

When Validation Defines Abnormality: Prediction Targets in Bayesian Accrual Measurement

Quan Pham

phamtruongquan@iuh.edu.vn

Industrial University of Ho Chi Minh City

Research Article

Keywords: residual-based measurement, abnormal accruals, validation target, predictive stacking, hierarchical models, construct operationalization

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10834137/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

When Validation Defines Abnormality: Prediction Targets in Bayesian Accrual Measurement

Abstract

Residual-based proxies are common in social science: researchers first estimate normal behavior and then interpret deviations from that benchmark as abnormality. When observations are clustered, however, predictive validation can evaluate different reference frames. This paper studies whether that choice becomes part of the measurement instrument itself. Using Bayesian stacking across established normal-accrual benchmark families, I hold the model space, estimation protocol, scoring rule, and information set fixed while changing only the held-out unit: individual firm-years versus entire firms. Row-level validation strongly rewards firm-random-effect benchmarks, whereas grouped-firm validation shifts weight toward population-level benchmarks. Simulation and restricted-model checks locate this reallocation primarily in firm-level heterogeneity rather than instability among pooled accounting families. The benchmark shift propagates to the generated residual: overall abnormal-accrual ranks remain related across targets, but top-tail classifications change materially, especially after uncertainty scaling. The evidence shows that a validation target can define the reference frame of a residual-based proxy, not merely its predictive score. The result is specific to generated measures built from predictively selected benchmarks and generalizes to clustered longitudinal settings in which predicted normal behavior is used to operationalize latent abnormality.

Keywords: residual-based measurement, abnormal accruals, validation target, predictive stacking, hierarchical models, construct operationalization

1 Introduction

Many empirical constructs in social science are observed only through a measurement mapping rather than directly. One common architecture estimates a reference level of “normal” behavior and interprets deviation from that reference as an observable trace of an underlying latent phenomenon ([Adcock and Collier 2001](#)). Abnormal accruals are a prominent accounting example: total accruals are observed, but normal accruals are model-generated and the residual is subsequently used as evidence about unusual reporting behavior ([Dechow et al. 2010](#); [Nezlobin et al. 2022](#)). The empirical measure therefore inherits the assumptions that define its counterfactual benchmark. This dependence matters whenever the benchmark is not fixed ex ante but selected

or weighted using predictive performance, because the rule used to reward candidate benchmarks can become part of the measurement architecture itself.

Accrual research has established two relevant features of this architecture. First, alternative Jones-type, performance-adjusted, cash-flow-mapping, and asymmetric-recognition models encode different notions of normal accrual behavior, while peer selection, firm shocks, and first-stage estimation can further contaminate the generated residual (Jones 1991; Dechow et al. 1995; Kothari et al. 2005; Dechow and Dichev 2002; Ball and Shivakumar 2006; Owens et al. 2017; Chen et al. 2018). Second, Bayesian approaches make parameter, model, and predictive uncertainty explicit and can combine candidate normal-accrual benchmarks according to out-of-sample predictive performance rather than committing to one specification *ex ante* (Breuer and Schütt 2023; Fiechter and Wallmeier 2026; Yao et al. 2018). More generally, work on generated regressors and generated outcomes shows that quantities estimated in one stage can transmit first-stage choices into later inference (Pagan 1984; Murphy and Topel 1985; Chen et al. 2023; Byzalov and Basu 2024). Together, these literatures establish that benchmark construction matters, but they leave open a prior question: what determines the benchmark when predictive validation itself is used to allocate weight across competing definitions of normality?

That question becomes nontrivial in clustered longitudinal data because predictive performance has more than one coherent target. Prediction for another observation from a cluster already represented in training differs from prediction for an unseen cluster, and conditional versus marginal predictive assessment formalizes the same distinction in hierarchical models (Vaida and Blanchard 2005; Merkle et al. 2019). Cross-validation under dependence therefore has an estimand: the held-out unit determines the reference frame represented by the train–test split (Rabinowicz and Rosset 2022; Bates et al. 2024; Roberts et al. 2017). In an accrual panel, row-level validation evaluates candidate benchmarks with same-firm history available, whereas grouped-firm validation evaluates them with the focal firm’s history unavailable. Existing validation research establishes that these are different predictive tasks; existing accrual research establishes that benchmark choice affects the generated measure. The unresolved bridge is whether changing the validation estimand can reallocate weight across accrual benchmarks and thereby change the residual before any downstream analysis begins (Breuer and Schütt 2023; Fiechter and Wallmeier 2026).

