via split-half reliability, Cronbach's α, or generalizability coefficients), which addresses whether trial averaging has adequately reduced random noise. This differs from test-retest reliability (consistency across separate testing

sessions), which assesses temporal stability but is less commonly examined in ERP research.

Reliability is not optional or supplementary; it represents a prerequisite for valid scientific inference. An unreliable measure, dominated by random error, cannot validly index the cognitive or neural process it purports to capture. Psychometric theory establishes that validity cannot exceed the square root of reliability, placing a mathematical ceiling on interpretability (Nunnally & Bernstein, 1994). Without established reliability, any reported effect remains fundamentally ambiguous, potentially representing stable neural signal or measurement artifact. This principle constitutes the foundation of sound measurement, and its systematic neglect threatens the integrity of scientific conclusions.

2.2 The critical vulnerability: mathematics of difference score reliability

The psychometric challenge intensifies due to ERP research's predominant reliance on difference waves. To isolate neural activity associated with specific cognitive processes (e.g., language switching), researchers subtract averaged ERPs from matched control conditions (e.g., language repetition) from experimental condition waveforms. This subtraction aims to cancel shared neural activity, preserving only process-specific components. While conceptually sound, this approach introduces severe psychometric vulnerability.

Difference score reliability does not simply inherit from constituent reliabilities. Classical Test Theory provides the formula for difference score reliability ($\rho_{DD'}$) when constituent variances are equal. As articulated by Lord and Novick (1968, pp. 66-67) and later discussed in the context of change measurement by Cronbach and Furby (1970), we have the general form shown in Equation 1:

Equation 1. General form

where $\rho_{XX'}$ and $\rho_{YY'}$ represent reliabilities of conditions X and Y, and ρ_{XY} represents their correlation. When constituent reliabilities are equal ($\rho_{XX'} = \rho_{YY'} = \rho$), Equation 1 simplifies algebraically. Substituting ρ for both $\rho_{XX'}$ and $\rho_{YY'}$:

Factor out common terms:

Cancel the factor of 2:

Equation 2. Simplified form for equal variance case

Critical assumption: Equation 2 applies when constituent conditions have equal variances ($\sigma^2_X = \sigma^2_Y$) in addition to equal reliabilities. This assumption is typically satisfied in well-designed ERP studies where conditions have similar trial counts, undergo identical preprocessing, and use the same electrode sites. When variances differ substantially, for example, when comparing conditions with unequal trial numbers or markedly different artifact rejection rates—researchers must use the general formula (Equation 1). The equal-variance assumption holds for the parameter combinations examined in this manuscript, making Equation 2 appropriate for the analyses presented here.

Verification of the equal-variance assumption. Equation 2 applies when constituent conditions have equal variances ($\sigma^2_X = \sigma^2_Y$) in addition to equal reliabilities. This assumption is typically satisfied in well-designed ERP studies where: (1) conditions have similar trial counts after artifact rejection (within ±20%), (2) preprocessing procedures are identical across conditions, (3) the same electrode sites and time windows are analyzed, and (4) experimental manipulations affect mean amplitude without systematically altering variance structure.

We examined trial count reporting in the four foundational studies (Table 3) to assess whether equal-variance assumptions are justified. **Critical limitation:** Original manuscripts do not consistently report condition-specific trial counts after artifact rejection, preventing direct variance

comparison. This represents an additional reporting gap beyond reliability assessment. However, the studies' methodological designs (balanced trial presentation, identical preprocessing pipelines, matched stimulus characteristics) suggest variance heterogeneity was minimal. When substantial variance differences exist (≥20% difference in retained trials, or condition-specific variance ratios >1.5:1), researchers must use the general formula (Equation 1) rather than the simplified form.

Understanding ρ_{XY}: the inter-condition correlation

A critical parameter in Equation 2 requires clarification. The term ρ_{XY} represents the correlation between individuals' ERP amplitudes across the two conditions, specifically, the extent to which individual differences are preserved across experimental manipulations. If a participant shows a large N2 amplitude in switch trials, do they also tend to show a large N2 in non-switch trials? High ρ_{XY} indicates yes (preserved individual differences); low ρ_{XY} indicates individual rankings change across conditions.

Terminological note: Throughout this manuscript, we use ρ (Greek rho) when discussing theoretical reliability coefficients in formulas, and r (italicized Latin) when reporting empirical correlation values from published studies. This distinction aligns with standard psychometric convention.

Rigorous experimental control is expected to produce high ρ_{XY} through a three-step mechanism. First, to isolate specific cognitive processes (e.g., language selection), researchers design conditions differing minimally except for the variable of interest, matching stimuli, timing, motor responses, and attentional demands. Second, this design ensures that most neural processing (visual encoding, lexical access, articulation, constituting the majority of variance based on task structure) remains identical across conditions. Third, because these shared processes dominate total variance, individual differences in these non-target processes are nearly perfectly preserved across conditions, producing high ρ_{XY}. According to Equation 2, high ρ_{XY} mathematically suppresses difference wave reliability. Consequently, standard ERP methodology may be optimized to produce potentially unstable primary measures.

Cross-domain precedent supports this inference. Infantolino et al. (2018; Table 2, p. 149) reported r = .97 for amygdala responses to faces versus shapes in fMRI, a tightly controlled within-subjects contrast analogous to language switching. Cofresí et al. (2022) found r = .85-.88 for P3 amplitudes to alcohol versus non-alcohol beverage cues, again, highly similar stimuli differing on a single dimension. Both studies demonstrate that rigorous experimental control can produce high inter-condition correlations. Clayson, Baldwin, and Larson (2021) directly confirmed in ERP research that "RewP difference scores yielded poor reliability due to the high correlation between the constituent reward and non-reward ERPs,"

providing empirical demonstration of the mathematical relationship we describe.

However, a critical empirical gap exists: despite the mathematical centrality of ρ_{XY} to difference wave reliability, inter-condition correlations are virtually never reported in published ERP research. We found no published studies reporting these correlations for N2 components in any paradigm, let alone language switching specifically. This absence represents not merely incomplete reporting but a systematic blind spot in ERP psychometrics. The inference that language switching paradigms exhibit $\rho_{XY} > .70$ rests on structural reasoning and cross-domain analogy rather than direct measurement. While these sources provide convergent plausibility, empirical verification remains an urgent priority.

Worked example

Suppose switch and non-switch conditions each have internal consistency reliability of .85 (excellent by standard criteria). If individuals' switch trial ERP amplitudes correlate with their non-switch trial ERP amplitudes at $\rho_{XY} = .80$ (meaning people with larger switch-trial N2s tend to have larger non-switch-trial N2s, preserving rank ordering across conditions), difference wave reliability becomes:

Despite excellent constituent reliabilities, the difference wave is essentially unreliable (ρ = .25), far below acceptable thresholds. Note that ρ_{XY} here represents the correlation between individuals' scores across the two conditions, not simply a generic similarity between conditions. This calculation is programmatically verified in the reproducible R script available at OSF, which confirms all worked examples in Table 2.

Formula implications

This formula reveals a critical relationship: difference score reliability decreases as inter-condition correlation increases, even when individual conditions exhibit excellent reliability. As ρ_{XY} approaches 1, difference score reliability approaches zero regardless of constituent reliabilities. This mathematical relationship creates systematic vulnerability in cognitive neuroscience methodology. The widespread failure to compute and report difference wave reliability represents not minor oversight but failure to verify whether rigorous experimental design has inadvertently rendered the dependent variable statistically meaningless. This constitutes a potential crisis embedded in methodological design, a fundamental blind spot extending beyond single components or domains, threatening validity of theoretical claims built upon unexamined difference wave foundations.

2.3 Empirical precedent: the faces-shapes reliability paradox

Empirical evidence from functional magnetic resonance imaging (fMRI) research demonstrates this problem's severity. Infantolino et al. (2018; see Table 2, p. 149; N = 139) examined amygdala activation differences between face and shape stimuli, a robust, highly replicable group-level effect. Individual condition reliabilities were excellent (Spearman-Brown corrected reliability > .94 for both faces and shapes conditions). However, difference score reliability (faces minus shapes) was essentially zero (r = -.06), because amygdala responses to faces and shapes correlated highly across individuals (r = .97).

