Residualised Scoring for Component-Process Isolation: A Regression-Based Framework for Layered Psychological Measurement

Emile Boullineau
Independent Researcher, United Kingdom
Correspondence: emile@judgmentalism.org
ORCID: 0009-0001-1500-4990

Preprint, May 2026 (Version 5). This version broadens the framework from layered cognitive abilities to layered psychological measurement. V5 adds two MIDUS real-data demonstrations, one on an ability test and one on a self-report instrument. V5 also adds a head-to-head comparison with a structural equation model latent residual and supporting simulations. The estimator, diagnostic workflow, and core simulation files were deposited on a public open-science archive on 19 May 2026 (Version 4), before the MIDUS demonstrations reported here. The V4 core is unchanged in V5. The R pipeline accompanies the deposit.

Abstract

Psychological ability tests and self-report instruments often have layered assessment structures. An upstream score captures detection and a downstream score captures the related interpretation. A single composite score conceals these stages. Subscores carry upstream variance rather than isolate the downstream component. Branch-Residualised Interpretation (BRI) regresses the downstream score on the upstream score in a calibration sample, resulting in a person-level downstream estimate. The algebra is standard residualisation, the first step of the Frisch-Waugh-Lovell theorem. The contribution is an evaluated scoring workflow with operating limits that are empirically mapped. The operating limits were evaluated using Monte Carlo simulations under varying reliability and measurement conditions, including ceiling compression, heavy-tailed error, and correlated measurement error. Predictive utility is demonstrated using two MIDUS applications: an ability test and a self-report instrument. In the Brief Test of Adult Cognition by Telephone, the retention residual predicted later cognitive performance over and above encoding, non-memory baseline ability, and demographics, $p < .001$, $N = 2{,}393$. In the marital demonstration, the residual predicted later relationship dissolution, OR = 2.60, 95% CI [1.92, 3.54], extending the BRI to self-report instruments. BRI is a simple person-level scoring procedure when layered assessments are strongly justified.

Keywords: psychological measurement, residualisation, regression, component processes, layered architectures, observed-score validity, simulation methods, Frisch-Waugh-Lovell

1. The Component-Process Entanglement Problem

Component processes are often conceptually separable in psychological measurement. However, these processes are also often operationally entangled. Using a working-memory span task as an example: the participant encodes the items, holds them, and reproduces them. An encoding error and a maintenance failure produce the same observable mistake. A theory-of-mind vignette prompts participants to read a mental state and reason from that state. However, a single composite score does not indicate which stage drove the answer. Ability tests of emotional intelligence have the same problem. Strong signal detection and/or strong attribution can be reflected in a high score with the underlying stages often having distinct correlates and distinct meanings.

The problem emerges whenever a task has a layered structure where the interpretation of downstream scores is dependent on an input upstream. Composite scoring collapses the stages together. Branch-level subscores do not fully solve this because the downstream score still includes variance from the upstream process. As a result, the score partly reflects the very input the investigator may wish to separate.

The Branch-Residualised Interpretation (BRI) procedure is designed to address this shortcoming. In a calibration sample, BRI regresses the downstream score on the upstream score. The result is a residual which represents the component of downstream performance not explained linearly by the upstream input. Serving as a person-level estimate of downstream-specific variance, the regression operates on person scores rather than items. BRI can be applied to any suitable pair of upstream and downstream task scores. The task battery itself remains the instrument. Figure 1 illustrates the logic.

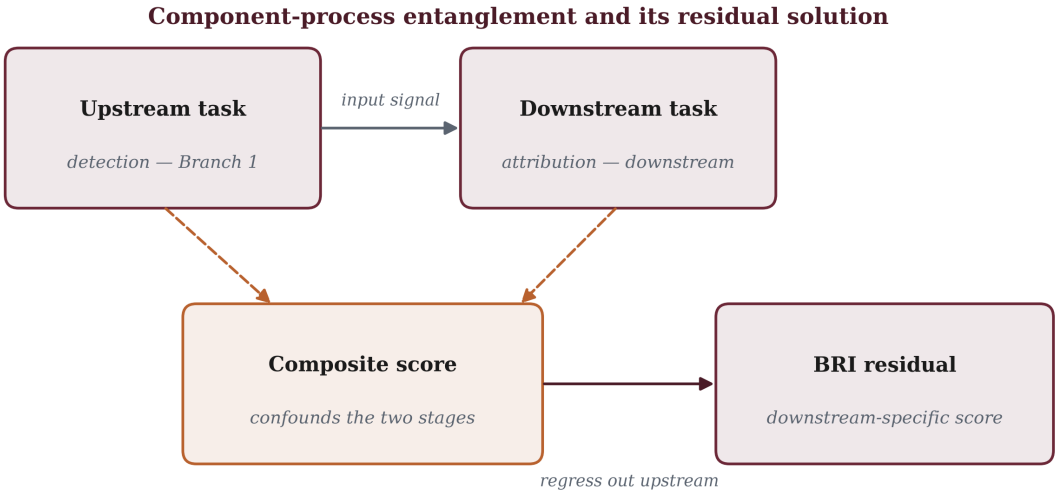

Figure 1. Component-process entanglement and the role of residualisation. The upstream task supplies an input signal that the downstream task must reason over. The shaded centre node represents a composite score that conflates variance from both stages. The Branch-Residualised Interpretation residual (right-hand node) is the variance in the downstream score that cannot be linearly explained by the upstream score, treated as a person-level score for the downstream component.