The resulting methodological puzzle is not which cross-validation rule is generically superior. It is whether alternative reference frames leave distinct, theoretically predicted traces in the measurement process. If persistent firm heterogeneity drives the difference, a row-level target should reward firm-specific benchmark structure more strongly than a grouped-firm target; if the contrast instead reflects unstable choice among accounting specifications, the reallocation should remain after firm-specific structure is removed. A second implication concerns propagation. Even when two target-specific residuals preserve a similar global ordering of firm-years, they may disagree near an extreme screening threshold, so rank association and tail overlap need not tell the same measurement story. Feature timing is a separate design dimension: withholding a firm and withholding future information answer different questions and

should not be conflated. These implications yield a testable chain from validation estimand to benchmark allocation, generated residual, and extreme-case classification. To examine this measurement problem, the study uses audited annual financial statements for 568 firms listed on HOSE and HNX over 2015–2024, with active measurement beginning in 2017. The repeated firm histories are useful because they permit two economically interpretable reference frames to be compared on the same observations and with the same candidate benchmark catalogue. The analysis first contrasts row-level and grouped-firm validation in their allocation of predictive weight across pooled and firm-specific normal-accrual structures, and then traces any benchmark difference into residual rankings and extreme-case membership. A no-lookahead measurement space is primary, with a retrospective space used to separate the held-out-unit question from feature timing. Simulations and restricted benchmark sets provide mechanism checks for whether any divergence is attributable to persistent firm heterogeneity rather than generic instability among accounting specifications.

The evidence resolves the two implications in different parts of the measurement chain. First, row-level validation assigns roughly 0.59–0.62 of aggregate weight to firm-random-effect benchmarks, compared with about 0.04 under grouped-firm validation; the mechanism checks attribute most of that reallocation to persistent firm heterogeneity rather than to unstable switching among pooled accounting families. Second, the target-specific residuals remain strongly related in their global ranking of firm-years (Spearman correlations of about 0.83), but they agree much less on which observations occupy the top five percent, where Jaccard overlap falls to about 0.65–0.69 and declines further after uncertainty scaling. These regularities support three bounded contributions. The first is conceptual: when predictive performance selects or weights the benchmark that generates a residual proxy, the validation estimand becomes part of the construct’s measurement mapping. The second is diagnostic: high global rank agreement does not imply stable classification at the tails where residual measures are often used for screening or sample construction. The third is methodological: mechanism checks can separate reference-frame dependence from generic model-family instability, implying that the held-out unit should be aligned with the intended empirical meaning of normality. The evidence therefore does not identify a universally superior validation rule; it shows that coherent reference frames can operationalize different versions of the same residual-based construct.

The remainder of the paper proceeds as follows. Section 2 develops the measurement and validation framework; Section 3 describes the data, Bayesian benchmark construction, and validation design; Section 4 reports the benchmark-allocation and classification evidence; and Section 5 discusses methodological implications, boundary conditions, and limitations.

2 Residual-Based Measurement and Validation

2.1 Generated accrual measures

Residual-based measures begin from a distinction that is fundamental to empirical measurement: a latent concept is not identical to the observable quantity used to

operationalize it ([Adcock and Collier 2001](#)). In accrual research, the latent object is the accrual that would have arisen under a relevant notion of normal operations. Because that counterfactual is unobserved, researchers estimate a benchmark from accounting fundamentals and define abnormal accrual as the deviation from that benchmark. The residual can therefore carry information about unusual reporting behavior without directly revealing managerial intent. This construct–proxy distinction is particularly consequential in accrual research because alternative accrual-quality measures embed different economic assumptions and can differ in construct validity ([Dechow et al. 2010](#); [Nezlobin et al. 2022](#)).