Note: The original paper does not report CIs for these correlations. They report: "the internal consistency of the activation difference between faces and shapes was nearly zero (SB = −.06)" and "the amygdala response to faces and shapes was highly correlated (r = .97)." No CIs are provided in the source paper itself.

Despite robust group-level effects and excellent individual-condition reliability, the difference score (the measure of theoretical interest) was completely unreliable. This finding powerfully illustrates that experimental robustness does not guarantee measurement reliability.

Recent evidence from ERP research demonstrates that this pattern extends to electrophysiological measurements. Clayson et al. (2021; see Figures 3 and 6), applying Generalizability Theory (G-theory, a framework that partitions variance into multiple sources including persons, trials, and

occasions) to N2 components recorded in a Go/NoGo task, found that internal consistency reached acceptable levels (dependability coefficients ≥ .70 in most conditions, within recommended thresholds for preliminary research), but test-retest reliability was substantially lower, falling below recommended thresholds. In Generalizability Theory terminology, the dependability coefficient serves the same function as the reliability coefficient in Classical Test Theory, while the generalizability coefficient addresses relative ranking of individuals. Critically, person × occasion variance was very large (comparable to between-person variance), illustrating precisely the mechanism by which high inter-condition correlations suppress temporal stability. These findings provide direct empirical demonstration that the theoretical vulnerability described here manifests in N2 measurements, supporting the hypothesis that inconsistent findings in bilingual switching studies may partially reflect measurement instability rather than solely genuine neural variability.

Cofresí et al. (2022; Session 1 N = 210, Session 2 N = 96) examined P3 responses to alcohol and non-alcohol beverage cues, finding excellent internal consistency (Alcohol Cue P3: r = .90, 95% CI [.88, .92]; NADrink Cue P3: r = .91, 95% CI [.89, .93]) and fair test-retest reliability (Alcohol Cue P3: r = .71, 95% CI [.59, .79]; NADrink Cue P3: r = .70, 95% CI [.58, .79]) for individual condition P3s. However, the alcohol cue-specific P3 difference score (Alcohol Cue P3 minus NADrink Cue P3) showed poor internal consistency (r = .37, 95% CI [.25, .48]) and poor test-retest

reliability (r = .20, 95% CI [.02, .37]), despite constituent conditions correlating highly (r = .85 to .88).

This demonstrates the mathematical vulnerability described in Section 2.2: high inter-condition correlations suppress difference wave reliability even when constituent measures are psychometrically sound. Comparable patterns have been documented for ERN difference scores, suggesting that the reliability degradation observed for N2 difference waves represents a general property of ERP methodology rather than a component-specific limitation.

2.4 The reliability paradox and implications for statistical inference

The observed pattern connects to what Hedge et al. (2018) termed the "Reliability Paradox" in the context of test-retest reliability. While Hedge and colleagues focused on temporal stability (consistency across testing sessions), similar mathematical principles apply to internal consistency reliability in ERP difference waves. In both cases, low between-person variance in the measure of interest (whether a behavioral effect across time or an ERP difference wave across trials) produces low reliability coefficients, even when measurement error is minimal.

Experimental tasks producing robust, easily replicable group-level effects become popular precisely because they minimize between-subject variability. When most participants show similar effects, error bars shrink

and statistical power for detecting main effects increases. However, this property (low between-subject variability in the effect size) is mathematically detrimental to reliability. For test-retest reliability, the intraclass correlation coefficient (ICC), a standardized measure of the degree of consistency or reproducibility of quantitative measurements, represents the proportion of total variance attributable to true between-subject differences, as shown in Equation 3:

Equation 3. Intraclass correlation coefficient

where $\sigma^2_{between}$ represents between-subject variance and σ^2_{error} represents error variance. When between-subject variance ($\sigma^2_{between}$) is low, ICC necessarily decreases even with minimal measurement error. In simple terms: tasks where everyone shows similar effect sizes produce small error bars for group averages (good for detecting main effects) but provide little information about stable individual differences (bad for correlational analyses or test-retest reliability assessment).

The same mathematical principle applies to internal consistency of difference waves. When most participants show similar difference wave amplitudes (small between-person variance), combined with high inter-condition correlations (as demonstrated in Section 2.2), the resulting difference wave reliability will be poor. Tasks optimized for demonstrating

robust group-level differences become unsuitable for analyses requiring reliable individual difference measures.

This reliability challenge has cascading implications for statistical inference and replication:

First, reduced statistical power: Unreliable measures require substantially larger sample sizes to achieve adequate power. A study adequately powered for reliable individual conditions may be severely underpowered for the difference wave, increasing Type II error rates and making genuine effects harder to detect.

Second, effect size instability: Unreliable measures produce unstable effect size estimates across samples. Even when true effects exist, observed effect sizes will vary substantially due to measurement error, producing high heterogeneity in meta-analyses even when paradigms are similar.

Third, contradictory findings become expected: When reliability approaches zero, observed effects reflect primarily random error rather than signal. Replication "failures" become likely even when true effects exist, because each study is essentially measuring noise rather than stable neural activity.

Fourth, publication bias interaction: Low reliability combined with publication bias means the published literature may systematically overestimate ERP difference wave effects. Studies finding significant effects

(due to random error favoring the hypothesis) get published, while null findings (equally likely with unreliable measures) remain in file drawers. This creates illusory robustness in published literature while masking underlying measurement instability.

Fifth, systematic underestimation of true relationships: Low reliability does not merely reduce statistical power; it systematically distorts observed effect sizes through attenuation. Classical Test Theory provides a correction formula:

where ρ_{xx} and ρ_{yy} represent the reliabilities of measures X and Y. Consider a plausible scenario: an observed correlation of r = .20 between an N2 difference wave (reliability ρ = .25) and a behavioral measure (reliability ρ = .80). The corrected estimate reveals the true relationship:

The observed correlation underestimates the true relationship by more than half. Many null findings in the literature, interpreted as evidence that neural and behavioral measures are dissociated, may reflect measurement artifacts rather than genuine absence of brain-behavior relationships. This has profound implications: studies concluding that "N2 amplitude does not correlate with switching costs" may be documenting measurement failure rather than theoretical insight.

Sixth, unstable group-level effects: The reliability crisis extends beyond individual differences research to threaten group-level inference itself. When difference waves are unreliable, the observed effect size in any particular sample reflects substantial measurement error, not just true signal. Consider a meta-analysis synthesizing 10 studies of the same N2 switching effect. If difference wave reliability is $\rho = .30$ (as predicted under high inter-condition correlation scenarios), each study's observed effect size is approximately:

Even when all studies measure the same true effect with identical paradigms, observed effects will vary by ±45% purely due to measurement error. Meta-analyses will report high heterogeneity ($I^2 > 75\%$), interpreted as genuine paradigm differences or moderator effects, when the variability actually reflects measurement instability. Researchers will develop increasingly complex theories to explain contradictions that are primarily psychometric artifacts. This isn't hypothetical: the N2 language switching literature exhibits exactly this pattern—similar paradigms producing divergent findings, spawning elaborate theoretical frameworks to reconcile results that may partially reflect unstable measurement.

The publication bias amplification effect: When measures are unreliable, publication bias doesn't merely select for significant findings—it systematically distorts the published literature in ways that compound

interpretive difficulties. Consider the mechanism: A laboratory collects data on N2 language switching effects with an unreliable difference wave ($\rho_{DD'} = .30$). Across repeated studies, approximately 50% will find "significant" switching effects by chance, while 50% will find null results (also by chance, since the measure is dominated by noise). The significant findings get published with theoretical interpretations ("N2 reflects inhibitory control"), while null findings remain unpublished. Other laboratories, attempting to replicate using different paradigm variations, encounter the same reliability problem but don't realize it. Some find "significant" effects (published with alternative theoretical interpretations: "N2 reflects conflict monitoring"), others find null results (unpublished).

The published literature now contains multiple significant findings with contradictory patterns, all from unreliable measurements, none reporting reliability coefficients. Researchers interpret this as theoretical complexity requiring sophisticated reconciliation. Meta-analysts note high heterogeneity and propose moderators (paradigm type, language proficiency, training history). The field invests years developing elaborate models to explain contradictions that partially reflect measurement instability.

This isn't speculation—it's the observed pattern in N2 language switching research. Breaking this cycle requires not merely reporting reliability but making reliability a gatekeeping criterion: studies with unverified or

inadequate measurement properties should not be interpretable as theoretical evidence, regardless of significance level.