Ability Emotional Intelligence is the underlying motivating case for the present framework. Under the four-branch model (Mayer & Salovey, 1997; Mayer et al., 2008), an upstream signal is supplied by Branch 1

(perceiving of emotions) from which later interpretation and attribution are derived. The Branch Dissociation Hypothesis (Boullineau, 2026a) proposes that strong affective signal recognition coupled with poor attribution can yield a distinctive interpersonal profile, one that current blended scoring approaches often obscure. Ability EI is one example of this kind of layered architecture. The framework is intended for any layered assessment. A separate paper will apply BRI directly to ability EI using MSCEIT data. The two real-data demonstrations in this paper apply the framework to different layered assessments. Section 5 applies BRI to an ability test in MIDUS that has a layered format. It uses immediate and delayed word-list recall to predict later cognition. Section 6 applies the same logic to a layered self-report structure. The same MIDUS dataset is used. This dataset examines marital strain, marital-risk assessment and subsequent relationship dissolution/rupture. The framework is intended to apply across both types of layered assessment.

The remainder of the paper is organised as follows. Section 2 specifies the estimator. Section 3 presents five Monte Carlo studies evaluating the operating boundary. These reports include a comparison with a latent-residual structural equation model (SEM). Section 4 describes a pre-validation diagnostic workflow for real-world application. Sections 5 and 6 present two real-data demonstrations. Section 7 outlines the scope, limitations and relevant decision criteria. An open R pipeline is documented in Section 8.

2. The Procedure

2.1 General specification

Suppose a task produces two scores per person. An upstream score X captures detection or basic input ability. A downstream score Y captures interpretation or reasoning conditioned on that detection. What the investigator usually wants is the part of Y that is not already in X.

The estimator is the residual from the calibration-sample regression of Y on X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \widehat{\mathrm{BRI}}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

The slope $\hat{\beta}_1$ has substantive content of its own. Tighter coupling produces steeper slopes. The residual ranks participants by how far their downstream score sits above or below the fitted line. The procedure runs in six steps. (1) Measure the upstream component with a task that isolates the input-side process. (2) Measure the downstream component with a task that requires the downstream process while keeping upstream demand approximately constant. (3) Fit the linear regression of Y on X in the calibration sample. (4) Extract the residual. (5) Pre-validate by reporting reliability of both components, the upstream-downstream scatter, slope and R^2, the nonlinearity test, the bootstrap residual-score stability, and (where multiple task forms exist) cross-task residual correlation and hold-out factor structure. Section 4 develops Step 5 in detail. (6) Use the residual as a predictor in subsequent analyses.

2.2 The ability-EI instantiation

In the ability-EI case, the upstream score is a branch-level measure of affective signal detection (the perceiving subscale of the MSCEIT, Mayer et al., 2002; the Reading the Mind in the Eyes Test, Baron-Cohen et al., 2001; the Diagnostic Analysis of Nonverbal Accuracy, Nowicki & Duke, 1994). The downstream score is a measure of situated causal attribution under self-relevant ambiguity. The Branch Dissociation Hypothesis predicts that the residual variance in situated attribution unexplained by detection is the construct of theoretical interest.

2.3 Prior art and the contribution

The underlying algebra is the first step of the Frisch-Waugh-Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963). The semi-partial correlation (Cohen et al., 2003) is the sample-level analogue of BRI's per-person residual. Residualised change scores (Cronbach & Furby, 1970; Steyer et al., 1997) use the same residual within one task across time. Latent-residual SEM (Bollen, 1989; Rosseel, 2012) handles measurement error explicitly but requires multiple indicators per side, identification constraints, and a structural commitment. Two-stage residual inclusion (Terza et al., 2008) targets causal identification given an instrument. Factor-score residuals (McDonald, 1999) capture indicator-specific variance within a fitted factor model. The contribution here is a simulation-based account of when an observed-score residual works as a defensible *person-level score*, with the diagnostics, reliability evidence, and portability infrastructure required to use it responsibly.

The "perils of partialling" critique (Lynam et al., 2006; Sleep et al., 2017) cautions against partialling closely-overlapping traits because the residual then carries method-variance artefacts. BRI differs in three respects. The upstream and downstream tasks are genuinely different tasks rather than two ways of measuring the same construct. The procedure is restricted to layered detect-then-reason architectures. Study 3 below quantifies the contamination risk under shared method variance and shows it to be modest.

2.4 Sample-relative calibration and cross-sample transfer

A BRI score is *sample-relative*. The slope and intercept come from the calibration sample. Three implications follow. BRI supports between-person comparisons within a defined sample. For cross-study scoring, the calibration regression should be estimated once and applied to the target sample. In small samples the calibration slope is unstable and destabilises BRI rankings.

A separate simulation (`08_calibration_transfer.R`) tests cross-sample transfer under five mild changes to how the target sample's data are generated. Rank stability between the calibration-applied residual and the target-refit residual sits at .993 to .998 across every tested condition. The within-target ranking is essentially identical whether BRI is computed from the imported calibration or refit on the target. The Section 2.4 recommendation is supported under mild distribution shift. Larger shifts (cross-cultural, cross-cohort) require empirical re-validation.

2.5 What residualisation does and does not establish

Residual scoring isolates the part of one score that is linearly unrelated to the other. It carries familiar failure modes. Reliability problems propagate to the residual (Lord, 1967). Nonlinear coupling biases the linear residual (Cohen et al., 2003). Heavy-tailed measurement error destabilises OLS slopes (Huber, 1981). Section 3 evaluates each. A causal caveat applies. The upstream and downstream language describes the *task architecture*, not a causal claim. BRI is a descriptive scoring tool, not a causal identification strategy. Investigators wanting causal claims should turn to instruments, longitudinal mediation, or experimental manipulation of the upstream stage rather than to a residual.

3. Simulation Evidence

Five Monte Carlo studies map the procedure's operating boundary. All studies use synthetic data generated from a known latent structure. Cell means are reported in the tables below. Monte Carlo standard errors and per-cell raw outputs are written to the `Code/results/` folder of the open archive.