The established accrual models illustrate how the measurement mapping is created. Jones and modified-Jones specifications model expected accruals using revenue changes and investment capacity; performance-adjusted models accommodate the relation between accruals and profitability; Dechow–Dichev and McNichols-type models use the mapping between accruals and cash flows; and Ball–Shivakumar-style models allow asymmetric accrual responses to adverse economic states ([Jones 1991](#); [Dechow et al. 1995](#); [Kothari et al. 2005](#); [Dechow and Dichev 2002](#); [McNichols 2002](#); [Ball and Shivakumar 2006](#)). These are not interchangeable descriptions of the same counterfactual. Each assigns a different portion of observed accounting variation to normal accruals and, by construction, leaves a different remainder in the residual.

A second literature shows why the remainder can contain more than the latent reporting phenomenon of interest. Balance-sheet accrual measures can be contaminated by non-articulation events ([Hribar and Collins 2002](#)); growth, performance, operating-cycle characteristics, and cash-flow dynamics can enter expected accruals imperfectly ([McNichols 2000](#); [Collins et al. 2017](#); [Frankel and Sun 2018](#); [Dichev and Owens 2025](#)); the estimation sample and peer definition can alter the comparison set ([Ecker et al. 2013](#); [Colas and Brousseau 2025](#)); and idiosyncratic shocks to underlying firm economics can appear as abnormal accruals when the benchmark is incomplete ([Owens et al. 2017](#)). Construct-validity work therefore treats the empirical measure as a mapping that must be evaluated, not as the construct itself ([Nezlobin et al. 2022](#)). The implication for the present study is not that residual measures are unusable, but that conclusions about abnormality are conditional on how normality is defined.

Once a model-generated quantity is reused, a separate methodological problem arises. Econometrics has long recognized that regressors or dependent variables generated in a first step inherit estimation uncertainty that can affect subsequent inference ([Pagan 1984](#); [Murphy and Topel 1985](#)). In accounting, this issue has been developed directly for regression residuals and other generated quantities: residual-based second-stage analyses can yield incorrect coefficient or error-rate inferences ([Chen et al. 2018](#); [Christodoulou et al. 2018](#)); generated regressors can understate standard errors when first-step uncertainty is ignored ([Chen et al. 2023](#)); and regression-based scores can create coefficient and interpretation problems when treated as dependent variables in a later stage ([Byzalov and Basu 2024](#)). Related work on discretionary accruals shows that first-stage specification choices can also alter which observations are classified as extreme ([Jackson 2018](#); [McNichols and Stubben 2018](#)). This literature establishes that generated measures matter downstream. The present paper moves one stage upstream: it asks whether the predictive validation target used to choose or weight the first-stage

benchmark can itself change the generated measure before any second-stage reuse occurs.

Bayesian accrual measurement makes this upstream architecture observable. Let $m = 1, \dots, M$ index candidate normal-accrual benchmarks. A model-averaged benchmark under validation target V can be written schematically as

$$\widehat{NDA}_{i,t}^V = \sum_{m=1}^M w_m^V \hat{\mu}_{i,t,m}, \quad (1)$$

where w_m^V is the predictive weight assigned to candidate m under target V . The generated abnormal-accrual proxy is then $DA_{i,t}^V = TA_{i,t} - \widehat{NDA}_{i,t}^V$. Bayesian stacking is useful because it estimates a continuous predictive allocation across candidate models rather than forcing a winner-take-all choice (Yao et al. 2018). Recent Bayesian accrual work shows how parameter, model, and predictive uncertainty can be incorporated into normal-accrual measurement (Breuer and Schütt 2023; Fiechter and Wallmeier 2026). Equation (1) isolates a different source of dependence: if the validation target changes w_m^V , it changes the normal benchmark and therefore the residual, even when the data, candidate models, and fitting procedures are otherwise held fixed. This logic is not uniquely Bayesian; stacking simply makes the allocation mechanism transparent.

