Residualised Scoring for Component-Process Isolation: A Regression-Based Framework, with Application to Ability Emotional Intelligence

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

Preprint, May 2026. Under peer review at Royal Society Open Science (*Psychology and Cognitive Neuroscience Section*). *The R pipeline and reproducibility materials accompany this deposit.*

Abstract

Many cognitive-ability tasks have a layered architecture. An upstream stage detects a signal. A downstream stage reasons from it. Composite scoring confounds the two, and branch-level subscores inherit the upstream signal rather than isolating the downstream one. Branch-Residualised Interpretation (BRI) addresses the problem directly. The procedure regresses the downstream score on the upstream score across a calibration sample and treats the residual as the person-level downstream score. The residual is the first step of the Frisch–Waugh–Lovell theorem, and no algebraic novelty is claimed. The contribution is a simulation-based map of when this person-level residual functions as a defensible score, paired with a pre-validation diagnostic workflow, regression-power guidance, and an open R pipeline that regenerates every reported number. Six Monte Carlo studies establish the operating boundary. Recovery sits between .77 and .94 once component reliabilities reach .80. It degrades under low downstream reliability and unmodelled nonlinear coupling, and is essentially flat across the heavy-tailed error conditions examined when component reliabilities are held fixed. A Lord-style disattenuation correction is reported only as a sensitivity check. Ability emotional intelligence is the illustrative application. The framework generalises to working memory, theory of mind, and other layered cognitive abilities.

Keywords: psychological measurement, residualisation, regression, component processes, ability tests, simulation methods, Frisch–Waugh–Lovell

1. The Component-Process Entanglement Problem

In psychological ability testing, component processes are often conceptually separable but operationally entangled. Consider three cases. A working-memory span task asks the participant to encode the items, hold them, and reproduce them, so an encoding error and a maintenance failure produce the same observable mistake. A theory-of-mind vignette asks the participant to read the protagonist's mental state and then reason forward from it. Read the state correctly but reason badly, and the composite score cannot tell you so. An ability test of emotional intelligence asks the participant to register an affective signal and then attribute its cause. A high score can reflect strong perception, strong attribution, or both.

The problem is general. When a task has a layered structure, with downstream reasoning conditioned on upstream detection, the composite score confounds the two stages. Investigators interested specifically in the downstream component need a scoring procedure that isolates the variance of interest. The motivation can be theoretical (a hypothesis about which stage drives an interpersonal outcome), psychometric (a need to disentangle correlated latent factors), or applied (an intervention targeting one stage). Composite scoring will not deliver it. Branch-level subscores, where they exist, do not deliver it either, because the downstream subscore inherits the variance of the upstream input.

The procedure developed here is *Branch-Residualised Interpretation* (BRI), a regression-based scoring approach. The downstream score is regressed on the upstream score across the calibration sample, and the residual is treated as the construct of interest. It carries the variance in the downstream measure that cannot be linearly explained by the upstream measure. The regression operates on person scores rather than items, so the procedure is an *estimator* applied to any pair of suitable upstream and downstream task scores. The task battery remains the instrument. BRI is the scoring framework that separates its components. Figure 1 illustrates the problem and the residual's role in resolving it.

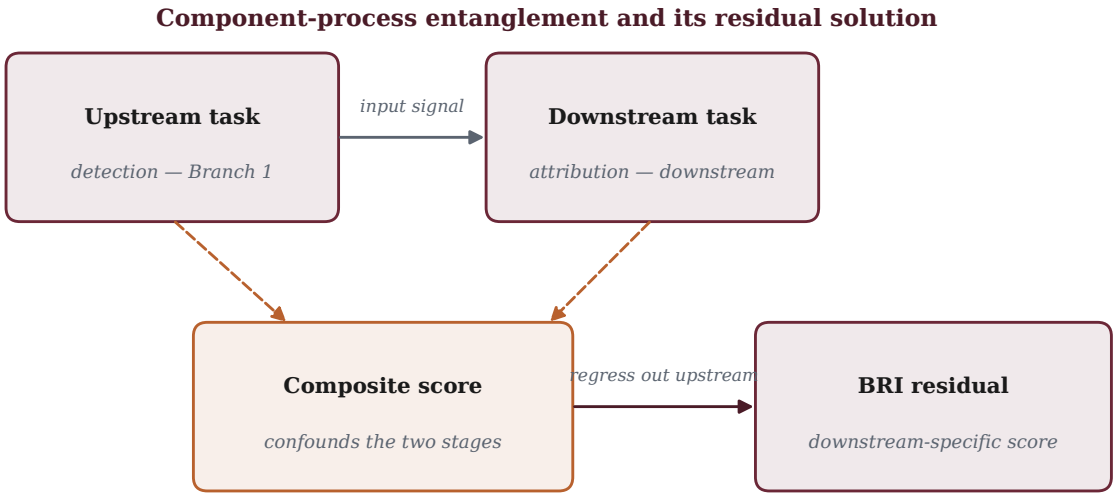

Figure 1. Component-process entanglement and the role of residualisation. The upstream task supplies an input signal that the downstream task must reason over. The shaded centre node represents a composite score that conflates variance from both stages. The Branch-Residualised Interpretation residual (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.

The worked case throughout is ability emotional intelligence. A recent theoretical formulation, the Branch Dissociation Hypothesis (Boullineau, 2026a), motivates the residualisation directly. In the four-branch model (Mayer & Salovey, 1997; Mayer, Salovey, & Caruso, 2008), Branch 1 (perceiving emotions) supplies an upstream signal that downstream interpretation must then attribute. The hypothesis predicts that high uptake combined with miscalibrated attribution under self-relevant ambiguity is interpersonally distinctive in a way that aggregate or branch-level scoring obscures. Testing the prediction empirically requires the kind of residualised scoring this paper develops. Ability EI is therefore both the substantive case for which BRI is needed and the worked example through which the general framework can be evaluated.

The estimator's relationship to classical econometrics is developed in Appendix A. That tradition runs from the Frisch–Waugh–Lovell theorem through factor-residual decomposition in empirical finance, two-stage residual inclusion in health econometrics, and layered detection-then-adjudication models in operational risk. Against that cross-disciplinary backdrop, the present paper does one specific thing. It characterises, by simulation, when the residual functions as a defensible person-level *score* in psychological ability data, and supplies the diagnostic and reproducibility infrastructure to check whether the necessary conditions hold in any given dataset.

The paper is organised as follows. Section 1.1 specifies the scope and intended audience. Section 2 specifies the estimator in general terms, situates it against six pre-existing methods, and then presents its EI-specific instantiation. Section 3 (Study 1) demonstrates that BRI recovers the latent downstream variance under ideal classical-test-theory assumptions. Section 4 (Study 2) characterises behaviour under unequal component reliabilities and develops a corrected estimator. Section 5 (Studies 3 through 5) stress-tests the residualisation against three realistic departures from idealised data. These are ceiling effects, heavy-tailed measurement error, and nonlinear coupling. Section 6 (Study 6) provides a regression-power table for detecting downstream effects on outcome variables. Section 7 walks through a worked example end to end with code. Section 8 documents the open-source R pipeline. Section 9 discusses generality, failure modes, and applicability beyond ability emotional intelligence. Appendix A develops the formal connection to Frisch–Waugh–Lovell residualisation and the cross-disciplinary tradition.

1.1 Scope and audience

This paper is methodological. Its primary readership is researchers working on the individual-differences measurement of layered cognitive abilities: working memory, theory of mind, executive function, ability emotional intelligence, and analogous domains in which component processes need to be separated for psychometric or theoretical reasons. The paper is *not* a contribution to clinical psychology. The ability-EI application is framed throughout as a cognitive-ability measurement problem rather than a clinical-assessment one. The Branch Dissociation Hypothesis (Boullineau, 2026a) is a theoretical claim about individual differences in cognitive ability rather than a clinical taxonomy.