3. Evidence from Bilingual Language Control Research

3.1 The N2 paradox: a two-decade theoretical debate

The N2 component in bilingual language switching has generated persistent controversy. Jackson et al. (2001) reported enhanced N2 amplitude for switch trials compared to non-switch trials in a predictable digit-naming paradigm, interpreting this as reflecting inhibitory control required to suppress the dominant language during switches to the weaker language. In contrast, subsequent studies using unpredictable picture-naming paradigms (e.g., Christoffels et al., 2007) reported different N2 patterns, with some showing attenuated or context-dependent effects.

These varying findings launched extensive theoretical efforts spanning two decades. Researchers proposed increasingly complex models invoking proactive versus reactive control mechanisms, context-dependent adaptation, training-induced plasticity, and individual differences in cognitive control capacity (Green & Abutalebi, 2013; Abutalebi & Green, 2016). The field treated this as a neurocognitive puzzle requiring sophisticated theoretical resolution.

Crucially, paradigm differences (predictable vs. unpredictable switching, digit naming vs. picture naming, varying cue-stimulus intervals) may indeed

produce legitimately different neural effects. Distinguishing genuine neurocognitive differences from measurement artifacts requires demonstrated reliability. If difference waves are unreliable, any pattern of results becomes possible regardless of underlying neural reality. The argument here is not that neural complexity is absent, but that reliability must be established to separate true effects from measurement noise. The field has constructed elaborate theoretical frameworks (proactive vs. reactive control, adaptive control hypothesis) without first establishing that the measures being compared are psychometrically sound. These contradictions may reflect measurement instability rather than neural complexity, though this possibility remains empirically unverified without direct reliability assessment.

However, examination of the empirical foundations reveals a striking omission: none of the foundational or subsequent major studies report psychometric reliability for their N2 difference waves.

3.2 The Pervasive Vulnerability: Problems Across the Parameter Space

Having established that difference wave reliability is mathematically determined by the Lord-Novick formula, we now examine whether this vulnerability manifests only under extreme conditions or across the plausible parameter space. Critically, our analysis reveals systematic

problems even when inter-condition correlations are moderate, not merely when they approach unity.

Table 2 demonstrates a consequential pattern: achieving adequate difference wave reliability ($\rho_{DD'} \geq .70$) requires constituent reliabilities approaching the ceiling of current ERP methodology, even when inter-condition correlations are moderate. Consider the implications at increasingly realistic correlation values:

At $\rho_{XY} = .50$ (moderate correlation): Constituent reliability must reach $\rho = .85$ to produce marginally adequate difference waves. While .85 falls within the typical range for well-decomposed ERP components (Clayson & Miller, 2017), it represents good—not routine—measurement quality, typically requiring 40+ trials per condition with rigorous artifact rejection.

At $\rho_{XY} = .60$: Required constituent reliability increases to $\rho = .88$, approaching the upper range of values reported in systematic generalizability studies (Clayson et al., 2021). Achieving this requires ≥60 trials per condition, optimal preprocessing, and favorable signal-to-noise characteristics.

At $\rho_{XY} = .70$: Required reliability of $\rho = .91$ exceeds typical values and requires near-optimal measurement conditions (>80 trials per condition, exceptional artifact rejection, strong component morphology).

At ρ_XY ≥ .80: Required reliabilities (ρ ≥ .94) approach theoretical limits, effectively unachievable with current ERP methodology regardless of trial count or preprocessing quality.

This sensitivity analysis reveals that the mathematical vulnerability we have demonstrated is not confined to extreme parameter combinations. Even optimistic assumptions about inter-condition correlations (ρ_XY = .50) require constituent reliabilities that, while achievable, are not routine. Across the entire plausible parameter space, difference wave reliability cannot be assumed from constituent conditions and must be empirically verified.

The critical implication: regardless of whether empirical measurements ultimately reveal ρ_XY values of .50, .70, or .90, the fundamental vulnerability persists. The two-decade absence of such measurements represents not merely incomplete reporting but systematic epistemic neglect—the field has operated without knowing whether its primary analytical approach produces reliable measurements.

Table 2. Sensitivity analysis: Required constituent reliability to achieve adequate difference wave reliability (ρ_DD' ≥ .70)

Inter-Condition Correlation	Required Constituent	Empirical Context

(ρ_XY)	Reliability (ρ)	
.50	.85	Good quality; typical for well-decomposed components with 40+ trials (Clayson & Miller, 2017)
.60	.88	Upper typical range; requires >60 trials/condition and optimal preprocessing (Clayson et al., 2021)
.70	.91	Exceeds typical values; requires >80 trials, exceptional artifact rejection, strong component morphology
.80	.94	Approaches theoretical maximum; rarely achieved even under optimal conditions
.85	.96	Beyond practical limits of current ERP methodology
.90	.97	Effectively unachievable with electrophysiological measurements

Note. Values calculated by algebraically solving the Lord & Novick (1968, pp. 66-67) formula for ρ when ρ_DD' = .70 (adequate reliability threshold for exploratory group comparisons). **Derivation:** Starting from ρ_DD' = (ρ – ρ_XY) / (1 – ρ_XY), multiply both sides by (1 – ρ_XY): ρ_DD'(1 – ρ_XY) = ρ – ρ_XY. Expand: ρ_DD' – ρ_DD'·ρ_XY = ρ – ρ_XY. Rearrange to isolate ρ: ρ = ρ_DD' + ρ_XY(1 –

ρ_DD'). Substituting ρ_DD' = .70: ρ = .70 + .30·ρ_XY. The "Empirical Context" column provides benchmarks from ERP reliability literature (Clayson et al., 2021; Clayson & Miller, 2017) rather than subjective assessments. Trial count estimates derive from Generalizability Theory studies showing that reliability increases with trial number but plateaus due to systematic variance components that trial averaging cannot eliminate. This sensitivity analysis demonstrates that reliability degradation remains consequential across the entire plausible parameter space. Even at ρ_XY = .50 (substantially lower than structural analysis suggests for tightly controlled paradigms), achieving adequate difference waves requires constituent reliability at the upper end of typical values. The mathematical vulnerability persists regardless of precise ρ_XY values. This analysis uses the equal-variance formula (Equation 2), which assumes $\sigma^2_X = \sigma^2_Y$. When constituent variances differ substantially (≥20% difference in trial counts after artifact rejection, or variance ratios exceeding 1.5:1), apply the general formula (Equation 1). The equal-variance assumption is typically satisfied in well-controlled ERP paradigms with balanced designs, though direct verification requires condition-specific variance reporting.

Confidence intervals not applicable: Table values are algebraic solutions to the deterministic Lord-Novick formula, not empirical estimates requiring uncertainty quantification. However, empirical reliability estimates should be reported with 95% CIs using bootstrapping or G-theory variance components.

3.2.1 Catastrophic Degradation Under Rigorous Experimental Control

While the sensitivity analysis demonstrates that problems persist even at moderate inter-condition correlations, the degradation becomes catastrophic when experimental design produces high correlations. If language switching paradigms exhibit ρ_XY ≥ .70—a plausible inference from structural analysis and cross-domain precedent, though empirically unverified in language switching specifically—then constituent reliabilities that appear excellent by standard psychometric criteria fail to produce adequate difference waves.

The structural reasoning: Rigorous experimental control is expected to produce high ρ_XY through a three-step mechanism. First, researchers design conditions differing minimally except for the variable of interest (e.g., language identity), matching stimuli, timing, motor responses, and attentional demands. Second, this design ensures that most neural processing (visual encoding, lexical access, phonological retrieval, articulation planning) remains identical across conditions, with these shared processes constituting the majority of variance based on task structure. Third, because shared processes dominate total variance, individual differences in these non-target processes are substantially preserved across conditions, producing high ρ_XY. Participants who show large N2 amplitudes in switch trials tend to show large N2 amplitudes in

non-switch trials, reflecting stable individual differences in the shared neural machinery.

This inference assumes that shared variance (common to both conditions) exceeds switching-specific variance (unique to the difference). However, if switching costs produce large, stable individual differences, this could reduce ρ_XY by introducing substantial between-person variance in the difference wave itself. Empirical measurement of ρ_XY is therefore critical to adjudicate between these alternatives.