2.2 Validation estimands under dependence

The relevant validation literature begins with the inferential target, not with a generic rule about cross-validation. In mixed-effects settings, prediction concerning the particular clusters already observed differs from prediction concerning the population or a new cluster (Vaida and Blanchard 2005). Bayesian model comparison makes a closely related distinction between conditional predictive assessment, which conditions on cluster-specific latent structure, and marginal assessment, which integrates over such structure when predicting a new cluster (Merkle et al. 2019). More generally, standard cross-validation can be inappropriate when observations are correlated and the dependence structure connecting training and test observations does not match the target prediction problem (Rabinowicz and Rosset 2022). Modern work therefore treats the held-out unit as part of the predictive estimand: structured folds are designed around temporal, spatial, hierarchical, or cluster dependence rather than being merely alternative computational partitions (Roberts et al. 2017; Adin et al. 2024; Liu et al. 2025; Qiu et al. 2025). Recent theory further emphasizes that cross-validation estimates a particular prediction target rather than a model-free, universal notion of out-of-sample error (Bates et al. 2024).

For an accrual panel, the firm is the natural cluster. Row-level validation can hold out firm-years while retaining other years of the same firm in training, allowing a hierarchical benchmark to exploit persistent firm structure when scoring the held-out observation. Grouped-firm validation withholds all observations of the focal firm, so candidate performance is evaluated without access to that firm’s realized history.

The two targets correspond to different predictive reference frames. Row-level validation asks how well a benchmark predicts another observation from a represented firm; grouped-firm validation asks how well it predicts when the focal firm is unrepresented. When those target-specific scores are subsequently used to weight full-sample benchmark fits, the resulting residual inherits the weighting reference frame without requiring the final estimation sample itself to exclude the focal firm. Neither reference frame is universally preferable. Which one is appropriate depends on the construct meaning assigned to the generated residual (Vaida and Blanchard 2005; Merkle et al. 2019; Liu et al. 2025).

This distinction is therefore not well described as a simple “leakage” problem. Same-firm history is legitimate information if the target is deviation from an observed firm’s learned trajectory. It is unavailable by design if the target is deviation from a population benchmark for an unseen firm. Treating all within-firm information as contamination would collapse two distinct estimands into one procedural rule; treating random row-wise validation as universally valid would make the opposite error. Dependence-aware validation instead requires the training–test partition to represent the intended reference frame (Roberts et al. 2017; Rabinowicz and Rosset 2022; Merkle et al. 2019).

Temporal information is a separate design axis. A model may restrict features to information available by time t yet still use row-level folds that retain other observations from the same firm. Conversely, a retrospective accrual model may use adjacent-period cash flows while still be evaluated under grouped-firm folds. Leave-future-out methods formalize the first problem by aligning the training information set with time-directed prediction (Bürkner et al. 2020); leave-group-out methods formalize the second by aligning the held-out unit with cluster-level generalization (Liu et al. 2025). The present study keeps these dimensions conceptually distinct: “no look-ahead” constrains what information enters the feature space, whereas row-level versus grouped-firm validation determines whose history can inform the benchmark.

2.3 Measurement implications

The two literatures reviewed above imply a specific measurement dependency. Hierarchical prediction distinguishes prediction for an already observed cluster from prediction for a new cluster because the former can condition on realized cluster structure whereas the latter must rely on population-level information or integrate over new-cluster uncertainty (Vaida and Blanchard 2005; Merkle et al. 2019; Rabinowicz and Rosset 2022). Predictive model combination then translates the chosen target into weights on candidate models (Yao et al. 2018; Bates et al. 2024). In accrual measurement, those candidates define the counterfactual level of normal accruals from which abnormal accruals are generated (Nezlobin et al. 2022; Breuer and Schütt 2023; Fiechter and Wallmeier 2026). The methodological implication is a propagation chain: a validation estimand can affect benchmark allocation and thereby alter the residual used to operationalize abnormality.

The first implication concerns the location of that reallocation. When other observations from the same firm remain in training, a hierarchical benchmark can exploit

persistent firm structure; when the entire firm is withheld, the realized firm effect is unavailable (Vaida and Blanchard 2005; Merkle et al. 2019; Liu et al. 2025). If persistent heterogeneity is the operative channel, the row-versus-grouped contrast should therefore be concentrated in the value assigned to firm-specific structure rather than dispersed arbitrarily across accounting families. Hierarchical accrual models provide a natural setting for this distinction because firm-level heterogeneity can be economically and predictively important (Breuer and Schütt 2023). A useful design must consequently distinguish reference-frame dependence from generic instability in model choice, for example by asking whether the contrast weakens when firm-specific structure is absent or removed from the candidate space (Rabinowicz and Rosset 2022; Bates et al. 2024).