Investigators in adjacent applied fields will recognise the procedure. Empirical finance uses factor-residual return decomposition (Fama & MacBeth, 1973; Fama & French, 1993) to separate idiosyncratic from systematic variance. Health econometrics uses two-stage residual inclusion (Terza, Basu, & Rathouz, 2008) to correct for endogeneity. Operational risk modelling routinely deploys layered detection-then-adjudication architectures, with transaction screening followed by analyst review, or alert generation followed by enhanced due diligence. All share the upstream-downstream structure that the present paper formalises for psychological ability data. Appendix A makes these connections explicit and locates the procedure within the residualisation tradition that begins with Frisch and Waugh (1933) and Lovell (1963).

2. The Procedure

2.1 General specification

Consider a layered task with two component scores per person: an upstream score X capturing detection or input-side ability, and a downstream score Y capturing interpretation or reasoning conditioned on that detection. The investigator's interest is in the variance of Y that is not explained by X. Equivalently, in the downstream ability that remains once the upstream score has been taken into account.

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

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

Two properties deserve emphasis. First, the slope $\hat{\beta}_1$ has substantive content of its own. It captures the average degree to which upstream ability predicts downstream ability across the sample. Tightly coupled processes produce steep slopes. Loosely coupled processes produce shallow ones. Second, the residual is a *between-person* score. BRI ranks participants by how far their downstream score sits above or below what their upstream score would predict.

BRI is computed in six steps.

Step 1. Measure the upstream component with a task or composite designed to isolate the input-side process. The score may be a proportion correct, a *d*-prime, or any standard psychometric score. The choice of metric leaves the residualisation untouched but shapes how the residual should be interpreted.

Step 2. Measure the downstream component with a task or composite designed to demand the downstream process while holding the upstream demand approximately constant.

Step 3. Across the sample, fit the linear regression of Y on X.

Step 4. Extract the residual; this is the participant's BRI score.

Step 5. Pre-validate. Before substantive interpretation, users should report the diagnostics supported by their design. At minimum, this should include upstream and downstream reliability evidence (such as internal-consistency or test–retest coefficients), the upstream–downstream scatter plot, the fitted slope and R^2, a nonlinearity check, and bootstrap stability of residual rankings. Where multiple task forms or indicators are available, cross-task residual correlation and hold-out factor-structure checks should also be reported. The R pipeline runs whichever of these diagnostics the supplied inputs allow, and clearly flags any that were skipped. Section 9 returns to these.

Step 6. Use the residual as a predictor in subsequent analyses. Section 6 develops the regression-power analysis for this step.

2.2 The ability-EI instantiation

The general procedure becomes the *Branch-Residualised Interpretation* of the Branch Dissociation Hypothesis when the upstream score is a branch-level measure of affective signal detection (e.g., the perceiving subscale of the MSCEIT, Mayer, Salovey, & Caruso, 2002; the Reading the Mind in the Eyes Test, Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; the Diagnostic Analysis of Nonverbal Accuracy, Nowicki & Duke, 1994; for review of ability-EI assessment more broadly, see MacCann & Roberts, 2008, and Roberts, MacCann, Matthews, & Zeidner, 2010), and the downstream score is a measure of situated causal attribution under self-relevant ambiguity (the Self-Referential Attribution Task, or SRAT, of Boullineau, 2026a, is the elicitation paradigm developed for this purpose).

The hypothesis predicts that the residual variance in situated attribution unexplained by detection is the construct of theoretical interest. Empirical testing of this prediction requires the residualisation procedure of Section 2.1. The worked example in Section 7 walks through this instantiation end to end.

2.3 Prior Art and the Contribution of BRI

Residualisation is a familiar tool in psychological measurement. A reviewer's first reasonable question is what BRI adds beyond methods that already exist. The estimator developed here sits within a well-established tradition of residual-based scoring and component decomposition. The contribution is not the use of a regression residual, which is decades old. It is the systematic evaluation of when, and under what conditions, a *person-level* regression residual usefully isolates downstream component variance in psychological ability data. To make the contribution precise, this section situates BRI against six relevant alternatives.

Semi-partial correlation and partial regression coefficients. The semi-partial correlation between an outcome and Y controlling for X quantifies the association between the outcome and the part of Y orthogonal to X. Its square corresponds to the incremental outcome variance explained by adding Y as a predictor after X (Cohen, Cohen, West, & Aiken, 2003). BRI produces the same orthogonalised quantity at the *person level*. The simple correlation between the BRI residual and an outcome equals the semi-partial correlation between that outcome and Y controlling for X. The two are therefore mathematically related but answer different questions, and are suited to different applications. The semi-partial correlation summarises an association across a sample. BRI is a per-participant score. It supports between-person comparisons, profile identification (for example, the high-detection low-attribution target cell in the worked example), and use as a single predictor across multiple outcome models.

Residualised change scores. Cronbach and Furby (1970), with subsequent elaboration by Steyer, Eid, and Schwenkmezger (1997), use a regression residual to capture change in Y relative to baseline X within the same individual over time. BRI is structurally identical, in that it is a residual from the linear regression of Y on X. The setting differs. Change scores apply within a single task across time. BRI applies between two tasks at one point in time. Two well-known critiques of residualised change scores carry over to BRI. The first is that they depend on the regression slope being a population property rather than an individual one. The second is that they inherit reliability problems of both inputs. Study 2 of this paper directly addresses one such critique through the disattenuation correction. Studies 3 through 5 address the others, namely nonlinearity, heavy tails, and ceiling effects.

Structural equation modelling with a latent-residual factor. A more theoretically rigorous alternative is to model X and Y as latent variables and estimate the latent-Y variance unique of latent-X using a bifactor or residual-factor specification (Bollen, 1989; Rosseel, 2012). This approach handles measurement error explicitly, identifies the residual at the latent level, and supports formal goodness-of-fit testing. BRI's relationship to this approach is one of convenience versus rigour. BRI operates on observed scores, requires no model specification beyond a linear regression, runs in a single line of R, and produces a person-level score. The SEM approach requires identification constraints, multiple indicators per latent construct, stable model identification, and typically larger samples than the simple regression. It also requires a structural commitment that the applied researcher may not be willing or able to make. The two are not in competition. BRI is the appropriate tool when the investigator wants a portable scoring procedure and acceptable reliability is plausible. A latent-residual SEM is appropriate when the investigator wants to estimate latent components directly, and has the sample size, indicator coverage, and modelling commitment to do so.

Two-stage residual inclusion (2SRI) and instrumental variables. In econometrics, 2SRI is used to correct for endogeneity by regressing an endogenous variable on an exogenous instrument, extracting the residual, and including it as a covariate in a structural model (Terza, Basu, & Rathouz, 2008). BRI's procedure is

mechanically similar. It runs a regression and then takes a residual. The inferential goal, however, is different. 2SRI requires a valid instrument and yields a causal estimate of an endogenous effect. BRI uses a measured upstream score and yields a descriptive variance decomposition, not a causal effect. Section 2.5 sets out the interpretive caveat. BRI is a scoring procedure, not a causal estimator. That distinction separates it clearly from 2SRI.