3.1 Study 1: Variance recovery and residual-score reliability

The first question is whether BRI actually isolates the downstream-specific variance when the data behave well. The second is whether the residual is reliable enough to use as a per-person score. The simulation answers both under classical measurement assumptions. For $N = 500$ participants per cell, draw a latent upstream score $T_X \sim \mathcal{N}(0, 1)$ and a latent downstream score T_Y correlated with T_X at one of three coupling strengths $\rho \in \{.30, .50, .70\}$. The downstream-specific component (the quantity BRI is meant to recover) is the part of T_Y orthogonal to T_X. Add independent normal measurement error scaled to give equal test reliabilities $r_{XX} = r_{YY} \in \{.70, .80, .90\}$, the standard range for ability tests in current use. Recovery is the correlation between the BRI residual and the true downstream-specific component. Parallel-forms reliability is computed across two independent draws of the measurement error.

	Coupling ρ		
Reliability r_{XX}	**.30**	**.50**	**.70**
.90	.94 / .88	.92 / .85	.87 / .77
.80	.88 / .77	.84 / .72	.77 / .62
.70	.82 / .67	.77 / .62	.68 / .51

Table 1. Recovery coefficient and parallel-forms residual-score reliability (recovery / reliability), 1,000 replications per cell.

At reliabilities of .80 or higher, recovery sits between .77 and .94 and residual reliability sits between .62 and .88. At a reliability of .70 with moderate-to-high coupling, residual reliability drops below .70 and the score should be used cautiously. Component reliabilities of .80 or higher are recommended.

3.2 Study 2: Reliability asymmetry and a corrected estimator

Real measurement instruments rarely deliver equally reliable scores on both sides of the layered architecture. The downstream component is often noisier than the upstream one, or vice versa. Study 2 asks how that asymmetry affects recovery, and whether a classical disattenuation correction repairs the damage. Hold coupling at $\rho = .50$ and vary upstream and downstream reliabilities independently across $r_{XX}^{(\text{up})}, r_{YY}^{(\text{down})} \in \{.60, .70, .80, .90\}$, the range that covers most existing instruments. Run 500 replications per cell. Compute recovery twice. Once with the uncorrected BRI residual. Once with a Lord-style disattenuation correction that scales the regression slope by the upstream reliability before extracting the residual, a standard adjustment for measurement-error attenuation in observed-score regressions.

	Upstream reliability $r_{XX}^{(\text{up})}$			
Downstream reliability $r_{YY}^{(\text{down})}$	**.60**	**.70**	**.80**	**.90**
.90	.88 / .85	.89 / .88	.91 / .90	.92 / .92
.80	.82 / .80	.83 / .82	.84 / .84	.86 / .85
.70	.77 / .75	.77 / .76	.78 / .78	.79 / .79
.60	.70 / .68	.71 / .70	.72 / .71	.72 / .72

Table 2. Recovery, uncorrected / corrected, 500 replications per cell.

Recovery is governed almost entirely by downstream reliability. The disattenuation correction reduces rather than improves recovery in most cells. The uncorrected estimator is the default. The correction is reserved for cases with independent reason to suspect restricted-range slope attenuation.

3.3 Study 3: Robustness sampler (ceiling, nonlinear coupling, correlated error)

Classical test theory assumes well-behaved measurement, but real instruments fail those assumptions in characteristic ways. Three failure modes matter most for layered architectures. The upper end of the scale can compress when a test is too easy (ceiling effects). The link between upstream and downstream can curve rather than run straight (nonlinear coupling). Shared method variance between the two scores can introduce correlated error. Study 3 stress-tests recovery against each in turn. Each panel uses 500 replications per cell with $r_{XX} = r_{YY} = .80$ and, where applicable, $\rho = .50$.

Panel A. Ceiling-effect compression.

Compression	0	.10	.20	.30	.50
Recovery	.85	.83	.82	.80	.77

Panel B. Quadratic coupling β_q.

β_q	0	.10	.20	.30	.40
Recovery	.84	.83	.80	.73	.64
r between residual and X^2	.00	.11	.22	.33	.44

Panel C. Correlated measurement error.

$\text{cor}(\varepsilon_X, \varepsilon_Y)$	0	.10	.20	.30	.50
Recovery of downstream-specific variance	.845	.852	.862	.872	.894
Common-method variance in residual	.002	.011	.017	.024	.035

Table 3. Recovery under three departures from ideal conditions, 500 replications per cell.

Ceiling effects degrade recovery modestly (.85 to .77 at fifty per cent compression). Nonlinear coupling produces two failure modes together. The linear residual misses an increasing share of the downstream-specific variance and absorbs an increasing share of upstream-quadratic variance. The Section 4 nonlinearity test catches this case. Positively correlated measurement error preserves recovery within the tested grid. Common-method contamination rises modestly with the error correlation, from .2 to 3.5 per cent of residual variance. Applied users should report any indication of shared method variance. A separate simulation in the open archive (`03_robustness.R`) shows recovery essentially flat across symmetric heavy-tailed error from t_3 to Gaussian.

3.4 Study 4: Regression power for downstream effects

A scoring procedure is only useful if it can detect the effects investigators care about at sample sizes investigators can actually collect. Study 4 asks what sample size a researcher needs to find a downstream-specific effect of a given size, using BRI as a covariate alongside the upstream score. Generate samples in which an outcome depends on the truly downstream-specific component with standardised effect $\beta_{\text{BRI}} \in \{.15, .20, .30, .40\}$, spanning small to medium effects in psychological convention (Cohen, 1988). Fit $Y_{\text{out}} \sim X + \text{BRI}$. Test the BRI coefficient at $\alpha = .05$. Replicate 500 times per cell at $r_{XX} = .80$.

	Sample size N				
β_{BRI}	**100**	**250**	**500**	**1000**	**2000**
.15	.22	.50	.82	.98	1.00
.20	.39	.78	.97	1.00	1.00
.30	.71	.99	1.00	1.00	1.00
.40	.95	1.00	1.00	1.00	1.00

Table 4. Power for the BRI coefficient at $\alpha = .05$, 500 replications per cell.