Cross-domain precedent: While direct measurements in N2 language switching are absent, analogous within-subjects contrasts in other domains provide convergent evidence. Infantolino et al. (2018) reported r = .97 for fMRI amygdala responses to faces versus shapes, a tightly controlled contrast differing on a single stimulus dimension. Cofresí et al. (2022) found r = .85-.88 for P3 amplitudes to alcohol versus non-alcohol beverage cues. Most critically, Clayson et al. (2021) demonstrated that N2 components in Go/NoGo tasks exhibit large person × occasion variance (comparable to between-person variance), providing direct electrophysiological evidence that N2 difference measures are vulnerable to the reliability degradation we describe.

Table 1 quantifies this degradation across parameter combinations that would be particularly consequential if they reflect actual measurement conditions in language switching research. The mathematical relationship

reveals an alarming pattern: high inter-condition correlations suppress difference wave reliability even when constituent conditions achieve excellent psychometric properties by standard criteria.

Table 1. Simulated difference score reliability as function of inter-condition correlation

Individual Condition Reliability (ρ)	**Inter-Condition Correlation (ρ_XY)**	**Difference Score Reliability (ρ_DD')**	**Interpretation**	**Plausible***
.85	.50	.70	Adequate for group comparisons	
.85	.70	.50	Inadequate for published research	*
.85	.85	.00	Completely unreliable	*
.85	.90	-.50	Mathematically unstable	*
.90	.70	.67	Marginal despite excellent	*

			conditions	
.90	.85	.33	Unacceptable reliability	*

Note. Calculations use the simplified formula: ρ_DD' = (ρ - ρ_XY) / (1 - ρ_XY) as derived by Lord & Novick (1968, pp. 66-67) for the equal-variance case. This formula assumes equal variances across constituent conditions (σ^2_X = σ^2_Y), an assumption typically satisfied in well-designed ERP studies with balanced trial counts and uniform preprocessing (assumption typically satisfied based on methodological designs (balanced trial presentation, identical preprocessing pipelines, matched stimulus characteristics) described in the four foundational studies, though direct variance verification would require condition-specific trial counts not consistently reported in original manuscripts). For conditions with substantial variance heterogeneity (≥20% difference in retained trials after artifact rejection), researchers should use the general formula: ρ_DD' = (ρ_XX' + ρ_YY' - 2ρ_XY) / [2(1 - ρ_XY)]. ρ_XY represents the correlation between individuals' ERP amplitudes across the two conditions—the extent to which individual differences are preserved across experimental manipulations. Values below .70 are generally considered inadequate for exploratory group comparisons, with higher thresholds (≥.80) recommended for confirmatory research and individual differences studies (Nunnally, 1978, pp. 245-246; Clayson & Miller, 2017, p. 61). Asterisk (*) indicates parameter combinations most consequential if tightly controlled paradigms produce high inter-condition correlations (ρ_XY ≥ .70), though direct

measurement of these correlations in language switching research is critically needed. The mathematical vulnerability is established; whether it manifests at hypothesized magnitudes requires empirical verification. However, sensitivity analysis (Table 2) demonstrates that problems persist even when ρ_XY is moderate, making this concern robust across the plausible parameter space rather than confined to extreme values. **Confidence intervals not applicable:** Table values are deterministic mathematical calculations from the Lord-Novick formula, not empirical estimates requiring uncertainty quantification. "For example, when constituent reliabilities are ρ = .85 (good quality), inter-condition correlations above .67 produce inadequate difference waves (ρ_DD' < .70), illustrating that the vulnerability manifests well before correlations approach unity."

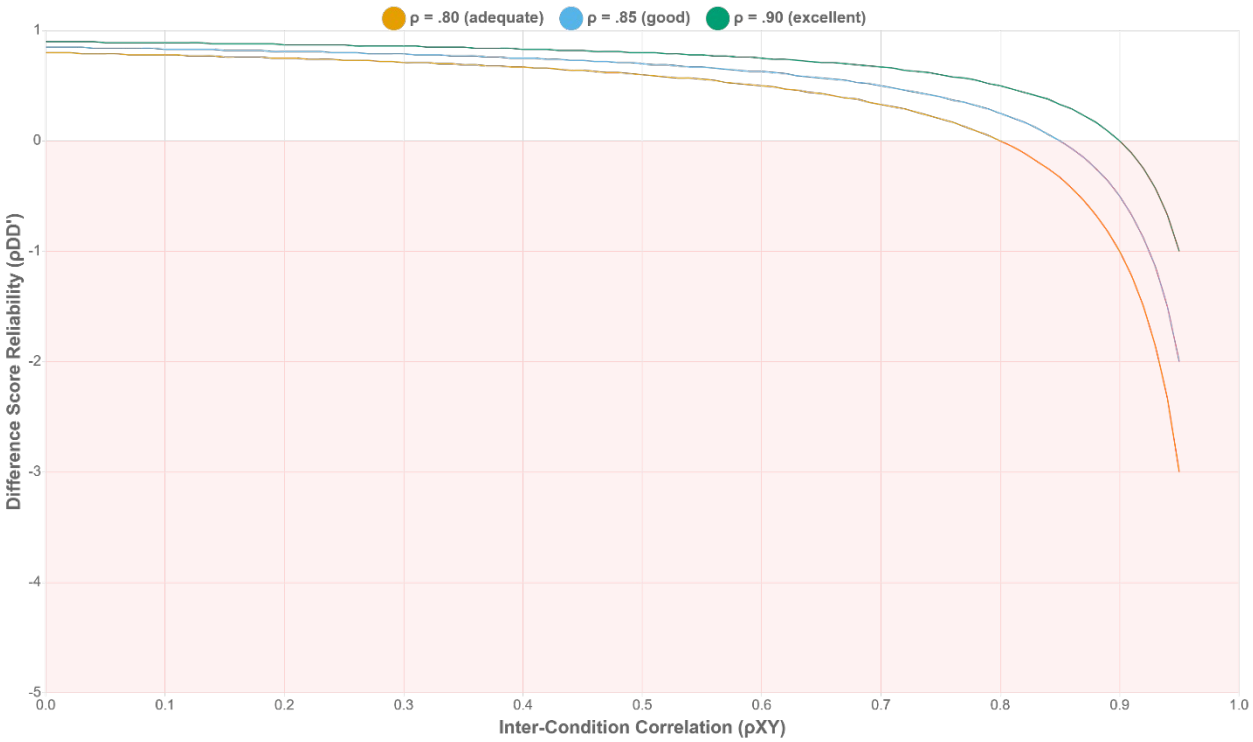

Figure 1. Difference score reliability degrades as inter-condition correlation increases, even when constituent conditions are highly reliable. Three curves show this relationship for constituent reliabilities of ρ = .80 (adequate), .85 (good), and .90 (excellent). As the correlation between individuals' ERP amplitudes across conditions (ρ_XY, x-axis) increases, difference score reliability (ρ_DD', y-axis) systematically declines. The horizontal dashed line indicates adequate reliability (ρ ≥ .70) for group comparisons. The left vertical line (ρ_XY = .67) marks where even excellent constituent reliability (ρ = .90) fails to produce adequate difference waves. The right vertical line (ρ_XY = .70) marks a plausible value for tightly controlled language switching paradigms based on structural considerations and cross-domain analogy, though direct empirical verification is needed. The shaded region indicates psychometric pathology: negative reliability means error variance exceeds true-score variance, producing maximally misleading measurements. Calculations use the Lord & Novick (1968, pp. 66-67) formula for equal-variance conditions. Complete data (288 points across three reliability curves) and R code available at OSF.

3.3 Evidence from foundational bilingual language control research

Study selection rationale. To evaluate whether reliability assessment preceded theoretical interpretation, we examined foundational studies that shaped the N2 debate in bilingual language control. Our selection criteria were: (1) Citation impact: Studies with >100 citations (as of October 2025)

indicating foundational influence on theoretical discourse; (2) Theoretical representation: Studies representing distinct positions in the N2 debate (switch > non-switch effects, context-dependent modulation, cue-stimulus interval effects, and training-induced plasticity); (3) Temporal span: Studies spanning 2001 to 2016 to capture the period during which core theoretical frameworks were established.