Hierarchical regression with the upstream score as a covariate. Consider the goal of predicting a single outcome from "downstream ability controlling for upstream". The most direct analysis is to include both X and Y as predictors of the outcome (Cohen et al., 2003). The coefficient on Y in such a regression is, by algebraic identity, equivalent to the coefficient on the BRI residual in a regression of Outcome $\sim X +$ BRI. This algebraic identity is itself an instance of the Frisch–Waugh–Lovell theorem, treated formally in Appendix A. BRI's contribution in this case is not statistical but practical. A person-level residual can be examined in diagnostic plots, used across multiple outcome models without re-fitting the regression of Y on X each time, used to identify profiles of interest in the sample, and shared across analyses as a single column of data. When only one outcome is at stake, hierarchical regression suffices. When multiple outcomes, profile identification, or diagnostic visualisation are involved, the person-score formulation has practical advantages and supports applications the coefficient-based approach does not.

Factor-score residuals and specific-factor scoring. Factor analysis can produce specific-factor scores that capture variance unique to each indicator (McDonald, 1999). The factor-scoring approach is theoretically grounded in a measurement model and supports inferences about latent structure. BRI is model-free at the latent level. It is consequently more portable to data where a full factor model has not been or cannot be estimated. The trade-off is clear. BRI gives up the latent-level grounding in exchange for that portability.

The Frisch–Waugh–Lovell theorem. The mathematical operation of residualising Y on X and using the residual in place of Y in subsequent regressions is the first step of the Frisch–Waugh–Lovell theorem (Frisch & Waugh, 1933; Lovell, 1963), a classical econometrics result. The theorem establishes that two-step residualisation reproduces multiple regression coefficients exactly. The present paper does not claim novelty for the residualisation operation itself. The contribution is the systematic evaluation of when the FWL residual functions as a defensible person-level *score* in psychological ability data. The diagnostic and reproducibility infrastructure needed to determine whether the necessary conditions hold accompanies the estimator. Appendix A develops the FWL connection formally. It also locates BRI within the cross-disciplinary tradition of factor-residual decomposition in empirical finance, two-stage residual inclusion in health econometrics, and layered risk-screening models in operational practice.

What BRI adds across these alternatives. The residualisation operation itself is not novel. The methodological contribution is the systematic simulation-based specification of when an observed-score residual functions as a defensible *person-level score* for downstream component variance in psychological ability data. The integrated package has five elements. First, the six-step scoring procedure of Section 2.1. Second, the simulation-based boundary characterisation of Sections 3 through 5, covering variance recovery, reliability asymmetry, ceiling effects, heavy-tailed error, and nonlinear coupling. Third, the regression-power guidance of Section 6. Fourth, the pre-validation diagnostic workflow of Step 5. Fifth, the open-source reproducible implementation documented in Section 8. None of these components alone is novel. Their integration into a single workflow with documented operating boundaries is the contribution. Subsequent sections support this framing rather than the underlying algebra.

2.4 Sample-relative calibration of BRI scores

An interpretive point critical to applied use. A BRI score is not an absolute trait measurement. It is a *sample-relative residual*. The slope and intercept used to produce the residual are estimated from the sample at hand. The residual for any given participant therefore depends on which sample they are embedded in. The residual ranks participants by how much higher or lower their downstream score is than the *fitted* upstream-downstream relation in the calibration sample would predict.

Three implications follow.

1. **BRI is suited to between-person comparisons within a defined sample** (e.g., a study cohort, an assessment cohort, a recruitment pool, or a validation sample) rather than to absolute scoring across heterogeneous populations. The same participant assessed in two samples with different upstream–downstream coupling will receive two different BRI scores.
2. **For applied scoring across studies, the slope and intercept should be estimated in a calibration sample and then applied to a target sample**, rather than re-estimated in each new sample. Stability of the calibration regression is therefore important, and cross-sample portability should be tested empirically rather than assumed.
3. **In small samples the slope is unstable, which directly destabilises BRI rankings.** Section 6 provides regression-power guidance for the downstream effect of interest. Users should additionally inspect the standard error of the calibration slope before drawing person-level inferences.

The R pipeline accompanying this paper provides bootstrap stability diagnostics for individual residual rankings via `pre_validate()`. Investigators applying BRI for individual-level inference should report bootstrap rank-stability alongside the substantive analysis.

2.5 What residualisation does and does not establish

Before proceeding to the simulations, an interpretive caveat applies to BRI and to every method in Section 2.3 that derives a residual.

The strengths of regression residualisation are well understood: it isolates the variance in one variable that cannot be linearly explained by another. Its failure modes are also well understood. When component reliabilities are low or unequal, the residual carries reliability problems of its own (Lord, 1967). When the relationship between the components is nonlinear, the linear residual is biased (Cohen et al., 2003). When measurement error is heavy-tailed, classical OLS estimates of the slope can be unstable, motivating robust alternatives (Huber, 1981). The simulations of Sections 3 through 6 carry out the systematic evaluation of when this estimator behaves well, when it fails, and what users should do in either case under the conditions actually encountered in psychological ability data.

A causal caveat. The upstream/downstream language used throughout this paper describes the *task architecture* (one task supplies an input that another reasons over), not a causal estimate. BRI is a scoring residual, not evidence that the upstream component causally determines the downstream component. A large residual, a systematic residual pattern across participants, or a non-zero regression slope is consistent with several causal structures, including genuine downstream-specific ability, common-cause variance, or mutual influence. Disambiguating among them requires additional evidence beyond the residualised score.

3. Study 1: Variance Recovery under Ideal Conditions

3.0 Reporting conventions for Sections 3–6

All entries reported in Tables 1–6 and Table 8 are means across replications. For the *recovery coefficients* of Tables 1–6 (correlations between the BRI estimate and the true downstream-specific component), cell means and Monte Carlo standard errors are written by each driver script to the `results/` directory as supplementary CSVs. The MCSEs are below approximately 0.01 in the recovery cells at the stated replication counts (1,000 replications per cell in Study 1; 500 per cell in Studies 2–5). For the *power estimates* in Table 8, the Monte Carlo standard error is the binomial standard error $\sqrt{p(1-p)/R}$; at $R = 500$ this ranges up to approximately 0.022 in the most variable cells (those closest to $p = .50$, where the binomial standard error is largest). Cell-by-cell binomial MCSEs are written alongside the power point estimates in the supplementary CSV. Investigators who require tighter precision in the power table can re-run `04_power.R` with `n_replications = 2500` or higher; the binomial standard error scales as $1/\sqrt{R}$, so 2,500 replications keeps the worst-case MCSE at or below 0.01.

3.1 Question

Does the residual recover the latent downstream-specific component under classical-test-theory assumptions: normally distributed true abilities, independent normal measurement error, linear coupling, and no ceiling effects?

3.2 Design

For each of $N = 500$ simulated participants, draw a true upstream ability $T_X \sim \mathcal{N}(0, 1)$ and a true downstream ability T_Y correlated with T_X at $\rho \in \{0.3, 0.5, 0.7\}$. The true downstream-specific component is the part of T_Y orthogonal to T_X. Observed scores are generated by adding independent normal measurement error scaled to give equal upstream and downstream reliabilities $r_{XX} = r_{YY} \in \{0.70, 0.80, 0.90\}$ (Study 2 relaxes this constraint). Apply BRI to the observed scores; compute the correlation between the extracted residual and the true downstream-specific component. Replicate 1,000 times per cell.

3.3 Results

	True coupling ρ		
Reliability r_{XX}	**0.30**	**0.50**	**0.70**
0.90	.94	.92	.87
0.80	.88	.84	.77
0.70	.82	.77	.68

Table 1. Recovery coefficient (correlation between the BRI residual and the true downstream-specific variance), 1,000 replications per cell.