N around 500 gives eighty per cent power for small effects ($\beta_{\text{BRI}} = .15$). N around 250 gives close to eighty per cent power for small-to-medium effects ($\beta_{\text{BRI}} = .20$; observed power .78 in the simulated grid).

3.5 Study 5: BRI versus a structural-equation-model latent residual

An applied reader at a measurement journal will reasonably ask why one would not just fit a structural equation model with a latent residual. Study 5 maps the operating range over which BRI is a transparent observed-score approximation to that estimator. Draw two or three indicators per side from a known latent structure with loading $\lambda \in \{.60, .80\}$, latent coupling $\rho = .50$, and a downstream-specific outcome effect $\beta_{\text{out}} \in \{.20, .40\}$. For each cell, recover β_{out} via BRI on observed-score composites and via `lavaan::sem` on a latent-variable model with the downstream construct regressed on the upstream construct and the outcome regressed on both. Report bias and root-mean-squared error of each estimator. 1,000 replications per cell. Sample size $N \in \{250, 500\}$.

Design corner	BRI bias	BRI RMSE	SEM bias	SEM RMSE	SEM convergence
Weakest (k = 2, λ = .60, N = 250)	−.060	.099	−.030	.108	.96
Mid (k = 2, λ = .80, N = 500)	−.054	.075	−.054	.076	1.00
Strong (k = 3, λ = .80, N = 500)	−.013	.057	−.042	.067	1.00
Grid mean across all sixteen cells	−.047	.087	−.038	.085	.99

Table 5. Summary of BRI and SEM estimators across the Study 5 design grid (one corner cell each plus a grid mean across the sixteen cells, with β_{out} averaged within each cell). The complete sixteen-cell table is reproduced as `study10_bri_vs_sem_table.csv` in the supplementary code archive. SEM convergence is the proportion of replications returning admissible estimates.

Three patterns matter. When SEM is well identified (three indicators per side, loadings of .80, N of 500), BRI and SEM produce similar RMSE, with bias differences depending on cell. When SEM is weakly identified (two indicators, low loadings), it occasionally fails to converge and BRI returns a more biased but stable estimate. RMSE is broadly comparable across the operating range. The applied recommendation is to fit SEM when the indicators, loadings, and sample support it, and to fall back to BRI otherwise with the bias

profile from Table 5 in view. Section 7 develops the decision criteria. Figure 2 plots bias and RMSE across the sixteen design cells.

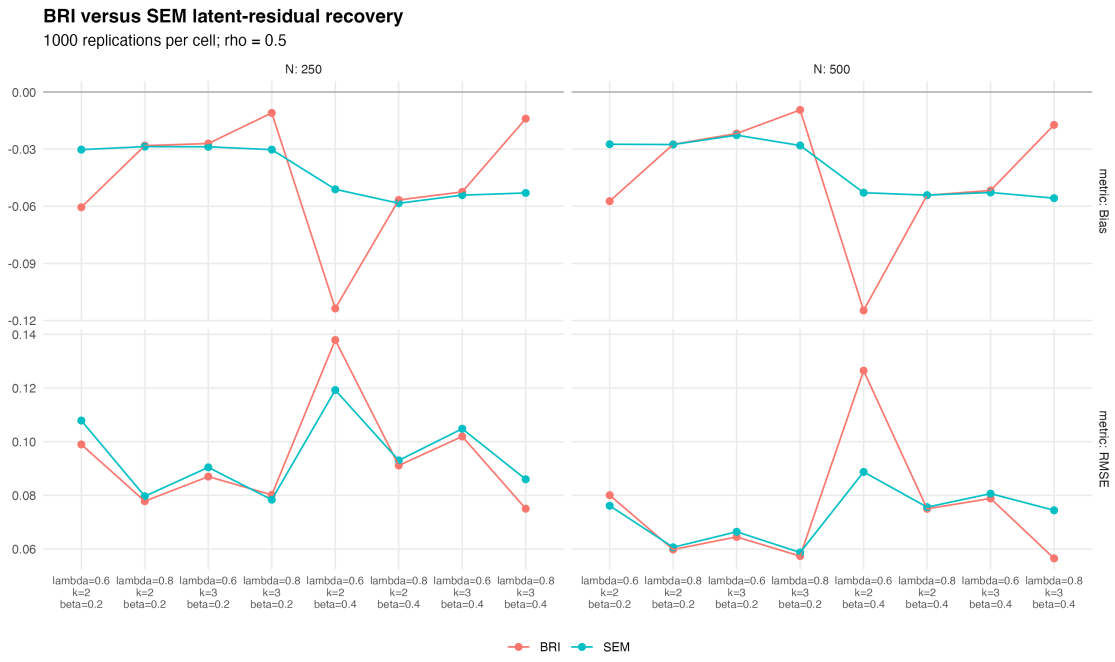

Figure 2. Bias (left) and RMSE (right) of the BRI observed-score residual estimator and the SEM latent-residual estimator for the downstream-specific outcome coefficient, across the sixteen design cells of Study 5. Lower is better on both panels. Reference line at zero bias.

4. Pre-Validation Diagnostics

The applied value of BRI depends on whether users can tell, on their own data, when the residual is and is not safe to interpret. The R pipeline provides each diagnostic through a single function (`pre_validate()`) that returns a structured report.

Component reliability. Report Cronbach α or test-retest reliability for both scores in the sample at hand. If reliabilities fall at or below .70 with moderate-to-high coupling, treat the residual as low-confidence and report bootstrap stability alongside any inference.

Upstream-downstream scatter and the calibration regression. Plot X against Y. Overlay the OLS regression line. Report the slope, R^2, and residual SD. A clean linear pattern licenses BRI. Visible curvature signals the nonlinearity test.