Temporal dependence is a competing source of fold sensitivity. Accrual reversals and persistent operating shocks can link adjacent firm-years even without a stable firm effect (Owens et al. 2017; Dichev and Owens 2025). Time-directed validation addresses the information set available across calendar time, whereas leave-group-out validation changes the identity of the held-out cluster (Bürkner et al. 2020; Liu et al. 2025). These mechanisms imply different observables: persistent heterogeneity predicts a sharp increase in the value of firm-specific structure when the firm’s history is available, whereas temporal dependence should operate through adjacent-period information even in the absence of stable firm effects. The empirical design therefore treats firm heterogeneity and temporal dependence as distinct explanations rather than collapsing them into a generic dependence problem.

The second implication concerns propagation from benchmark allocation to the generated accounting quantity. Prior research shows that model-generated residuals and scores can transmit first-stage specification choices into later inference and classification (Chen et al. 2018, 2023; Byzalov and Basu 2024), while accrual-specific studies show that alternative first-stage mappings can change which observations appear extreme (Jackson 2018; McNichols and Stubben 2018; Nezlobin et al. 2022). The relevant diagnostic is therefore two-dimensional. Global rank association captures whether the broad ordering of firm-years is preserved, whereas matched-tail overlap captures whether the observations classified as most abnormal remain the same. A pattern of high rank correlation but materially lower tail overlap would indicate a concentrated measurement effect rather than wholesale instability of the residual measure. Posterior-uncertainty scaling is examined separately because it introduces an additional generated denominator and therefore creates a distinct measurement object (Breuer and Schütt 2023; Fiechter and Wallmeier 2026).

Finally, disagreement between two generated measures is not by itself evidence that one more closely reveals managerial intent. Construct-validity logic instead asks whether alternative operationalizations leave different, theoretically relevant observable traces while preserving the distinction between proxy and latent construct (Adcock and Collier 2001; Nezlobin et al. 2022). Subsequent cash flow, earnings, and profitability are therefore used only to assess whether target-specific extreme sets contain economically different firm-years, not to identify earnings-management motives or to rank one validation target as uniquely correct. The analysis thus follows a single methodological sequence—validation estimand $\rightarrow$ benchmark allocation $\rightarrow$ generated

Table 1 Sample construction and measurement spaces

Sample stage	Firm-years	Firms	Years
Licensed workbook	5,680	568	2015–2024
Post-continuity base	4,544	568	2017–2024
No-lookahead measurement space	4,494	568	2017–2024
Ex-post measurement space	3,936	568	2017–2023

Notes: The two active spaces use the same 568-firm post-continuity base and differ in eligible firm-years and information timing. The ex-post space permits lead cash-flow variables; the no-lookahead space excludes future-year accounting covariates. Detailed coverage and eligibility diagnostics appear in the Supplementary Information.

residual $\rightarrow$ extreme-case classification—and evaluates whether each transition leaves the discriminating pattern implied by the corresponding mechanism.

3 Methods

3.1 Data and benchmark models

The empirical panel contains audited annual financial-statement data for firms listed on HOSE and HNX, obtained from a licensed FiinPro/FiinPro-X workbook spanning 2015–2024. The preprocessing pipeline excludes 2015 before constructing contiguous lags and leads, so 2016 is the initial source year and active measurement begins in 2017. The unit of analysis is the firm-year. Raw data cannot be redistributed under the licence. Table 1 reports the sample flow. The post-continuity base contains 568 firms. The primary no-lookahead space contains 4,494 firm-years over 2017–2024; a retrospective ex-post space contains 3,936 firm-years over 2017–2023.