At reliabilities of .80 or higher, BRI recovers downstream-specific variance at .77 to .94, depending on the underlying coupling. At a reliability of .70 the range is .68 to .82. Recovery degrades modestly with increasing ρ, as expected: residualisation has less unique variance to recover when the components share more underlying structure. BRI does not require near-orthogonality of the components to function, but its effective ceiling falls as coupling tightens. The central practical implication is that component reliabilities of .80 or higher are recommended for substantive applications.

4. Study 2: Reliability Asymmetry and the Corrected Estimator

4.1 Question

Does the residual behave differently when the two component reliabilities are unequal? Specifically, does the linear regression slope absorb so much variance that the residual systematically underestimates downstream-specific variance, or vice versa?

4.2 Design

Hold the true coupling at $\rho = 0.50$; vary the upstream reliability $r_{XX}^{(\text{up})}$ and the downstream reliability $r_{YY}^{(\text{down})}$ independently across $\{0.60, 0.70, 0.80, 0.90\}$. For each cell, run 500 replications and compute recovery against the true downstream-specific component. Run twice: once with the uncorrected residual, once with a Lord-style disattenuation correction in which the regression slope is divided by the upstream reliability before the residual is extracted. Because scaling the slope shifts the implied intercept, the corrected intercept is recomputed at the sample means rather than retained from the uncorrected fit:

$$\hat{\beta}_{1,\text{cor}} = \hat{\beta}_1 \,/\, r_{XX}^{(\text{up})}, \qquad \hat{\beta}_{0,\text{cor}} = \bar{Y} - \hat{\beta}_{1,\text{cor}}\,\bar{X}, \qquad \widehat{\text{BRI}}_{i,\text{cor}} = Y_i - (\hat{\beta}_{0,\text{cor}} + \hat{\beta}_{1,\text{cor}} X_i).$$

(If scores are mean-centred before residualisation the intercept-recomputation step is unnecessary; the R pipeline handles both cases.)

4.3 Results

	Upstream reliability $r_{XX}^{(\text{up})}$			
Downstream reliability $r_{YY}^{(\text{down})}$	**0.60**	**0.70**	**0.80**	**0.90**
0.90	.88	.89	.91	.92
0.80	.82	.83	.84	.86
0.70	.77	.77	.78	.79
0.60	.70	.71	.72	.72

Table 2. Uncorrected recovery, 500 replications per cell.

	Upstream reliability $r_{XX}^{(\text{up})}$			
Downstream reliability $r_{YY}^{(\text{down})}$	**0.60**	**0.70**	**0.80**	**0.90**
0.90	.85	.88	.90	.92
0.80	.80	.82	.84	.85
0.70	.75	.76	.78	.79
0.60	.68	.70	.71	.72

Table 3. Corrected recovery (Lord disattenuation applied to the regression slope), 500 replications per cell.

Recovery is governed almost entirely by the *downstream* reliability: rows of both tables are nearly flat, with upstream reliability contributing only secondary effects. The disattenuation correction *reduces* recovery rather than improving it in most cells of the parameter space examined here.

The practical implication is unambiguous: the uncorrected estimator should be used as the primary analysis in typical ability-test data with reliabilities in the .70–.90 range. The disattenuation correction is reported here as a sensitivity check, not as a recommended default. Its use is restricted to settings where an investigator has independent reason to suspect that restricted-range sampling has biased the regression slope downward. That is a specific identification problem rather than a general improvement. Readers interpreting Table 3 should note that the slight degradation of recovery is the expected price of attempting a correction that targets a problem absent from the simulation conditions in this study.

5. Studies 3 through 5: Robustness under Non-Ideal Conditions

The simulations of Sections 3 and 4 use idealised data: normal latent abilities, independent normal measurement error, linear coupling. Real psychological ability data violate these assumptions in three predictable ways. Studies 3, 4, and 5 stress-test BRI under each in turn.

5.1 Study 3: Ceiling effects

Ability tests in highly verbal samples often produce right-skewed score distributions: a substantial fraction of participants score near the maximum. We simulate this by compressing the upper tail of both X and Y scores, applying a piecewise transform that retains 20% of variation above a threshold quantile. We vary the proportion of the sample subject to compression across $\{0.00, 0.10, 0.20, 0.30, 0.50\}$, with 0.00 serving as the uncompressed baseline.

Compression	0.00	0.10	0.20	0.30	0.50
Recovery	.85	.83	.82	.80	.77

Table 4. Recovery under ceiling-effect compression of the upper tail, 500 replications per cell, $r_{XX} = .80$.

Recovery degrades modestly: at 50% compression, recovery falls from .85 to .77. BRI is robust to ceiling effects of the kind common in practice, but applied users should still inspect their score distributions and, where ceiling effects are severe, consider transforming scores onto a logit or probit scale before residualisation.

5.2 Study 4: Heavy-tailed measurement error

Ability data often contain occasional extreme scores from participants who are disengaged, misunderstand the instructions, or are guessing. These produce heavy-tailed measurement error that violates the OLS assumption of normal residuals. We replace the Gaussian error term with a t-distributed error, varying degrees of freedom across $\{3, 5, 10, 30\}$ and a Gaussian baseline. Lower df produces heavier tails.

Error distribution	t_3	t_5	t_{10}	t_{30}	Gaussian
Recovery	.85	.84	.84	.85	.84

Table 5. Recovery under heavy-tailed measurement error, 500 replications per cell, $r_{XX} = .80$. Monte Carlo standard errors are below .002 in every cell.

Recovery is essentially flat across the heavy-tail conditions examined. From t_3 through the Gaussian baseline, mean recovery sits in the .84 to .85 band, indistinguishable within Monte Carlo standard error. The result is initially surprising. It reflects the simulation's error-generation mechanism: the t-distributed error draws are first standardised to unit variance and then scaled to give the target measurement-error variance implied by $r_{XX} = .80$. This holds the effective signal-to-noise ratio constant across error distributions, so the OLS slope of Y on X is little disturbed by occasional extreme observations and the residual ranking is preserved.

A supplementary head-to-head comparison of OLS against M-estimation (Huber, 1981; `compute_bri(..., robust = TRUE)`, via `MASS::rlm`) is written to the `results/` directory as `study4_heavytail_ols_vs_robust.csv`. The two estimators produce recovery values that differ by less than .001 across the heavy-tail grid. Robust regression is therefore not required under the conditions modelled here.

Two caveats apply. First, this finding holds within the simulation's design, where measurement error is symmetrically heavy-tailed and component reliabilities are fixed. Real ability data can carry asymmetric outliers (for example, careless responders who underperform on most items but not all) and reliability that itself varies with the contaminating process. BRI's behaviour under those richer conditions is not characterised here. Second, the simulation does not test settings where extreme observations are leverage points rather than vertical outliers. Robust regression remains the appropriate first response in any applied dataset showing extreme-value diagnostic flags.

5.3 Study 5: Nonlinear coupling

The OLS residual assumes a linear relationship between X and Y. If the true relationship is nonlinear, the linear residual misses the curvature and absorbs detection-related variance that should have been removed. We simulate quadratic coupling by generating $T_Y = \rho T_X + \beta_q(T_X^2 - 1) + e$, varying the quadratic coefficient β_q across $\{0.00, 0.10, 0.20, 0.30, 0.40\}$, with $\beta_q = 0.00$ serving as the linear baseline. The truly downstream-specific component is the noise term e alone (orthogonal to all functions of T_X).

Quadratic coupling β_q	0.00	0.10	0.20	0.30	0.40
Recovery of e	.84	.83	.80	.73	.64
r between residual and X^2	.00	.11	.22	.33	.44

Table 6. Recovery of the truly orthogonal downstream component under nonlinear coupling, and spurious upstream-quadratic variance absorbed into the residual, 500 replications per cell, $r_{XX} = .80$.