Nonlinearity test. Refit the calibration with a quadratic term in X. Report the quadratic coefficient, p-value, and ΔR^2. The trigger rule is a significant quadratic p-value at conventional alpha. ΔR^2 is reported as severity context, not as a separate trigger threshold. If the trigger fires, the recommended response is to refit the calibration with the quadratic term included and use the quadratic-orthogonalised residual in the substantive model. The MIDUS BTACT demonstration in Section 5 shows the workflow in action even when the severity is modest.

Bootstrap residual-score stability. Resample with replacement, refit the calibration, recompute the residual, report the mean Pearson correlation of residual scores across resamples. Above .95 indicates stable participant rankings. Below .90 indicates slope instability and should prompt a sample-size reconsideration.

Cross-task residual correlation (where available). If two parallel task forms exist, compute BRI on each and correlate the residuals across forms within participants. A correlation of .30 or higher indicates a stable per-person quantity rather than task-specific noise.

Hold-out factor-structure check (where multiple indicators exist). Fit a one-factor model to each construct in a holdout subsample. Report loadings and fit. The check rules out multidimensionality that would invalidate single-score residualisation.

`pre_validate()` runs the subset of diagnostics implemented in the bundled function (cross-task correlation, factor structure, bootstrap residual stability, nonlinearity test) and flags any that were skipped because inputs were missing. The remaining items in this section (component reliability, calibration scatter, hold-out factor structure) are reported alongside `pre_validate()` outputs in the BTACT demonstration script as a worked template. Investigators who report BRI without at least component reliability, the calibration scatter, the nonlinearity test, and bootstrap stability have not established that the residual is interpretable on their data.

5. Real-Data Demonstration: MIDUS BTACT

5.1 Dataset and architecture

The Midlife in the United States (MIDUS) study follows a national sample of American adults aged 25 to 74 at baseline across three waves (M1, 1995/96; M2, 2004/05; M3, 2013/14). The Brief Test of Adult Cognition by Telephone (BTACT; Lachman et al., 2014) was administered at M2 (N = 4,512) and M3 (N = 3,291) and covers episodic memory, working memory, executive function, inductive reasoning, and processing speed. A standardised composite (C3TCOMP at M3) aggregates the five abilities. MIDUS main-study and Cognitive Project files used here were obtained via the MIDUS Colectica Portal (https://midus.colectica.org) under a Data Use Agreement (University of Wisconsin-Madison Institute on Aging, n.d.). The MIDUS programme is described in Brim et al. (2004). Downloaded source files were hash-locked at intake; SHA-256 hashes are recorded in the supplementary code archive for provenance.

The analysis treats word-list immediate recall as the upstream encoding stage and word-list delayed recall as the downstream retention stage. The analysis is a secondary-data worked example, not a preregistered substantive test. The method was deposited on Zenodo on 19 May 2026 before the BTACT analysis was run.

5.2 Analytic frame

Upstream is M2 Word List Immediate Recall total unique recalled (`B3TWLITU`), on the raw word-count scale. Downstream is M2 Word List Delayed Recall total unique recalled (`B3TWLDTU`), also raw. The outcome is M3 BTACT composite cognitive performance (`C3TCOMP`), MIDUS-standardised. Baseline cognitive control is an M2 BTACT non-memory composite, constructed by z-scoring the five non-word-list subtests (backward digit span, category fluency, number series, backward counting, attention-switch reaction time, with reaction time reverse-scored) and averaging. The non-memory composite avoids partialling out the very word-list variance from which BRI is constructed. The full M2 BTACT composite is reported below as an over-controlled sensitivity. Demographic covariates are age and education at M2 (z-scored) and a female-versus-male indicator. Listwise deletion yields N = 2,393.

5.3 Calibration and pre-validation

The calibration regression of M2 delayed on immediate recall returns slope $\hat{\beta}_1 = 0.911$ and $R^2 = .615$. The word-list totals are single observed counts per administration, so a formal internal-consistency reliability is not available. The Pearson correlation between immediate and delayed totals at M2 is $r = .784$ across the calibration sample (N = 4,276), reported to be transparent about the pair rather than as a reliability coefficient. Bootstrap residual-score stability across 500 resamples returns a mean correlation of .9999 between the original residuals and the bootstrap residuals. The nonlinearity test flags a significant quadratic term ($p = 5.5 \times 10^{-16}$) with a small severity ($\Delta R^2 = .006$). The pre-specified p-value rule from Section 4 triggers the quadratic-residual sensitivity reported below. Non-memory baseline composite reliability is acceptable ($\alpha = .72$).

5.4 Profile cleavage

The Section 4 cell logic flags participants in the upper tertile of M2 immediate recall and the lower tertile of the BRI residual. *Rapid-forgetting* participants encoded well but retained below what encoding would predict. *Strong-retention* participants encoded well and retained above what encoding would predict.

Cell	Label	N	Mean M3 BTACT composite
High immediate, low BRI	Rapid-forgetting profile	269	+.25 SD
High immediate, high BRI	Strong-retention profile	390	+.41 SD
Remainder	All other participants	1,734	−.08 SD

Table 6. M3 BTACT composite cognition by BRI cell. N = 2,393. C3TCOMP is the MIDUS-standardised M3 BTACT composite.

Both high-immediate cells outperform the remainder, as expected. The strong-retention cell adds another 0.16 SD over the rapid-forgetting cell at the same encoding level. The BRI residual exposes that retention-specific increment, which composite or branch-level scoring would absorb into the encoding variance. Figure 3 shows the scatter.