The two spaces serve different measurement purposes. The no-lookahead space is primary because future-year covariates are excluded from candidate models. It therefore isolates the validation-unit question from feature lookahead. The ex-post space allows lead operating cash flow and supports Dechow–Dichev or McNichols-type retrospective cash-flow mappings. The spaces are not interpreted as a clean “real-time versus ex-post” experiment because their admissible benchmark catalogues differ. Instead, they provide two environments in which the same held-out-unit comparison can be examined.

Total accruals are defined from the cash-flow statement rather than balance-sheet changes:

$$TA_{i,t} = \frac{NI_{i,t} - CFO_{i,t}}{A_{i,t-1}}, \quad (2)$$

where NI is net income, CFO is operating cash flow, and $A_{i,t-1}$ is lagged total assets. This choice avoids a known non-articulation problem in balance-sheet accrual measures (Hribar and Collins 2002). Continuous predictors are scaled as specified by the benchmark family, winsorized at the 1st and 99th percentiles, and standardized before Bayesian fitting. The baseline uses globally estimated transformation constants

within each declared analysis space; fold-local winsorization and standardization are used as a sensitivity so that preprocessing is not assumed to be innocuous.

The candidate catalogue represents alternative accounting views of normal accruals rather than an unrestricted search over predictive algorithms. It includes Jones and modified-Jones variants, performance adjustment, Dechow–Dichev and McNichols-type cash-flow mappings, Ball–Shivakumar-style asymmetric recognition, a no-lookahead cash-flow/Jones adaptation, and secondary volatility and operating-cycle extensions (Jones 1991; Dechow et al. 1995; Kothari et al. 2005; Dechow and Dichev 2002; McNichols 2002; Ball and Shivakumar 2006). The purpose of this breadth is not to declare one family universally correct. It is to allow predictive validation to allocate weight across economically distinct definitions of normal accruals.

Each admissible family is estimated under pooled and firm-random-intercept structures where feasible. The firm-structured arm is

$$TA_{i,t} = X_{i,t}\beta + \delta_t + u_i + \epsilon_{i,t}, \quad u_i \sim N(0, \sigma_u^2), \quad (3)$$

where δ_t captures year structure and u_i captures persistent firm heterogeneity. Pooled candidates instead use population-level industry/year structure without a firm-specific random intercept. Candidate cells are therefore indexed jointly by accounting family and heterogeneity arm. This joint indexing matters for the benchmark-allocation analysis because a validation target can change either the accounting mapping, the value assigned to firm-level structure, or both. The full model registry, covariate eligibility rules, priors, and code identifiers are reported in Supplementary Table S1.

3.2 Bayesian estimation and validation

All candidates use the same Bayesian estimation protocol within a measurement space. The baseline likelihood is Student- t , chosen to accommodate heavy-tailed scaled accruals after winsorization. Continuous predictors are standardized so that weakly regularizing priors have a common scale interpretation. The replicated configuration uses $N(0, 0.10)$ priors for standardized slopes and the intercept, exponential scale priors for residual and firm-effect standard deviations, and a Gamma prior for the Student- t degrees of freedom. Prior adequacy is evaluated through prior predictive simulation rather than target-specific tuning. The same prior registry is applied to row-level and grouped-firm validation.

Models are fitted in `brms`/Stan using four chains, 12,000 iterations per chain, 4,000 warmup iterations, `adapt_delta`=0.99, and `max_treedepth`=15. Prior plausibility is assessed by prior predictive simulation. Primary fit reliability is screened using $\hat{R}$, bulk and tail effective sample size, divergent transitions, and treedepth warnings; candidates that do not satisfy the declared reliability gates are excluded from the primary exact-K-fold stacking comparison. Posterior predictive checks are retained as secondary model-adequacy diagnostics rather than as an additional exact-K-fold inclusion gate. These computational choices are held constant across validation targets so that the treatment of interest is the held-out unit rather than a change in estimation quality.