Two failure modes appear together. As nonlinear coupling grows, BRI's linear residual misses an increasing share of the truly downstream-specific variance (recovery falls from .84 to .64), and simultaneously absorbs an increasing share of upstream-quadratic variance into the residual. The residual is no longer attribution-specific; it is contaminated with a function of detection. Applied users should test for nonlinearity before relying on the linear residual: include X^2 as an additional predictor and check whether its coefficient is significant, or fit a generalised additive model and inspect the smooth term (Wood, 2017). If nonlinearity is detected, the estimator should be extended (e.g., to a polynomial regression or a kernel-based residualisation) rather than applied as-is. The R pipeline includes a `test_nonlinearity()` diagnostic.

5.4 Practical guidance from the robustness studies

The three studies together identify the conditions under which BRI can be applied as a primary analysis and the conditions under which it requires extension or supplement:

Condition	BRI suffices?	Recommended action
Ceiling effects up to approximately 30%	Yes	Inspect distributions; transform if severe
Heavy-tailed measurement error (t_3 through Gaussian, with standardised reliabilities)	Yes	Standard OLS adequate; supplementary OLS-vs-robust comparison differs by less than .001 across the grid
Asymmetric outliers or extreme leverage points (not modelled here)	No	Use robust regression and flag extreme cases at the screening stage
Linear or near-linear coupling	Yes	Standard OLS adequate
Detected nonlinear coupling	No	Extend procedure or report sensitivity

Table 7. Practical guidance from Studies 3 through 5.

6. Study 6: Regression Power for Detecting Downstream Effects

6.1 Question

The substantive use of BRI is to predict downstream variance in outcome variables: relationship satisfaction, interpersonal conflict, occupational performance, or analogues in other domains. What sample size is needed to detect a downstream effect of a given magnitude on an outcome?

6.2 Design

Generate samples in which an outcome variable Y_{out} depends on the truly downstream-specific component with standardised effect $\beta_{\text{BRI}} \in \{0.15, 0.20, 0.30, 0.40\}$, spanning the range from small to medium-large by conventional benchmarks (Cohen, 1988). Apply BRI to extract the residual, fit $Y_{\text{out}} \sim X + \text{BRI}$, and test the BRI coefficient at $\alpha = .05$. The raw downstream score is omitted from the regression because BRI is by construction a linear function of X and the raw downstream score, producing a rank-deficient design otherwise. Replicate 500 times per cell across $N \in \{100, 250, 500, 1000, 2000\}$ at $r_{XX} = .80$.

6.3 Results

	Sample size N				
Effect size 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 8. Power for the BRI coefficient at $\alpha = .05$, 500 replications per cell.

For small effects ($\beta_{\text{BRI}} = .15$), $N \approx 500$ is needed for 80% power. For small-to-medium effects ($\beta_{\text{BRI}} = .20$), $N = 250$ is approximately sufficient. Re-analyses of published datasets with smaller individual study sizes may benefit from meta-analytic combination across studies.

7. Worked Example: Ability EI End to End

A self-contained worked example demonstrates the scoring framework on simulated branch-level ability-EI data. The full code runs in under a minute on a current laptop. The R pipeline accompanying this paper executes this example via `05_worked_example.R`.

```
# 1. Simulate a sample with realistic ability-EI parameters
sample_data <- simulate_branches(N = 600, rho_13 = 0.50,
                                 r_det = 0.82, r_att = 0.78)

# 2. Apply BRI: regress attribution on detection, extract residual
bri <- compute_bri(sample_data$X_det, sample_data$Y_att)
sample_data$BRI <- bri$residual

# 3. Inspect the regression slope and R^2
cat("Slope:", round(bri$slope, 3),
    " R-squared:", round(bri$r2, 3),
    " Residual SD:", round(sd(bri$residual), 3))
# Slope: 0.411   R-squared: 0.175   Residual SD: 0.978
```

```r
# 4. Identify the profile of theoretical interest (high detection, low BRI)
sample_data$cell4 <- identify_cell4(sample_data$X_det, sample_data$BRI)
table(sample_data$cell4)
# FALSE  TRUE
#   533    67    (11.2% of sample)

# 5. Pre-validation: cross-task correlation
set.seed(2027)
e1_var <- (1 - 0.82) / 0.82
e3_var <- (1 - 0.78) / 0.78
X_taskB <- sample_data$T1 + rnorm(600, 0, sqrt(e1_var))
Y_taskB <- sample_data$T3 + rnorm(600, 0, sqrt(e3_var))
bri_taskB <- compute_bri(X_taskB, Y_taskB)
cross_task_correlation(sample_data$BRI, bri_taskB$residual)
# r = 0.707, passes >= .30 threshold (same participants, independent task noise)
```

The worked example produces a scatter plot in which detection accuracy is on the x-axis, attribution accuracy on the y-axis, the regression line is dashed, and Cell 4 members appear as below-the-line outliers at the high-detection end (Figure 2). Investigators applying BRI to their own data should reproduce this plot as an interpretive aid before any substantive regression analysis.

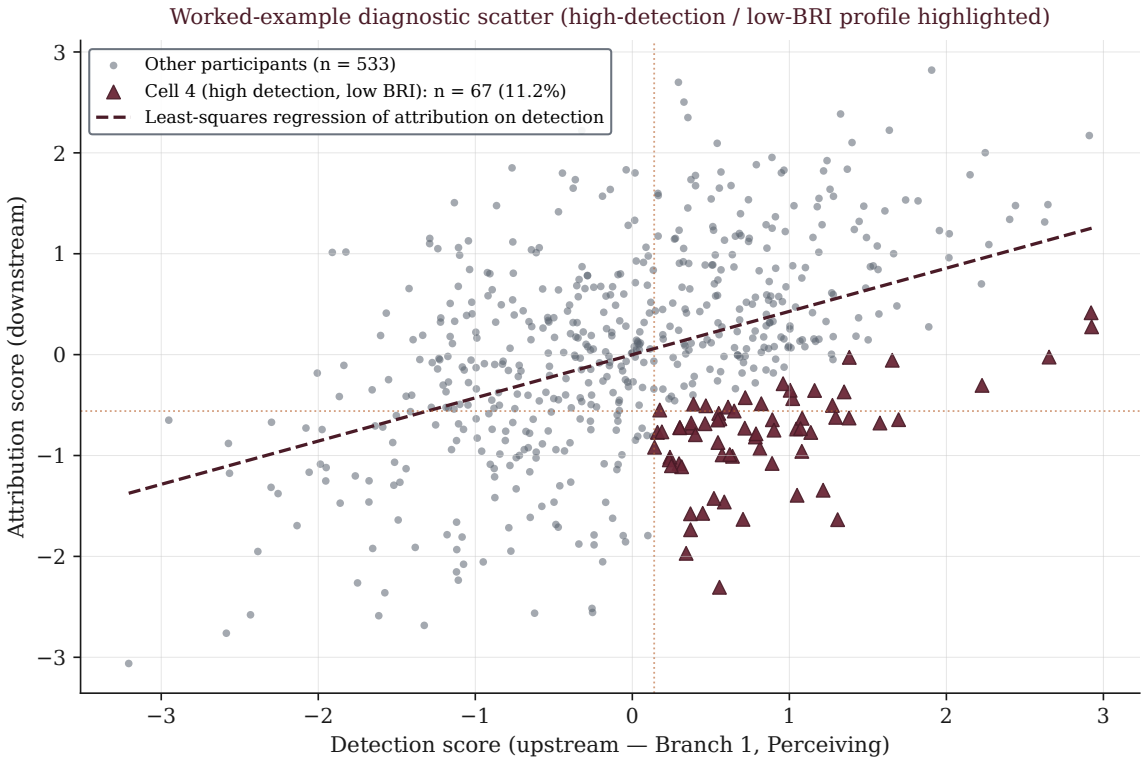

Figure 2. Diagnostic scatter plot from the worked example. Detection scores (upstream, Branch 1) on the x-axis; attribution scores (downstream) on the y-axis. The dashed line is the least-squares regression of attribution on detection; the BRI score for each participant is the signed vertical distance from this line. Red triangles flag the Cell 4 profile of theoretical interest (high detection, low BRI residual), representing approximately 11% of the sample, consistent with the value reported by the worked-example code in Section 7. Dotted reference lines mark the thresholds used to identify the cell. Investigators applying BRI to their own data should reproduce this plot before substantive analysis.