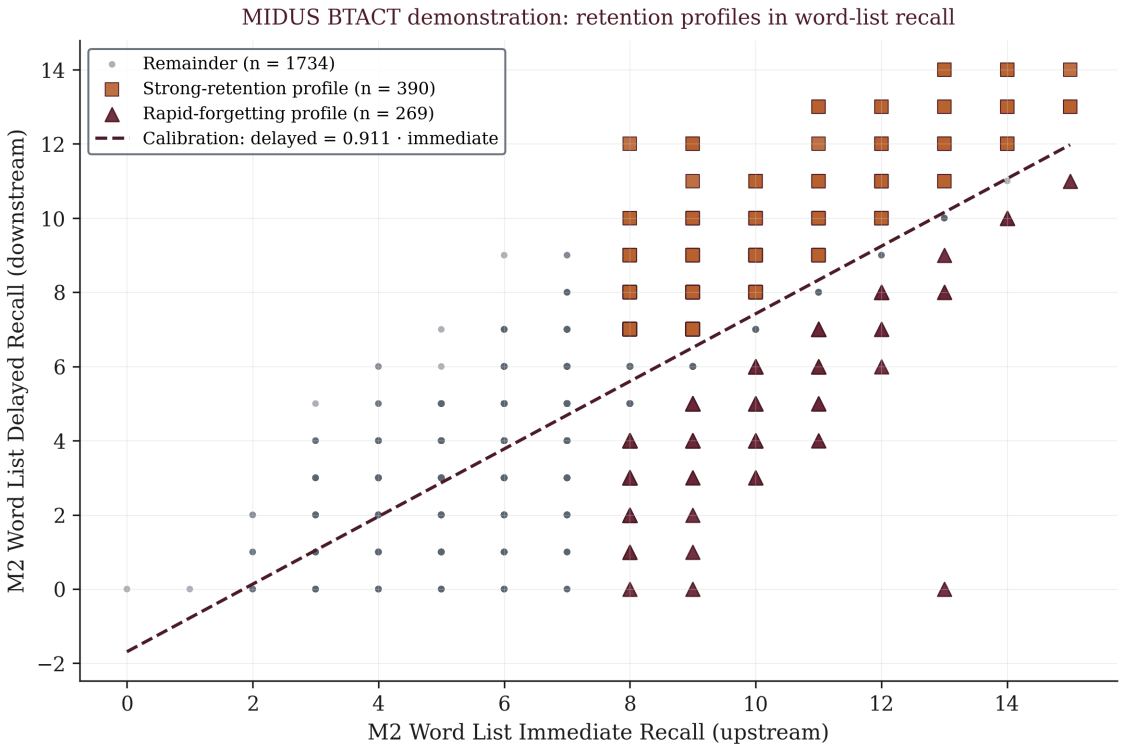

Figure 3. MIDUS BTACT real-data demonstration. M2 Word List Immediate Recall (upstream, raw word count) plotted against M2 Word List Delayed Recall (downstream, raw word count) for the analytic sample (N = 2,393). The dashed line is the calibration regression of delayed on immediate (intercept $-$1.78, slope 0.911, $R^2 = .615$). Filled triangles flag the rapid-forgetting profile, defined as high immediate recall with a low BRI residual (N = 269). Filled squares flag the strong-retention profile, defined as high immediate recall with a high BRI residual (N = 390).

5.5 Primary outcome model and sensitivities

The primary model regresses M3 BTACT composite on the BRI residual, M2 immediate recall, the M2 non-memory baseline composite, and demographics, with HC3 robust standard errors (Zeileis, 2004).

Predictor	Coef.	HC3 SE	p
BRI retention residual (M2, raw words)	.0242	.00541	7.9e-06
Immediate recall (M2, raw words)	.0394	.00433	1.9e-19
Non-memory baseline composite (M2, z)	.669	.0169	less than 1e-200
Age (z)	−.171	.0101	1.1e-60
Sex (F vs M)	.059	.0175	7.0e-04
Education (z)	.061	.00907	1.6e-11

Table 7. MIDUS BTACT primary outcome model. Outcome is M3 BTACT composite (C3TCOMP, MIDUS-standardised). N = 2,393, $R^2 = .661$. *HC3 robust standard errors. BRI residual and immediate recall are on the raw word-count scale; the non-memory composite is the average of five z-scored non-word-list subtests; age and education are z-scored.*

The BRI retention residual is orthogonal to immediate recall by construction. It is not by construction independent of the other covariates, although the estimated effect is robust across the sensitivity models reported next.

Sensitivity	BRI coef.	BRI p
No baseline cognitive control	.0362	4.3e-07
Memory-specific outcome (M3 episodic memory, C3TEM)	.0887	3.5e-18
Full M2 BTACT composite as baseline	−.0090	.096
Diagnostic-triggered quadratic-residual sensitivity	.0259	2.5e-06

Table 8. MIDUS BTACT sensitivity models. Same covariates and HC3 SEs as Table 7.

The first three rows test robustness of the primary BRI effect. Dropping the non-memory baseline control strengthens the BRI coefficient slightly, as expected when a strongly predictive control is removed. Replacing the outcome with M3 episodic memory (the construct most directly downstream of word-list retention) yields a substantially larger BRI effect, consistent with the residual capturing retention-specific variance. The full-composite sensitivity flips sign and approaches the null, which is the expected over-control pattern. The composite contains the very word-list variance from which BRI is constructed, so once the composite enters as a covariate the residual partials out the only variance that distinguishes it from noise. The diagnostic-triggered quadratic-residual sensitivity addresses the small nonlinearity flagged in Section 5.3 and preserves the primary conclusion.

5.6 Interpretation

When a layered measurement architecture is substantively justified, BRI provides a simple observed-score residual estimator that can be diagnosed, reported, and used as a person-level predictor. In MIDUS BTACT, that residual captures retention beyond encoding and predicts later cognition. The pre-specified diagnostic workflow caught a real nonlinearity, the recommended sensitivity addressed it, and the conclusion survived.