This purposive sampling addresses a specific research question: Did the studies that established current theoretical frameworks verify their measurements? This differs from systematic review objectives (prevalence estimation across all N2 research) or meta-analytic objectives (pooled effect size estimation). We emphasize that this analysis targets theory-shaping research rather than estimating field-wide prevalence. These four studies collectively accumulated ~1,575 citations and shaped fifteen years of foundational theoretical discourse on bilingual language control (2001-2016), establishing paradigms and interpretive frameworks that subsequent research built upon. If foundational studies establishing core theoretical frameworks proceeded without reliability verification, subsequent research building on these frameworks inherits this uncertainty. We acknowledge this approach does not establish field-wide prevalence rates. A comprehensive systematic review documenting prevalence rates across all N2 language switching studies represents critical future work that would complement the present analysis. Table 3 documents this pattern across representative foundational studies.

Table 3. Psychometric reliability reporting in foundational N2 language switching studies

Study	Citation Count (Approx.)	Paradigm Type	N2 Finding	Reliability Reported	Reliability Value
Jackson et al. (2001)	~414	Predictable sequence, digit naming	Switch > Non-switch	No	Not Reported
Christoffels et al. (2007)	~660	Unpredictable, picture naming	Context-dependent	No	Not Reported
Verhoef et al. (2009)	~396	Unpredictable, varied CSI	CSI-dependent	No	Not Reported
Liu et al. (2016)	~105	Training manipulation	Training effect on N2	No	Not Reported

Note. Selection criterion: Studies with >100 Google Scholar citations (as of October 2025) representing distinct theoretical positions in the N2 language switching debate. These four studies collectively accumulated ~1,575 citations across predictable versus unpredictable paradigms (Jackson; Christoffels; Verhoef), digit-naming versus picture-naming tasks

(Jackson; Christoffels; Verhoef), and training-based manipulations (Liu), demonstrating that psychometric verification was absent even in research that established foundational theoretical frameworks. CSI = cue-stimulus interval. "Not reported" indicates that reliability coefficients for difference waves or inter-condition correlations were absent from published manuscripts. This purposive sample illustrates systematic non-reporting in influential studies; field-wide prevalence estimation requires comprehensive systematic review (acknowledged limitation, Section 4.4).

Our analysis examines studies published through 2016. Whether reliability reporting has increased substantially in the most recent literature (2017-2025) requires systematic documentation beyond this manuscript's scope. However, previous assessments found minimal improvement despite availability of tools. Clayson's (2020) meta-analysis documented "routine failure of most ERN studies to report internal consistency" three years after initial methodological guidelines. His (2024) review concluded that psychometric principles remain "often neglected in most fields of neuroscience, including psychophysiology" seven years post-ERA Toolbox release. These assessments suggest persistent non-reporting despite available solutions. Nonetheless, documenting current practice (2024-2025) represents important future work that would establish whether our recommendations reflect ongoing or historical gaps.

Taken together, this consistent absence creates a fundamental interpretive ambiguity. Without reliability coefficients, any N2 effect remains uninterpretable; it may represent either a stable neurocognitive signal or a measurement artifact.

We cannot definitively conclude that unreported reliability was poor. Internal psychometric assessments may have occurred but remained unpublished. However, the mathematical demonstration (Sections 2 through 3.2) establishes that tightly controlled paradigms are highly vulnerable to reliability degradation, making poor reliability a plausible hypothesis. More fundamentally, science operates through transparent reporting and independent verification, not private internal checks. Even if researchers confirmed adequate reliability privately, the field cannot build cumulative knowledge on unverified foundations. The interpretive ambiguity created by non-reporting is itself sufficient to warrant the measurement practices we recommend, regardless of what internal assessments might have revealed.

The field has invested substantial effort in reconciling findings that may partly reflect measurement instability. Jackson et al. (2001) used a predictable digit-naming paradigm that exemplifies the reliability paradox described in Section 2.4. The highly constrained, closed-set response context produced robust group-level N2 effects that were widely cited and theoretically influential. However, these same features likely minimized

individual variance, reducing reliability for difference measures. Consequently, the field's preference for paradigms that yield strong group effects may have inadvertently promoted psychometrically weak measures, fueling decades of theoretical debate built on unstable findings.

4. Implications and the Path Forward

4.1 Reinterpreting theoretical debates

The systematic absence of reliability reporting carries profound implications for theoretical interpretation. The N2 "paradox" may not represent solely a neurocognitive puzzle but also a consequence of comparing measurements of unknown psychometric quality. When primary dependent variables have unknown reliability, replication failures become as likely as successes, and contradictory findings become expected regardless of true neural differences.

This does not invalidate all prior work. Studies may have measured genuine effects. However, without demonstrated reliability, this remains unknowable. The theoretical edifice may rest partly on unverified empirical foundations. Recognizing this constitutes not scientific failure but necessary course correction.

4.1.1 Distinguishing measurement artifacts from neurocognitive complexity

As established in Section 3.1, paradigm variations may produce genuine neural differences. However, without reliability verification, distinguishing these from measurement artifacts remains impossible. This confounding has specific implications: high measurement error can produce spurious contradictions, masked true effects, false replication failures, and theoretical proliferation where researchers develop increasingly complex models to reconcile contradictions that partially reflect measurement artifacts.

The N2 paradox may exemplify this confounding. Jackson et al.'s (2001) predictable digit-naming paradigm and Christoffels et al.'s (2007) unpredictable picture-naming paradigm differ substantially in design, stimulus materials, and cognitive demands. These differences could produce legitimately different N2 patterns. However, neither study reported difference wave reliability or inter-condition correlations, leaving open the alternative explanation: their findings partially reflect measurement instability. Distinguishing these possibilities requires the data that has not been collected.

A note on constructive criticism: This analysis presents forceful critique of current practices, but the intent is strengthening rather than dismissing the field. The studies examined here represent rigorous, influential research that advanced understanding of bilingual cognition. The researchers involved are not being accused of negligence; they followed the prevailing

methodological standards of their time. The problem is that those standards were systematically inadequate regarding psychometric verification. Recognizing this creates opportunity: by adopting more rigorous measurement practices going forward, we can build upon existing findings with stronger empirical foundations. Science progresses not by abandoning prior work but by refining the methods that produced it. it can be used somewhere else

4.2 A framework for methodological reform

Recognizing systemic problems enables systematic solutions. Following Parsons' (2022) call to transform measurement crisis into measurement revolution, the following multi-tiered framework can guide ERP research toward greater rigor, replicability, and theoretical progress:

4.2.1 Tier 1: Mandatory reliability reporting (immediate implementation)

Journals, reviewers, and funding agencies should mandate reporting of psychometric reliability for all key ERP effects, particularly difference waves. This requires routine application of established psychometric principles, not novel complex analyses.

Specific requirements:

- Report internal consistency coefficients (split-half reliability, Cronbach's α, or generalizability coefficients) for all primary ERP measures
- Specify calculation method clearly (e.g., odd-even split with Spearman-Brown correction)
- Justify minimum trial counts based on reliability goals rather than arbitrary cutoffs
- Include reliability assessment in pre-registration documents
- Critically, report inter-condition correlations when analyzing difference waves, enabling readers to evaluate potential reliability degradation

Evidence-based thresholds:

The appropriate reliability threshold depends on research context. Following principles from Nunnally (1978, pp. 245-246) and ERP-specific guidelines from Clayson & Miller (2017, p. 61), we establish the following standards:

- Preliminary/exploratory research: minimum $\rho \geq .70$
- Confirmatory research and individual differences studies: minimum $\rho \geq .80$
- Clinical applications and high-stakes decisions: minimum $\rho \geq .90$

Studies falling below appropriate thresholds must explicitly acknowledge limitations and temper theoretical claims accordingly.

Practical implementation considerations:

We note that implementing these recommendations faces practical obstacles. Many legacy datasets have insufficient trials for stable reliability estimates, raw data are often unavailable for retrospective assessment, and funding structures typically do not reward methodological validation studies. However, these barriers should motivate prospective reform rather than excuse continued neglect. Researchers designing new studies can implement reliability assessment at minimal cost during pilot testing. Journals can require verification for new submissions while grandfathering existing literature. Funding agencies can allocate modest resources for large-scale validation studies that would benefit entire research domains. The path forward is clear even if obstacles exist.