The worked example uses simulated data because, as Section 9.6 develops, recovery of a latent component cannot be benchmarked in real data where the latent components are precisely what is unobserved. A real-data demonstration on existing branch-level MSCEIT data is the subject of a companion empirical paper, pending data-access agreements (see Section 9.7, Limitations).

8. Software

The accompanying R pipeline (`bri-pipeline`) implements every study in this paper, plus the estimator for applied use. Its structure is a research compendium: each study is a numbered R script that loads shared functions from a `functions/` folder, runs its simulations, and writes results to a `results/` directory under the project working directory. Every table in this paper is reproducible from the pipeline. Tables 1 through 6 and Table 8 are produced by the four numbered driver scripts in turn; Table 7 is editorial guidance distilled from Studies 3 through 5, and is emitted as a static CSV by `03_robustness.R` (`results/study7_guidance_table.csv`) so the output set is complete.

Core functions:

- `simulate_branches()`: generate synthetic samples with specified true coupling and component reliabilities. Optional arguments allow ceiling-effect compression, heavy-tailed errors, and nonlinear coupling for the robustness studies.
- `compute_bri()`: extract the residual from the upstream-downstream regression. Optionally applies the disattenuation correction of Section 4 or robust regression for heavy-tailed conditions.
- `pre_validate()`: convenience wrapper that runs whichever Step 5 diagnostics the supplied inputs allow (cross-task correlation when a second task's residuals are provided, factor structure when a residuals matrix is provided, bootstrap residual stability when the upstream and downstream scores are provided), additionally invokes the nonlinearity test from Section 5.3 when scores are supplied, and clearly flags any diagnostic that was skipped for missing inputs.
- `test_nonlinearity()`: test for quadratic coupling before relying on the linear residual.
- `power_bri()`: regression-power analysis as in Section 6.

The pipeline is portable across macOS, Windows, and Linux; runtime for the full set of studies depends on the host: under a minute on an Apple Silicon MacBook Pro running R 4.6, and up to approximately eighteen minutes on a current MacBook Air. Per-script session captures are written to `Code/results/session_info_*.txt` after each run. A user manual targeted at first-time R users on macOS is included in the project distribution.

The implementation is distributed as a research compendium (Boullineau, 2026b). A future package release may provide an additional software citation. Code, simulation scripts, manual, and reproducibility materials are released under the MIT licence and archived on Zenodo under the concept DOI https://doi.org/10.5281/zenodo.20007669, which resolves to the latest version of the deposit. Running the four numbered driver scripts (`01_simulate.R`, `02_reliability.R`, `03_robustness.R`, `04_power.R`) with the included seeds regenerates Tables 1 through 6 and Table 8, subject only to minor platform-level or package-version numerical differences. Re-running the simulations with different seeds will produce Monte Carlo sampling variability whose magnitude is reported by the cell-level MCSEs written to the supplementary CSVs. `03_robustness.R` additionally writes Table 7 as a static guidance CSV.

9. Discussion

9.1 What the procedure contributes

The contribution is not the regression residual itself. That operation is the first step of the Frisch–Waugh–Lovell theorem (Appendix A) and has appeared in psychological measurement under various names for decades. The contribution lies in the integration. Five components are specified together: the residual as a *person-level scoring procedure* for ability tests with layered detection-then-reasoning architectures, a simulation-based boundary characterisation that says when this scoring procedure behaves well and when it does not, a diagnostic workflow for applied users to verify that the conditions hold in their data, a regression-power table for downstream effect detection, and an open reproducible implementation. None of these components alone is novel. Their integration into a single workflow with documented operating boundaries, calibrated to the conditions actually encountered in psychological ability data, is what the present paper adds. Appendix A locates that workflow in the cross-disciplinary methodological tradition running from classical econometrics through empirical finance to health econometrics.

9.2 When the procedure works and when it does not

The robustness studies of Section 5 give the practical answer. BRI works when:

- component reliabilities are at least .70, ideally .80
- the relationship between upstream and downstream components is approximately linear
- ceiling effects affect at most approximately the top 30% of the score distribution
- the dataset does not contain extreme leverage points or asymmetric outliers (these are conditions Study 4 does not directly model)

When any of these conditions fail substantially, BRI requires supplementation. Nonlinearity calls for polynomial or kernel residualisation. Severe ceiling effects call for score transformation. Very low component reliabilities call for a structural-equation-model alternative that estimates latent components directly (Section 2.3). Asymmetric outliers or visible leverage points call for robust regression. The R pipeline supports the formal nonlinearity diagnostic. It also provides cross-task, factor-structure, and bootstrap-stability validation checks via `pre_validate()`, and an optional robust-regression mode in `compute_bri()` for users who wish to verify their results on data containing extreme observations. Ceiling effects and severe distributional skew are screened by the investigator using descriptive statistics and visual inspection of the score distributions prior to residualisation. Score transformation is applied where indicated.

9.3 Distinguishing the estimator from the instrument

A point worth emphasising for methodological clarity. The instrument is the task battery, meaning the items, stimuli, and response procedures that produce the upstream and downstream scores. BRI is the *estimator* applied to the scores produced by that instrument. A poor instrument cannot be rescued by clever residualisation. A strong instrument with composite scoring can be substantially improved by it. Investigators contemplating applications of BRI should first invest in the quality of their upstream and downstream tasks. The residualisation is the second-order improvement, not the first.

9.4 The ability-EI worked example: illustrative, not evidential

The worked example in this paper uses ability emotional intelligence for three reasons. It is a clean case of a layered cognitive ability with a well-developed measurement tradition. Branch-level scoring already exists in established instruments (MSCEIT and successors). The detect-then-attribute architecture is theoretically motivated, and reasonable simulations of the data can be specified from published reliability estimates. What the residualised score *means* in psychological terms is the subject of a separate theoretical paper (Boullineau, 2026a) and a companion empirical paper currently in preparation. Neither of those is required for the present methodological argument. That argument stands or falls on whether the simulation evidence in Sections 3 through 6 supports the procedure's use as a person-level downstream-component score. Readers without an interest in ability EI can substitute working memory (encoding then maintenance), theory of mind (mental-state inference then prediction), executive function (general fluid ability as upstream input), or any analogous layered cognitive ability. The methodological argument is identical.

9.5 Beyond the present application

Three other domains illustrate the same logic. In *empathy research*, cognitive empathy and affective empathy correlate substantially, and isolating the unique variance of one given the other is a recurring challenge (Reniers, Corcoran, Drake, Shryane, & Völlm, 2011). In *executive-function research*, multiple component scores load on a general factor (Miyake et al., 2000). Residualising specific components on the general factor is one analytic option among several. In *self-regulation research*, performance-based and self-report measures are weakly correlated (Duckworth & Kern, 2011), and isolating the unique variance of each given the other is similarly fraught. The framework developed here, and the open-source pipeline that implements it, can be adapted to those cases. The conditions are that the upstream and downstream scores are theoretically separable, sufficiently reliable, and approximately linearly related (see Sections 3 through 5). Appendix A discusses the parallel applied traditions in empirical finance and health econometrics, where the same residualisation is well established under different names.