6. Generality Demonstration: MIDUS Marital Appraisal

A second MIDUS demonstration shows that the framework generalises beyond ability tests to layered self-report architectures. The architecture is structurally analogous to BTACT. A respondent reports the upstream property of a relationship (perceived spouse strain). They then evaluate the relationship's standing (marital-risk appraisal). The methodological question is whether the variance in appraisal unexplained by strain predicts a later observable outcome (marital rupture between M2 and M3).

The analytic frame is the MIDUS M2 first-marriage subsample. Upstream is M2 spouse strain composite (`B1SSPCRI`). Downstream is M2 marital-risk appraisal composite (`B1SMARRS`). Outcome is M3 marital rupture (separation or divorce between waves). Widowhood is excluded rather than coded as rupture. Covariates are age, sex, education, and marriage duration at M2. The outcome-complete subsample (strain, appraisal, and rupture coding) is N = 1,301 with 76 rupture events. The primary regression sample, after listwise deletion on covariates, is N = 1,296 with the same 76 events (5.9 per cent).

The calibration regression of appraisal on strain returns slope $\hat{\beta}_1 = 0.585$ and $R^2 = .342$. The two MIDUS composites are documented in the MIDUS scale notes. Spouse strain (`B1SSPCRI`) is a four-item index with documented internal consistency in the published MIDUS scale notes (MIDUS scale notes, https://midus.colectica.org). Marital-risk appraisal (`B1SMARRS`) is a published MIDUS index with comparable documentation. Per-item alphas were not recomputed here because the analyses use the MIDUS-supplied composites as the operational units. The pre-validation nonlinearity test flags quadratic coupling and the diagnostic-recommended sensitivity is run. The Cell-4 logic identifies two profiles. The *high-strain / high-residual-appraisal* cell (interpretive gloss, *over-warning*) reported high strain and downstream appraisal above what strain would predict. The *high-strain / low-residual-appraisal* cell (interpretive gloss, *under-warning*) reported high strain and downstream appraisal below what strain would predict. M3 rupture rates differ by approximately an order of magnitude across the two profiles. The high-residual-appraisal cell ruptures at 16.1 per cent. The low-residual-appraisal cell ruptures at 1.4 per cent. The remainder ruptures at 4.0 per cent.

The primary outcome logistic regression returns the BRI residual at OR = 2.60, 95% CI [1.92, 3.54], p less than .001, controlling for strain, age, sex, education, and marriage duration with HC3 robust standard errors (Zeileis, 2004). A diagnostic-triggered quadratic-residual sensitivity preserves the conclusion (OR = 2.57). A Firth-style bias-reduced logistic sensitivity (Firth, 1993; Kosmidis & Firth, 2021) addresses the rare-event concern (76 events) and likewise preserves the conclusion (OR = 2.56). An exploratory agreeableness-by-BRI interaction test is null (OR = 0.91, p = .51), which rules out the interpretation that the residual is just an agreeableness confound.

The marital demonstration is included for generality. The framework was developed for layered cognitive ability tests but the underlying logic (composite scoring confounds an upstream stage and a downstream stage) applies equally to layered self-report architectures. The marital and BTACT analyses are applications of the same scoring procedure to two different architectures within the same dataset. The two analytic samples are drawn from the same MIDUS cohort and likely overlap in participants. They share an upstream and downstream measurement frame, not the substantive constructs.

7. Discussion

7.1 What the procedure contributes

When a layered measurement structure is substantively justified, BRI provides a simple observed-score residual estimator that can be reported, diagnosed, and used as a person-level predictor. In MIDUS BTACT, the residual captured retention beyond encoding and predicted later cognition. The validated workflow that surrounds the residual is the core contribution of this paper. The simulation evidence in Section 3 maps the operating boundary. Section 4 includes a diagnostic workflow which guides applied users on when the residual is and is not appropriate to interpret. The two real-data demonstrations show the procedure in two different layered architectures within the same public dataset. Every reported number is reproduced by the open R pipeline.

7.2 When to use BRI and when not to

Table 9 summarises the decision criteria. The choice between BRI and a structural-equation-model latent residual is the most common applied question. Study 5 maps the operating range over which the two agree empirically.

Criterion	Use BRI	Use SEM latent residual
Indicators per construct	Single observed score	Multiple indicators (three or more)
Identification assumptions	Linear regression only	Plausible single-factor structure per side
Sample size	See Section 3.4 power table	Typically 200 or more (Kline, 2023)
Goal	Person-level score, portable	Latent variance estimate, model fit testing
Measurement error	Conditioning constraint	Modelled explicitly
Reproducibility burden	One regression line	Identification, specification, fit checks
Modelling commitment	Observed score	Latent variable

Table 9. Choosing between BRI and a structural-equation-model latent residual. The two are complementary rather than competing.

7.3 Limitations

Five limitations warrant acknowledgement. The simulations cover specified data-generating assumptions. Departures we have not modelled remain open. These include item-level reliability heterogeneity, missing-data patterns, and multilevel sampling. Both real-data demonstrations come from a single cohort. Substantive claims beyond MIDUS require replication elsewhere. The regression-power analysis assumes a well-specified outcome model. Study 3 Panel C tests positively correlated error under one reliability-coupling combination. Negative or asymmetric correlated error is not characterised. Study 5 maps BRI versus SEM within a sixteen-cell grid. Outside that grid the relative performance should be re-evaluated.

7.4 The broader point

Composite scoring confounds component processes. Residualisation is one principled route to separating them. BRI is a sample-relative observed-score residual, not an absolute trait measurement. That has practical consequences. Applied users should report the calibration sample alongside any inference, treat generalisation to new samples as an empirical question rather than assume it, and report bootstrap residual-score stability via `pre_validate()`. The simulation evidence, the diagnostic workflow, and the two real-data demonstrations together support the procedure as a person-level scoring tool within its documented operating boundary.