Retrospective reliability assessment:

Researchers with existing ERP datasets can retrospectively assess difference wave reliability using straightforward procedures:

1. **Split-half approach:** Randomly split trials into odd and even subsets, calculate difference waves for each subset, correlate them, and apply Spearman-Brown correction: reliability = 2r/(1 + r)

2. **ERA Toolbox:** Use the freely available ERP Reliability Analysis (ERA) Toolbox (Clayson et al., 2016) to partition variance across multiple

sources and provide generalizability coefficients

3. **Transparent reporting:** Report results in online supplements or brief reports, even when reliability proves inadequate. Acknowledge post-hoc nature and note as limitation. Transparent acknowledgment that published findings may rest on unreliable measurements advances the field more than continued silence.

4.2.2 Tier 2: Adopting superior psychometric frameworks

While Classical Test Theory metrics constitute crucial first steps, Generalizability Theory (G-theory) offers more sophisticated and appropriate frameworks for ERP data (Shavelson & Webb, 1991; Brennan, 2001; Clayson et al., 2021). In Classical Test Theory, this metric is called reliability; in Generalizability Theory, the equivalent measure for absolute decisions is called the dependability coefficient (G-theory), while the generalizability coefficient addresses relative decisions (ranking individuals within CTT frameworks).

Advantages of G-theory:

- Partitions variance into multiple sources (persons, trials, occasions) rather than conflating error sources
- Provides nuanced understanding of reliability determinants

- Handles unbalanced designs common in ERP research
- Distinguishes relative decisions (generalizability coefficient for ranking individuals) from absolute decisions (dependability coefficient for absolute score interpretation)

Practical implementation:

The ERP Reliability Analysis (ERA) Toolbox provides freely available, user-friendly MATLAB software specifically designed for implementing G-theory with ERP data (Clayson et al., 2016). This tool eliminates practical barriers to adoption. Continued failure to assess reliability constitutes active choice to ignore available solutions rather than technical limitations.

4.2.3 Tier 3: Moving beyond averaging (long-term paradigm shift)

While ensuring averaged waveform reliability is necessary, averaging discards valuable information in trial-to-trial variability. Advanced statistical methods leverage this variability for increased power and mechanistic insight.

Single-trial approaches:

Hierarchical Linear Models (mixed-effects models) analyze data at the single-trial level, modeling ERP amplitude as a function of categorical and continuous predictors while accounting for multiple variance sources (Smith & Kutas, 2015; Volpert-Esmond et al., 2021; Bagiella et al., 2000). These methods offer important advantages:

- **Transform rather than eliminate psychometric challenges:** By modeling condition effects directly rather than through subtraction, single-trial approaches avoid the specific mathematical vulnerability we have demonstrated. However, reliable estimation of condition effects still requires adequate signal-to-noise ratios at the trial level, and individual differences in condition effects (random slopes) require sufficient within-subject trial sampling.

- **Increase statistical power:** Utilizing all trials rather than averaged waveforms provides more information for parameter estimation, particularly when incorporating trial-level covariates (response time, previous trial type, stimulus characteristics).

- **Enable complex modeling:** Simultaneous estimation of multiple effects and their interactions becomes tractable, avoiding the proliferation of difference waves required in traditional approaches.

Single-trial methods represent a valuable complement to, not replacement for, careful psychometric assessment of averaged measures. Different research questions require different analytical approaches: component identification and time-course characterization often rely on averaged waveforms, while individual differences and brain-behavior correlations may benefit from single-trial modeling. The key principle remains consistent

across methods: verify that your measurements are reliable before building theoretical interpretations upon them.

Alternatives to difference wave subtraction:

Beyond improving difference wave reliability or adopting single-trial approaches, researchers should consider analytical methods that avoid categorical condition subtraction entirely. Regression-based ERP analysis models amplitude as a continuous function of multiple predictors simultaneously (Smith & Kutas, 2015; Hauk et al., 2006), avoiding the need to categorize trials into discrete conditions. Multivariate pattern analysis (MVPA) classifies trials based on distributed spatial patterns rather than univariate amplitude differences, providing an alternative metric (decoding accuracy) that may exhibit better psychometric properties (King & Dehaene, 2014). Time-frequency analysis decomposes EEG signals into constituent neural oscillations, offering mechanistically transparent measures: frontal midline theta (4-8 Hz) indexes conflict monitoring, alpha-band suppression (8-13 Hz) reflects active inhibition, and beta-band modulation (13-30 Hz) tracks motor preparation (Cohen, 2014). Each frequency band can be analyzed independently without requiring difference waves, potentially avoiding the reliability degradation we have demonstrated.

These methods complement rather than replace traditional ERP analysis. Difference waves remain valuable for isolating specific cognitive processes

when properly validated. However, triangulating findings across multiple analytical approaches (difference waves with verified reliability, single-trial mixed models, time-frequency decomposition, and multivariate pattern analysis) provides more robust evidence than relying solely on unvalidated subtraction measures. The diversity of analytical methods available strengthens rather than undermines the case for systematic reliability assessment: each method requires its own psychometric validation.

4.2.4 Addressing Implementation Barriers and Stakeholder Incentives

Implementing mandatory reliability reporting confronts practical obstacles that require systematic analysis and stakeholder-specific solutions. The ERA Toolbox has been freely available since 2016, yet adoption remains limited (Clayson, 2020, 2024). Understanding why previous tools failed to achieve widespread uptake is essential for designing effective reform.

Barrier 1: Legacy datasets and retrospective assessment

Many existing datasets were not designed with reliability assessment in mind, lacking sufficient trial counts for stable estimates or appropriate trial-splitting procedures. Raw data are often unavailable for retrospective reanalysis due to storage limitations, participant consent restrictions, or institutional data retention policies.

Solution: Implement requirements prospectively while grandfathering published work. Journals should establish explicit cutoff dates (e.g., "Studies submitted after January 2026 must report difference wave reliability") allowing legacy literature to remain valid while improving future practice. For influential paradigms where retrospective assessment is scientifically valuable, funding agencies could support targeted replication-with-verification studies using modern psychometric frameworks.

Barrier 2: Perverse incentives and publication bias

Researchers may reasonably fear that transparently reporting poor reliability will reduce publication chances, creating incentive structures that reward suppression over transparency. Junior researchers lacking job security face particular pressure to produce "clean" findings that maximize publication success.

Solution: Journals should explicitly commit to publishing well-designed studies regardless of reliability outcomes, provided results are transparently reported and appropriately interpreted. This requires policy statements such as: "This journal values methodological rigor and transparent reporting. Manuscripts will not be rejected solely because reliability estimates fall below optimal thresholds, provided authors acknowledge limitations and temper theoretical claims accordingly." Registered Reports format, where study designs are peer-reviewed and provisionally accepted before data collection, separates methodological

quality from outcome-dependent decisions, eliminating publication bias against null findings or suboptimal psychometric properties.

Barrier 3: Computational accessibility and expertise gaps

The ERA Toolbox requires MATLAB (commercial license >$2,000 for individuals, though institutional licenses may be available), creating access barriers particularly in under-resourced institutions and international contexts. Additionally, psychometric expertise may be limited in research groups with primarily cognitive or linguistic training, making reliability assessment technically challenging.

Solution: Develop and maintain open-source implementations in R (freely available) and Python (widely used in neuroscience), expanding accessibility beyond MATLAB users. Clear tutorials with worked examples reduce technical barriers (several are emerging; Clayson et al., 2021 provides detailed walkthroughs). Encourage methodological collaboration where expertise in experimental design and psychometrics can be pooled—co-authorship with quantitative specialists benefits all parties. Some institutions have established psychometric consultation services; expanding this model could provide expertise without requiring every researcher to become a psychometrician.

Barrier 4: Reviewer burden and inconsistent enforcement

Requiring reviewers to evaluate reliability reports increases cognitive load without necessarily improving review quality if reviewers lack psychometric training. Inconsistent enforcement across papers and journals creates perception of unfairness: some authors face demands for reliability evidence while others do not.

Solution: Develop standardized checklists and decision trees for editors and reviewers, translating psychometric principles into concrete evaluation criteria. Implement automated screening in manuscript submission systems (analogous to GRIM tests for summary statistics or word count enforcement) that flags missing reliability information before manuscripts reach reviewers, ensuring consistent application without human judgment variability. Provide reviewer training resources: brief primers on interpreting reliability coefficients, recognizing adequate versus inadequate values given study context, and distinguishing measurement issues from theoretical disputes.