9.6 Why a simulation-first validation is appropriate here

The validation programme of Sections 3 through 6 is entirely simulation-based. This choice is principled rather than convenient. A scoring procedure that claims to recover a latent component-specific variance can only be validated against the latent structure that produced the data. That structure is observable only in simulated data, where the data-generating mechanism is known by construction. In real data, the latent components are precisely what is unobserved. Without a simulation step there is no benchmark against which recovery can be measured. The problem is general to residualised scoring procedures, and it is the standard motivation for simulation-first work in psychometrics and statistical-methods evaluation more broadly (e.g., Lance, Cornwell, & Mulaik, 1988; Cole & Maxwell, 2003; Morris, White, & Crowther, 2019). Luijken, Lohmann, Alter, and colleagues (2024) demonstrate the same logic in a recent simulation-only methods paper at this journal. The simulations of this paper establish that the procedure recovers the construct it claims to recover, under specified assumptions, and characterise its behaviour as those assumptions are relaxed. That is the load-bearing methodological contribution. Real-data demonstration is a complementary step, not a substitute for it (see Section 9.7).

9.7 Limitations

Six limitations warrant explicit acknowledgement.

First, the simulations characterise BRI under specified data-generating assumptions; departures we have not modelled (item-level reliability heterogeneity, missing-data patterns, multilevel sampling structures) are not addressed and remain open methodological questions.

Second, no real-data demonstration is reported in this manuscript, by design. The methodological argument is load-bearing on the simulation evidence (Section 9.6 explains why). The empirical application, which applies the estimator to existing branch-level MSCEIT data and to a new self-referential attribution-task pilot, is the subject of a separate companion paper currently in preparation pending data-access agreements. The substantive findings of that empirical work do not affect the methodological conclusions reported here: the simulations establish what the estimator does and does not recover under known data-generating mechanisms; the empirical paper will report what is found when the estimator is applied to a specific real dataset.

Third, the regression-power analysis in Section 6 assumes that the substantive analyst's outcome model is well-specified; misspecification of the outcome model will produce power estimates that are themselves miscalibrated.

Fourth, the prior-art comparison in Section 2.3 is necessarily summary rather than exhaustive. BRI's relationship to less-common decomposition methods (e.g., partial least squares, regularised residualisation, machine-learned residual estimators) is not characterised here and is left to subsequent methodological work.

Fifth, the calibration-sample logic emphasised in Section 2.4 (estimate slope and intercept in one sample; apply to a target sample) is recommended on principled grounds but is not itself stress-tested through a dedicated calibration-to-target transfer simulation in the present paper. A formal cross-sample portability study, generating calibration and target samples from the same data-generating process, applying the calibration regression to the target, and characterising recovery and rank stability under mild distribution shift, would convert the Section 2.4 recommendation from cautioned to evidenced. This is a clear next step.

Sixth, the present simulations assume independent measurement error on X and Y. Ability tests in practice often share method variance (common response format, common rater, shared verbal demand, shared motivation or careless responding), inducing correlated measurement error between the upstream and downstream scores. Under such correlated error, BRI would be expected to absorb a portion of the shared method variance into the residual, treating it as downstream-specific. The magnitude of this contamination is unmodelled here and is a high-priority addition for subsequent work; a correlated-error sensitivity study with $\text{cor}(\epsilon_X, \epsilon_Y) \in \{0, .10, .20, .30, .50\}$ is the natural next simulation.

9.8 The broader point

Composite scoring confounds component processes. Component-process residualisation is one principled route to disentangling them. The procedure developed here is one of many possible. The broader recommendation is simple. Ability tests of psychologically composite constructs benefit from explicit residualisation, validated by simulation, before substantive inferences are drawn from composite scores.

One practical caveat. Because BRI is a sample-relative residual (Section 2.4) rather than an absolute trait score, applied users should report the calibration sample alongside any BRI-based inference, and treat cross-sample portability as a question to test rather than assume. The accompanying R pipeline includes bootstrap rank-stability diagnostics for this purpose. Two methodological extensions are high priorities for follow-up work. A correlated-error sensitivity simulation, because shared method variance is the most plausible

psychometric threat the present simulations do not yet model. And a calibration-to-target transfer study, to put the sample-relative scoring logic on simulation footing rather than principled recommendation alone.

Declarations

Funding. The author received no external financial support for the research, authorship, or publication of this article.

Independent scholarship and institutional disclosure. This manuscript reports **independent personal research** by the author. It was not commissioned, supported, or directed by any employer, institution, or external entity, and the work falls outside the scope of any role held by the author. The author is currently completing an MSc in Psychology by distance learning at the University of Northumbria. This manuscript is not submitted in fulfilment of any degree requirement, is not supervised under the University's research programme, and is independent of any institutional research stream. Any affiliation listed is for correspondence and identification only and does not imply institutional sponsorship of, or responsibility for, the work.

Conflict of interest. The author declares no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Ethics approval. Not applicable. This is a methods article. No human or animal participants were involved in any of the studies reported. All datasets analysed are simulated.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Use of generative AI tools. The author used generative AI assistants (Claude from Anthropic, Codex from OpenAI) for copy-editing, structural review of the prose, code troubleshooting, implementation checking, and reference-format checking. AI tools were also used as search and brainstorming aids for locating relevant statistical terminology and implementation options. They were not used to generate hypotheses, determine the substantive argument, make final simulation-design choices, analyse or interpret data, or draw conclusions. The author reviewed and verified every output, retains records of prompts and outputs, and remains fully responsible for the content of the manuscript. This disclosure follows COPE, ICMJE, and Royal Society Open Science guidance on AI-assisted authoring.

Author contributions. Sole-author manuscript. The author conceived, designed, simulated, drafted, and revised the work. Full responsibility for the content rests with the author.

Code availability. All code is open-source under the MIT licence and accompanies the manuscript as supplementary material; see Open Practices Statement below.

Open Practices Statement

This is a methods article. No participant data were collected. All studies in this paper use Monte Carlo simulated data with known data-generating structure; the open-source pipeline accompanying the manuscript regenerates every table in the paper (Tables 1 through 6 and Table 8 from the four numbered simulation drivers; Table 7 emitted as a static guidance CSV by `03_robustness.R`). The pipeline includes the estimator (`compute_bri()`), simulation generators (`simulate_branches()`), pre-validation diagnostics, regression-power routines, and a worked example, together with a user manual targeted at first-time R users on macOS. Code, manual, and supporting materials are released under the MIT licence and archived on Zenodo (concept DOI https://doi.org/10.5281/zenodo.20007669, resolving to the latest version). This study was not formally pre-registered; pre-registration is appropriate for empirical applications of the procedure rather than for the methodological evaluation reported here.

Data Availability Statement

No new participant data were collected. All datasets analysed in this article were generated by Monte Carlo simulation using the R scripts provided in the open-source pipeline (see Section 8 and the Open Practices Statement above). Running the four numbered driver scripts in the pipeline (`01_simulate.R`, `02_reliability.R`, `03_robustness.R`, `04_power.R`) with the included seeds regenerates every reported number in Tables 1 through 6 and Table 8, subject only to minor platform-level or package-version numerical differences. Re-running with different seeds will produce Monte Carlo sampling variability; cell-level MCSEs are written to the supplementary CSVs alongside the point estimates. `03_robustness.R` additionally emits Table 7 as a static guidance CSV (`results/study7_guidance_table.csv`). Each script sets its own random-number seed at the top.