Table 2 Validation targets as measurement reference frames

Validation target	Held-out unit	Same firm in training?	Measurement reference frame
Row-level exact K-fold	Firm-year	Yes (primary set)	Represented-firm benchmark weighting
Grouped-firm exact K-fold	Entire firm	No	Out-of-firm benchmark weighting

The primary comparison uses exact row-level K-fold and exact grouped-firm K-fold with $K = 5$, fixed fold assignments, and exact refits. Table 2 states the implied reference frames. Row-level folds are constructed within firm, and the primary score is restricted to held-out firm-years for which at least one observation from the same firm remains in the training set. Grouped-firm K-fold instead assigns all observations from a firm to the same fold. The former therefore estimates predictive performance for another observation from a represented firm; the latter estimates performance when the focal firm’s own history is unavailable.

For a firm-random-intercept candidate, the baseline grouped score is population-level. The fitted random intercept of the held-out firm is unavailable, so its group-specific contribution is not conditioned on an estimated firm effect. The expected grouped benchmark is therefore

$$E(\mu_{new,t} \mid x_{new,t}, D) = E(x_{new,t}\beta + \delta_t \mid D). \quad (4)$$

A separate sensitivity evaluates the fully marginal predictive density for a genuinely new firm by integrating over a newly drawn $u_{new} \sim N(0, \sigma_u^2)$. This is a stricter out-of-firm predictive object. The sensitivity is kept separate from the baseline because it changes the predictive density while preserving the held-out-firm target.

Predictive stacking assigns nonnegative weights to candidate predictive distributions and chooses the mixture that maximizes held-out log predictive density (Yao et al. 2018). For validation target V ,

$$\hat{w}^V = \arg \max_w \sum_{j \in V} \log \left[\sum_{m=1}^M w_m p_m(TA_j \mid D_{-V(j)}) \right], \quad w_m \geq 0, \sum_m w_m = 1, \quad (5)$$

where $V(j)$ denotes the row fold or firm fold containing observation j . Benchmark allocation is evaluated primarily through two objects: the target-specific leading benchmark cell and the aggregate weight assigned to the firm-random-effect arm. The latter is especially informative for the focal mechanism because the two validation targets differ in whether same-firm history can contribute to held-out prediction.

A pooled-only sub-stacking sensitivity then removes all firm-random-effect candidates. If the main contrast is primarily a heterogeneity effect, row-level and grouped-firm family allocations should become similar once the heterogeneity arm is held fixed.

The exact-refit K-fold comparisons therefore remain the governing validation evidence throughout the analysis.

3.3 Abnormal-accrual measures and classification

Exact-K-fold scoring produces target-specific stacking weights. To construct the generated measure, these cross-validated weights are then applied to posterior draws from the corresponding candidate models refitted on the full analysis sample. Candidate draws are combined by sampling models according to the target-specific stacking vector, preserving model and parameter uncertainty rather than reducing each candidate to a point estimate. Thus, the validation target determines the benchmark-weighting reference frame, while the final generated measure uses the common full-sample estimation step. Let $\widehat{NDA}_{i,t}^V$ denote the posterior mean normal accrual under validation target V . The primary generated measure is

$$DA_{i,t}^{raw,V} = TA_{i,t} - \widehat{NDA}_{i,t}^V. \quad (6)$$

This is a model-generated proxy for deviation from the chosen normal-accrual reference frame. It is not interpreted as directly observed discretion.

Measurement propagation is evaluated by comparing row- and grouped-target versions of this object using two complementary statistics. Spearman correlation captures the stability of the full ranking. Jaccard similarity for matched top-five-percent sets captures agreement in the observations selected by an extreme screen. Using both avoids conflating global ordering with classification stability. A high rank correlation can coexist with substantial turnover near the tail cutoff, which is precisely the pattern relevant to studies that partition observations into high- and low-abnormality groups.

A second measurement object scales the raw residual by posterior uncertainty in the expected benchmark:

$$DA_{i,t}^{z,est,V} = \frac{DA_{i,t}^{raw,V}}{SD(\mu_{i,t}^{V,(s)})}. \quad (7)$$

The scaled measure asks whether a deviation is large relative to uncertainty about the expected normal accrual. It is not treated as the same construct expressed in different units because the denominator is itself estimated and target-dependent. Denominator floors, caps, and a predictive-uncertainty alternative therefore provide boundary evidence about how much of the scaled-score divergence comes from the residual numerator versus the uncertainty denominator.