Barrier 5: Field-specific norms and disciplinary inertia

Publication practices reflect decades of accumulated conventions. Senior researchers trained when reliability reporting was uncommon may view it as unnecessary burden rather than essential verification. Review processes institutionalize these norms through gatekeeping: manuscripts deviating from conventional practices face greater scrutiny.

Solution: Prestigious journals can catalyze change by example. When high-impact outlets (e.g., *Nature Neuroscience*, *Journal of Neuroscience*, *Psychophysiology*) adopt mandatory reliability reporting, the practice gains legitimacy and other journals follow. Learned societies (Society for Psychophysiological Research, Cognitive Neuroscience Society) can establish best-practice guidelines with community input, building consensus around minimum standards. Award committees can explicitly value methodological rigor and transparent reporting, shifting incentive structures at the disciplinary level.

Realistic implementation timeline:

Widespread adoption will require sustained effort over 5-10 years, not immediate transformation. However, incremental progress is achievable:

- **Years 1-2 (2026-2027):** Major journals adopt "strongly encouraged" language in author guidelines; R/Python ERA implementations released; tutorial workshops at conferences
- **Years 3-5 (2028-2030):** Mandatory reporting for new studies in leading journals; Registered Reports format expands; methodological review papers synthesize reliability evidence
- **Years 6-10 (2031-2035):** Retrospective validation of influential paradigms; meta-analyses incorporate reliability as moderator; community consensus on context-appropriate thresholds

The absence of perfect solutions should not justify continued neglect. Transparency about current measurement properties, even when suboptimal, advances scientific progress more than perpetuating unmeasured assumptions. Every incremental improvement—one lab adopting G-theory, one journal requiring reliability reporting, one training workshop teaching psychometric principles—contributes to the measurement revolution this field requires.

4.2.5 Exemplars of Rigorous Practice: What Good Reliability Looks Like

While this manuscript has focused on gaps in psychometric reporting, emerging research demonstrates that rigorous reliability assessment is achievable within standard ERP workflows. These exemplars provide concrete models for implementation:

Clayson et al. (2021) applied Generalizability Theory to N2 components in Go/NoGo tasks across multiple studies, systematically partitioning variance into person, trial, and occasion components. Their work demonstrated that: (1) internal consistency can reach acceptable levels (dependability ≥ .70) with sufficient trial counts (typically 40-60 trials per condition), (2) test-retest reliability is substantially lower than internal consistency due to large person × occasion variance, and (3) different ERP components exhibit different reliability profiles requiring component-specific optimization.

Critically, they transparently reported when reliability fell below optimal thresholds and discussed implications for interpretation.

Cofresí et al. (2022) examined P3 responses to alcohol cues with comprehensive psychometric assessment, reporting internal consistency, test-retest reliability, and inter-condition correlations across two independent samples. Their transparency about reliability degradation in difference scores (internal consistency r = .37, test-retest r = .20, despite excellent individual condition reliabilities) exemplifies the honest reporting this field requires.

Carbine et al. (2021) demonstrated generalizability assessment for food-related ERPs, showing that reliability varies systematically across stimulus categories, electrode sites, and participant characteristics. Their work illustrates that adequate reliability is achievable but requires deliberate optimization rather than default parameter selection.

These studies share critical features: (1) Reliability assessment was integral to the research design, not an afterthought; (2) Results were reported transparently regardless of outcomes; (3) Limitations were explicitly acknowledged when reliability proved suboptimal; (4) Theoretical claims were tempered appropriately given measurement properties. This represents the standard toward which ERP research should aspire.

Importantly, none of these studies required extraordinary resources or years of additional data collection. Reliability assessment added days or weeks to the analytical timeline, not months. The barrier is not technical difficulty but disciplinary norms that have not yet made psychometric verification mandatory.

4.3 Implementation Guidelines by Stakeholder

For researchers designing new studies:

- Calculate required trial counts based on reliability goals (not arbitrary cutoffs)
- Verify reliability during pilot testing before full data collection
- Pre-register reliability assessment procedures and minimum thresholds

For researchers with existing data:

- Conduct retrospective reliability analyses using methods detailed in Section 4.2.1
- Report findings transparently, even when reliability proves inadequate
- Acknowledge limitations explicitly in manuscripts and presentations

For reviewers:

- Request reliability coefficients and inter-condition correlations for all difference wave effects
- Question theoretical conclusions based on unmeasured psychometric properties
- Recommend increased trial counts or alternative analyses if reliability inadequate
- Distinguish measurement quality from theoretical merit

For journal editors:

- Establish reliability reporting as mandatory submission requirement
- Include reliability and inter-condition correlation assessment in review criteria and checklists
- Consider Registered Reports format for ERP studies requiring reliability verification before data collection
- Commit to publishing well-designed studies regardless of reliability outcomes when limitations are transparently acknowledged

For funding agencies:

- Require reliability assessment in grant proposals
- Support methodological validation studies alongside hypothesis-driven research
- Fund development of open-source psychometric tools (R/Python implementations)

- Incentivize multi-laboratory collaborations for paradigm validation

4.4 Limitations and future directions

This analysis examines four foundational studies to illustrate systemic issues in theory-shaping research but does not constitute a comprehensive systematic review of all N2 language switching literature. A complete quantitative synthesis examining reporting rates across the full ERP literature (all components, paradigms, and domains) represents critical future work. Such systematic documentation would provide definitive evidence for field-wide reporting practices and enable meta-analytic assessment of how unmeasured reliability has affected theoretical conclusions.

Additionally, this analysis critiques the absence of reliability reporting and inter-condition correlation measurement without directly demonstrating that N2 reliability is inadequate. Definitive conclusions require reanalysis of raw data from multiple laboratories calculating difference wave reliability using standardized procedures. This represents essential future work that could confirm or refute the theoretical predictions presented here. However, the core issue transcends whether specific N2 measurements prove reliable: the field has proceeded for decades without demanding evidence that its primary measures are sound. This constitutes fundamental methodological oversight regardless of what empirical testing might reveal.

The current findings situate measurement reliability concerns within a broader framework of methodological heterogeneity in ERP research. Recent analysis has demonstrated that between-study variability in analytical choices (time windows, electrode selection, baseline correction) contributes substantially to divergent N2 patterns in bilingual switching (Mojumdar, 2025). The present manuscript identifies a complementary within-study source of instability: unreported difference wave reliability. Future research should address both levels simultaneously, establishing standardized analytical protocols across laboratories while verifying psychometric properties within individual studies.

The ERP research community has initiated efforts to address these concerns. The ERA Toolbox (Clayson et al., 2016, 2021) was developed to facilitate reliability assessment, and major journals (*Psychophysiology*, *International Journal of Psychophysiology*) have encouraged (though not mandated) reliability reporting in author guidelines post-2017. However, subsequent research indicates that reliability reporting remains rare in practice. The foundational studies analyzed in Table 3 span 2001-2016, and systematic evaluation indicates that reliability reporting has not become standard practice despite available tools and journal encouragement. Critical future work includes documenting whether reliability reporting has increased in recent publications (2020-2025) and whether adoption of available tools has extended beyond measurement-focused specialists.

Confidence interval reporting limitations. Several foundational studies cited for empirical reliability values (Infantolino et al., 2018; Cofresí et al., 2022) did not report confidence intervals for all correlation coefficients. Infantolino et al. (2018) reported correlations (r = .97 for faces-shapes similarity; SB = -.06 for difference score reliability) without accompanying uncertainty estimates. While Cofresí et al. (2022) provided 95% confidence intervals for reliability coefficients in their Tables 3-4, inter-condition correlations reported in footnotes lacked uncertainty quantification. This reflects a broader reporting gap in the ERP reliability literature and underscores the need for comprehensive uncertainty reporting in psychometric assessments. Where possible, we have included confidence intervals from source papers; where unavailable, we acknowledge this limitation while noting that the absence of CIs in original publications does not diminish the theoretical validity of the mathematical relationships we demonstrate.

Beyond reliability, difference waves assume measurement invariance: that ERP components reflect equivalent constructs across conditions (Meredith, 1993). Establishing invariance requires factor analytic approaches rarely applied in ERP research. This represents an additional interpretive consideration beyond the reliability issues we emphasize.