Appendix A: Connections to Factor-Residual Econometrics

The estimator developed in Section 2 has direct antecedents in econometric residualisation, most notably the Frisch–Waugh–Lovell (FWL) theorem (Frisch & Waugh, 1933; Lovell, 1963). This appendix makes the connection explicit, both to acknowledge prior art at the formal level and to clarify exactly what BRI adds beyond the underlying mathematical identity.

A.1 The Frisch–Waugh–Lovell theorem

Consider a multiple linear regression of an outcome variable Z on two predictors, an upstream score X and a downstream score Y:

$$Z_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Y_i + u_i. \quad (A.1)$$

The FWL theorem states that $\hat{\gamma}_2$ can equivalently be obtained by:

1. Regressing Y on X to obtain the residual $\widetilde{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.
2. Regressing Z on X to obtain the residual $\widetilde{Z}_i = Z_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_i$.
3. Regressing $\widetilde{Z}$ on $\widetilde{Y}$; the slope coefficient equals $\hat{\gamma}_2$ exactly.

In short: the coefficient on Y in a multiple regression that also includes X is the same as the simple coefficient on the residualised Y in a regression of the residualised Z on it. Two-step residualisation gives the same coefficient as one-step multiple regression.

A.2 BRI as the first step of FWL deployed for scoring

BRI is precisely the residual $\widetilde{Y}_i$ from Step 1 of the FWL procedure, but redeployed as a *person-level score* rather than as an algebraic device for coefficient estimation. The mathematical operation is identical to FWL's first step. The contribution lies in what is done with that residual afterwards.

Three uses distinguish BRI from FWL-as-coefficient-tool:

(i) **Person-level interpretation.** The residual $\widetilde{Y}_i$ ranks individuals by how much higher or lower their downstream score is than their upstream score would predict. This ranking is the substantive object of interest, not a means to estimating a coefficient. The FWL theorem treats residuals as a step in coefficient identification; BRI treats them as a measurement.

(ii) **Use across multiple outcome models.** A BRI residual computed once can serve as a predictor in many subsequent regressions (outcome A, outcome B, outcome C) without re-fitting the residualisation. In a pure FWL setting, each new outcome would conceptually require a new full multiple regression. The pre-computed residual is the practical efficiency.

(iii) **Profile identification and visualisation.** The residual supports diagnostic plotting (e.g., the high-detection, low-residual cell in the worked example of Section 7), the identification of subgroups with theoretically interesting profiles, and visual inspection of the upstream–downstream relationship. These uses operate at the population descriptive level, not at the level of coefficient hypothesis testing.

A.3 Conditions under which the FWL residual functions as a defensible score

FWL holds as an algebraic identity for any OLS specification. It does not require the residual to be useful as a score. The substantive question for BRI is when the residual has properties (reliability, recovery of a meaningful latent component, bias-tolerance under departures from idealised assumptions) that make it defensible. Sections 3 through 6 of the body characterise those conditions through simulation. Briefly:

- Variance recovery is governed primarily by downstream reliability (Section 4), with recovery coefficients of .77 to .94 at $r_{XX} \geq 0.80$ (Section 3).
- The estimator is robust to ceiling effects of up to approximately 30% of the score distribution and is essentially flat across the symmetric heavy-tailed measurement-error grid examined (from t_3 through Gaussian, with reliabilities held fixed). It requires extension under detected nonlinear coupling and under asymmetric outliers, leverage points, or contamination processes that alter reliability (Section 5).
- A Lord-style disattenuation correction (Section 4) is available for restricted-range applications.

These properties are what BRI documents and what FWL alone does not address. The methodological contribution is the integration of the FWL residual with the simulation-based characterisation of when it functions as a meaningful person-level score.

A.4 Cross-disciplinary correspondence

The same residualisation appears under different names across applied quantitative fields.

In *empirical finance*, factor-model residualisation appears in two distinct but related guises. Time-series regression of an asset's return on market and style factors decomposes return into systematic exposure (the factor loadings) and an unexplained component, namely the intercept (Jensen's alpha; Jensen, 1968) and the residual (idiosyncratic return). The Fama and MacBeth (1973) two-pass procedure separately uses estimated factor loadings from first-pass time-series regressions as the *predictors* in a second-pass cross-sectional regression of average returns, with the focus on estimating factor risk premia. Modern multi-factor extensions (Fama & French, 1993; Carhart, 1997) follow the same logic. The structural parallel with BRI is the time-series first-pass step: in both cases an observed quantity (an asset's return, or an individual's downstream score) is regressed on an upstream variable (market factors, or an upstream task score) and the residual is interpreted as the component unexplained by the upstream. The terminology differs across these traditions, with "idiosyncratic return", "alpha", and "residual" referring to distinct quantities in finance, but the underlying residualisation operation is the same. The applied finance literature has accumulated substantial machinery for the robust estimation, multi-factor extension, and inference for residual-based quantities, much of which is conceptually parallel to the issues this paper addresses through simulation.

In *health econometrics*, two-stage residual inclusion (Terza, Basu, & Rathouz, 2008) uses the same residual to correct for endogeneity in models of treatment effects, though the inferential goal (causal effect identification via an instrument) differs from BRI's (descriptive scoring without a causal claim). The mechanical similarity nonetheless underwrites the family resemblance.

In *operational risk modelling*, layered detection-then-adjudication architectures share the same upstream-downstream architecture BRI is built for. Examples include transaction screening followed by analyst review, fraud detection followed by case-management decision, and anti-money-laundering alert generation followed by enhanced due diligence. Industry practice in these settings already separates "alert quality" (an upstream-detection variable) from "case-disposition quality" (a downstream-reasoning variable), though typically without the formal residualisation framework that BRI provides.

These applications are not the present paper's contribution. Each has its own substantial literature and its own established estimators. But they situate the procedure within a recognisable cross-disciplinary methodological tradition. The contribution of the present paper, again, is the systematic simulation-based characterisation of when the FWL residual functions as a defensible person-level score in psychological ability data, with the diagnostic and reproducibility infrastructure required to determine whether the necessary conditions hold in any given dataset.

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 (concept DOI; resolves to latest version)

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

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

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.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 558–577. https://doi.org/10.1037/0021-843X.112.4.558

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

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, 45(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3), 607–636. https://doi.org/10.1086/260061

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.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2), 389–416. https://doi.org/10.1111/j.1540-6261.1968.tb00815.x

Lance, C. E., Cornwell, J. M., & Mulaik, S. A. (1988). Limited information parameter estimates for latent or mixed manifest and latent variable models. *Multivariate Behavioral Research*, 23(2), 171–187. https://doi.org/10.1207/s15327906mbr2302_3

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

Luijken, K., Lohmann, A., Alter, U., Claramunt Gonzalez, J., Clouth, F. J., Fossum, J. L., Hesen, L., Huizing, A. H. J., Ketelaar, J., Montoya, A. K., Nab, L., Nijman, R. C. C., Penning de Vries, B. B. L., Tibbe, T. D., Wang, Y. A., & Groenwold, R. H. H. (2024). Replicability of simulation studies for the investigation of statistical methods: The RepliSims project. *Royal Society Open Science*, 11(1), 231003. https://doi.org/10.1098/rsos.231003

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, 8(4), 540–551. https://doi.org/10.1037/a0012746

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.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086

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

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, 93(1), 84–95. https://doi.org/10.1080/00223891.2010.528484

Roberts, R. D., MacCann, C., Matthews, G., & Zeidner, M. (2010). Emotional intelligence: Toward a consensus of models and measures. *Social and Personality Psychology Compass*, 4(10), 821–840. https://doi.org/10.1111/j.1751-9004.2010.00277.x

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

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

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279