8. Open Materials, Software, and Data Availability

All code is open-source under the MIT licence. The R pipeline (`bri-pipeline`) is deposited on Zenodo under concept DOI https://doi.org/10.5281/zenodo.20007669 as the archival version of record. A substantively equivalent code copy accompanies the journal submission. The pipeline includes the estimator (`compute_bri()`), simulation generators (`simulate_branches()`), pre-validation diagnostics (`pre_validate()`), regression-power routines (`power_bri()`), the BRI versus SEM head-to-head simulation (`10_bri_vs_sem.R`), the BTACT demonstration script (`99_BRI_MIDUS_BTACT_demo.R`), the marital demonstration script (`99_BRI_MIDUS_demo.R`), and a user manual. A CHANGELOG accompanying the public open-science deposit lists the simulation-pipeline files that are byte-identical between the prior (v4, 19 May 2026) deposit and the current deposit, with per-file SHA-256 hashes. The supplementary code archive bundled with this submission carries the files themselves; the hash record lives with the public deposit and on the title page.

The simulation studies are fully reproducible from the supplied code and synthetic outputs. The MIDUS demonstrations use public-use MIDUS data obtained via the MIDUS Colectica Portal under terms that prohibit redistribution of participant-level data. Accordingly, the archive includes analysis scripts, variable provenance, sample-flow tables, diagnostics, model outputs, and aggregate result tables, but not raw MIDUS files or participant-level derived MIDUS files. Authorised users can reproduce the MIDUS analyses by obtaining the relevant MIDUS files from the MIDUS Colectica Portal and running the supplied scripts locally.

The BTACT and marital demonstrations were not preregistered as substantive analyses. The methodological procedure was deposited on the open-science archive on 19 May 2026 prior to the demonstration analyses being run, which is the relevant provenance claim for a methods demonstration.

Declarations

Funding. None.

Conflict of interest. None.

Ethics approval. Not applicable. Simulation studies use Monte Carlo synthetic data. The Section 5 and Section 6 analyses use public-use de-identified MIDUS data. Original ethical approval is held by the MIDUS investigators.

Independent scholarship. This manuscript reports independent personal research by the author. The author is concurrently completing an MSc in Psychology by distance learning at the University of Northumbria. The present work is independent of that programme and is not submitted toward any degree.

Author contributions. Sole-author manuscript.

Data availability. Data underlying the analyses are publicly available but require an institutional or personal data-access agreement. Wave-level survey and disposition data are available via the MIDUS Colectica Portal (https://midus.colectica.org). SHA-256 hashes for the locked simulation-pipeline files are recorded in CHANGELOG.md inside the supplementary archive. The simulation studies are fully reproducible from the supplied code and synthetic outputs.

Use of generative AI tools. The author used Claude (Anthropic) and Codex (OpenAI) for technical assistance with the manuscript and the accompanying R pipeline, including content-checking edits, citation formatting, and literature identification. All cited references were verified by the author against primary sources. The estimator, theoretical framing, and interpretation are the author's, who bears full responsibility for the content.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The 'Reading the Mind in the Eyes' Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Boullineau, E. (2026a). *Seeing the Signal, Missing the Meaning: A Branch Dissociation Hypothesis of Affective Perception and Misattribution in Emotional Intelligence* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.19958635

Boullineau, E. (2026b). *bri-pipeline: R code accompanying "Residualised Scoring for Component-Process Isolation"* [Software]. Zenodo. https://doi.org/10.5281/zenodo.20007669

Brim, O. G., Ryff, C. D., & Kessler, R. C. (Eds.). (2004). *How healthy are we? A national study of well-being at midlife*. University of Chicago Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

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

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1(4), 387–401. https://doi.org/10.2307/1907330

Huber, P. J. (1981). *Robust statistics*. Wiley. https://doi.org/10.1002/0471725250

Kosmidis, I., & Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1), 71–82. https://doi.org/10.1093/biomet/asaa052

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

Lachman, M. E., Agrigoroaei, S., Tun, P. A., & Weaver, S. L. (2014). Monitoring cognitive functioning: Psychometric properties of the Brief Test of Adult Cognition by Telephone. *Assessment*, 21(4), 404–417. https://doi.org/10.1177/1073191113508807

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68(5), 304–305. https://doi.org/10.1037/h0025105

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304), 993–1010. https://doi.org/10.1080/01621459.1963.10480682

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The perils of partialling: Cautionary tales from aggression and psychopathy. *Assessment*, 13(3), 328–341. https://doi.org/10.1177/1073191106290562

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). Basic Books.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). *Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT): User's manual*. Multi-Health Systems.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2008). Emotional intelligence: New ability or eclectic traits? *American Psychologist*, 63(6), 503–517. https://doi.org/10.1037/0003-066X.63.6.503

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410601

Nowicki, S., Jr., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior*, 18(1), 9–35. https://doi.org/10.1007/BF02169077

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sleep, C. E., Lynam, D. R., Hyatt, C. S., & Miller, J. D. (2017). Perils of partialing redux: The case of the Dark Triad. *Journal of Abnormal Psychology*, 126(7), 939–950. https://doi.org/10.1037/abn0000278

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research*, 2(1), 21–33.

Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3), 531–543. https://doi.org/10.1016/j.jhealeco.2007.09.009

University of Wisconsin-Madison Institute on Aging. (n.d.). *Midlife in the United States (MIDUS) Project 1 Survey Data, Waves 1–3* [Data set]. MIDUS Colectica Portal. https://midus.colectica.org

University of Wisconsin-Madison Institute on Aging. (n.d.). *Midlife in the United States (MIDUS) Project 3 Cognitive Project (Brief Test of Adult Cognition by Telephone), Waves 2–3* [Data set]. MIDUS Colectica Portal. https://midus.colectica.org

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10), 1–17. https://doi.org/10.18637/jss.v011.i10