5. Conclusion

The N2 component in bilingual language switching may represent not merely a neurocognitive paradox requiring theoretical resolution but also a measurement problem requiring methodological correction. Two decades of apparent contradiction may reflect both cross-paradigm inconsistency and systematic neglect of psychometric validation. In the absence of reliability estimates or inter-condition correlations, it remains empirically uncertain whether these contradictions arise from unstable measurement or genuine neural complexity.

This issue likely extends beyond bilingual research. Apparent contradictions in ERP findings across cognitive neuroscience may similarly result from unrecognized methodological confounds. The mathematical fragility of difference scores, combined with near-total absence of psychometric reporting, undermines confidence in theoretical disputes.

Recognizing this methodological uncertainty constitutes scientific progress rather than failure. Establishing clear paradigm specifications and adopting rigorous measurement standards are prerequisites for cumulative understanding. The absence of empirical data on inter-condition correlations does not weaken this argument; it defines its urgency. The field has not yet collected the psychometric evidence necessary to evaluate the reliability of its most frequently used analytical tools.

The path forward requires renewed commitment to foundational measurement principles, including systematic psychometric validation of

ERP measures, routine reporting of inter-condition correlations for difference waves, the application of appropriate statistical frameworks such as generalizability theory and mixed-effects models, transparent documentation of reliability estimates and trial-count justifications, and explicit acknowledgment when measures fall below acceptable reliability thresholds.

By embracing psychometric discipline, ERP research can progress from debating potential artifacts toward constructing mechanistic understanding of the neural basis of cognition. This represents not abandonment of prior work but building upon it with stronger methodological foundations. Following Parsons' (2022) call for a measurement revolution, we must recognize when our measures may not capture what we assumed they did and take systematic steps to ensure they do going forward. The question is not whether ERP difference waves can be reliable (with appropriate paradigm design and verification, they certainly can) but whether we will demand evidence that they are reliable before building theoretical edifices upon them. Two decades of debate over the N2 paradox demonstrates the cost of not asking this question and not measuring the parameters that determine psychometric quality. We cannot afford to repeat this pattern in other domains without learning this fundamental methodological lesson.

References

Abutalebi, J., & Green, D. W. (2016). Neuroimaging of language control in bilinguals: Neural adaptation and reserve. *Bilingualism: Language and Cognition*, *19*(4), 689–698. https://doi.org/10.1017/S1366728916000225

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*(1), 13–20. https://doi.org/10.1111/1469-8986.3710013

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-3456-0

Carbine, K. A., Clayson, P. E., Baldwin, S. A., LeCheminant, J. D., & Larson, M. J. (2021). Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores. Part 2: Application to food-based tasks and stimuli. *International Journal of Psychophysiology*, *166*, 188–198. https://doi.org/10.1016/j.ijpsycho.2021.02.015

Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research*, *1147*, 192–208. https://doi.org/10.1016/j.brainres.2007.01.137

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clayson, P. E. (2020). Moderators of the internal consistency of error-related negativity scores: A meta-analysis of internal consistency estimates. *Psychophysiology*, *57*(6), e13583. https://doi.org/10.1111/psyp.13583

Clayson, P. E. (2024). The psychometric upgrade psychophysiology needs. *Psychophysiology*, *61*(4), e14522. https://doi.org/10.1111/psyp.14522

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2021). Evaluating the internal consistency of subtraction-based and residualized difference scores: Considerations for psychometric reliability analyses of event-related potentials. *Psychophysiology*, *58*(4), e13762. https://doi.org/10.1111/psyp.13762

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2016). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. *International Journal of Psychophysiology*, *111*, 68–79. https://doi.org/10.1016/j.ijpsycho.2016.10.012

Clayson, P. E., Carbine, K. A., Baldwin, S. A., Olsen, J. A., & Larson, M. J. (2021). Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores. Part 1: Algorithms, framework, and implementation. *International Journal of Psychophysiology*, *166*, 174–187. https://doi.org/10.1016/j.ijpsycho.2021.01.006

Clayson, P. E., & Miller, G. A. (2017). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*, *111*, 57–67. https://doi.org/10.1016/j.ijpsycho.2016.09.005

Cofresí, R. U., Piasecki, T. M., Hajcak, G., & Bartholow, B. D. (2022). Internal consistency and test-retest reliability of the P3 event-related potential (ERP) elicited by alcoholic and non-alcoholic beverage pictures. *Psychophysiology*, *59*(10), e14030. https://doi.org/10.1111/psyp.13967

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. MIT Press.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68–80. https://doi.org/10.1037/h0029382

Declerck, M., Özbakar, E., & Kirk, N. W. (2021). Is there proactive inhibitory control during bilingual and bidialectal language production? *PLoS ONE*, *16*(9), e0257355. https://doi.org/10.1371/journal.pone.0257355

Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, *25*(5), 515–530. https://doi.org/10.1080/20445911.2013.796377

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*(4), 1383–1400. https://doi.org/10.1016/j.neuroimage.2005.11.048

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage*, *173*, 146–152. https://doi.org/10.1016/j.neuroimage.2018.02.024

Jackson, G. M., Swainson, R., Cunnington, R., & Jackson, S. R. (2001). ERP correlates of executive control during repeated language switching. *Bilingualism: Language and Cognition*, *4*(2), 169–178. https://doi.org/10.1017/S1366728901000268

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of*

Chiropractic Medicine, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Liu, H., Liang, L., Dunlap, S., Fan, N., & Chen, B. (2016). The effect of domain-general inhibition-related training on language switching: An ERP study. *Cognition*, *146*, 264–276. https://doi.org/10.1016/j.cognition.2015.10.004

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Mojumdar, P. (2025, June 20). *An enigma of our own making: How methodological heterogeneity created the N2 reversal in bilingual language switching*. OSF Preprints. https://doi.org/10.31234/osf.io/p56ey_v1

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill Book Company.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill, Inc.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Parsons, S. (2022). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *Meta-Psychology*, *6*, MP.2020.2577. https://doi.org/10.15626/MP.2020.2577

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157–168. https://doi.org/10.1111/psyp.12317

Verhoef, K., Roelofs, A., & Chwilla, D. J. (2009). Role of inhibition in language switching: Evidence from event-related brain potentials in overt picture naming. *Cognition*, *110*(1), 84–99. https://doi.org/10.1016/j.cognition.2008.10.013

Volpert-Esmond, H. I., Page-Gould, E., & Bartholow, B. D. (2021). Using multilevel models for the analysis of event-related potentials. *International Journal of Psychophysiology*, *162*, 145–156. https://doi.org/10.1016/j.ijpsycho.2021.02.006

Author Contributions

Projnya Mojumdar: Conceptualization (lead), theoretical framework (lead), literature review (lead), writing (original draft, lead), writing (review and editing, lead), visualization (figure design), project administration. Purba Mojumdar: Methodology (mathematical modeling, simulation design), software (R code development and validation), formal analysis (computational verification), data curation (simulation datasets), writing (review and editing, technical sections).

Both authors approved the final manuscript and agree to be accountable for all aspects of the work.

Acknowledgments

The authors thank the open science community for developing tools and frameworks that enable transparent, reproducible research. Special appreciation to developers of the ERP Reliability Analysis (ERA) Toolbox for making psychometric assessment accessible to the ERP community, and to Sam Parsons for articulating the call for a measurement revolution in Meta-Psychology. The authors emphasize that critique of methodological practices should not be construed as questioning the integrity or rigor of individual researchers. The studies examined followed prevailing standards, and this analysis aims to strengthen collective practice going forward. All materials are openly available at https://osf.io/q2dkj/overview under CC-BY 4.0 license.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare no conflicts of interest

Data Availability Statement

This theoretical analysis does not involve primary data collection. All computational materials, including R code, generated figures, verification data, and reproducibility documentation, are publicly available via the Open Science Framework under a CC-BY 4.0 license.

Open Practices Statement

This manuscript presents a theoretical framework and methodological critique and does not involve original empirical data collection; all cited studies are available through their respective publishers. All supplementary materials are publicly available on the Open Science Framework at https://osf.io/q2dkj/overview under a CC-BY 4.0 license, permitting reuse with appropriate attribution. The repository includes annotated R code that programmatically verifies Tables 2 and 3 and generates Figure 1, as well as a comprehensive README file that serves as the primary source for reproducibility, explaining the formula, assumptions, and instructions. Preregistration was not applicable for this conceptual analysis.