3.4 Mechanism and sensitivity analyses

The mechanism analysis is organized around the strongest alternative explanations rather than a catalogue of generic robustness checks. The first simulation varies persistent firm heterogeneity in

$$TA_{i,t} = X_{i,t}\beta + \eta_{industry(i),t} + u_i + \epsilon_{i,t}, \quad (8)$$

with σ_u set to zero, 0.1, or 0.3. Panel length is also varied. The discriminating prediction is straightforward: if row-level validation rewards access to persistent same-firm structure, the row-minus-grouped premium for firm-random-effect candidates should be close to zero when $\sigma_u = 0$ and should increase sharply once $\sigma_u > 0$.

A second simulation block introduces autoregressive same-firm shocks over a grid of persistence parameters. This asks whether the empirical contrast could instead be produced by temporal dependence in the absence of persistent firm heterogeneity. The purpose is not to prove that time dependence is irrelevant. Accrual reversals and operating shocks make some temporal dependence expected. The test asks which mechanism better reproduces the observed reallocation and identifies boundary cells in which temporal structure may contribute.

The remaining analyses test specific links in the measurement chain. Fully marginal new-firm scoring asks whether the grouped conclusion survives a predictive rule that integrates over a new random effect. Pooled-only stacking tests whether accounting-family substitution explains the target difference after firm heterogeneity is removed from the candidate space. Fold-local preprocessing checks whether target differences are induced by transformation constants estimated outside a fold. Denominator variants test whether the uncertainty-scaled measurement-propagation result depends on the scaling convention rather than the residual numerator. Finally, descriptive regressions compare subsequent cash flow, earnings, and ROA across observations common to both tails and those selected only by one target. These regressions are used as discriminating observable traces of tail composition; they do not rank the validation targets by managerial intent or causal validity.

AI-assisted preparation.

A large language model was used during manuscript preparation for language editing, LaTeX assembly, consistency checking, drafting assistance, and code-review support used to audit correspondence between the manuscript and the analysis pipeline. The reported statistical analyses were executed by the documented R/Stan pipeline. The author reviewed the analytical choices, numerical outputs, citations, interpretations, and final text and remains fully responsible for the work.

4 Results

4.1 Benchmark allocation by validation target

Table 3 reports the primary benchmark-allocation result, beginning with the no-lookahead space. Row-level exact K-fold assigns 0.5927 of aggregate stacking weight to firm-random-effect benchmarks, compared with 0.0368 under grouped-firm validation, a row-minus-grouped reallocation of 0.5559. The retrospective ex-post space shows the same pattern: 0.6207 under row-level validation versus 0.0405 under grouped-firm validation, a difference of 0.5801. Put differently, the Firm-RE share under row-level validation is about fifteen to sixteen times its grouped-firm share in both spaces. The contrast is also visible at the leading-cell level. In the primary no-lookahead

Panel A: Aggregate heterogeneity weights

Table 3 Validation target and normal-accrual benchmark allocation

Space	Row Firm-RE	Grouped Firm-RE	Difference	Ratio
No-lookahead	0.5927	0.0368	0.5559	16.09
Ex-post	0.6207	0.0405	0.5801	15.31

Panel B: Leading benchmark cells

Space	Row-level leading cell	Grouped-firm leading cell
No-lookahead	Ball-Shivakumar, (0.4521)	Firm-RE No-lookahead, pooled (0.8332)
Ex-post	Ball-Shivakumar, (0.4375)	Firm-RE McNichols performance, pooled (0.8195)

Notes: Firm-RE columns report aggregate stacking weight assigned to firm-random-effect candidates. Difference is row-level minus grouped-firm weight. Parentheses in Panel B report the stacking weight of the leading benchmark cell.

space, a Ball-Shivakumar-style Firm-RE benchmark receives the largest row-level weight (0.4521), whereas grouped-firm validation concentrates on the no-lookahead pooled family (0.8332). The ex-post space yields the corresponding contrast between Ball-Shivakumar Firm-RE (0.4375) and pooled McNichols performance (0.8195).

The relevant inference is not that one accounting family dominates another. The discriminating feature is the heterogeneity arm. Row-level validation scores a held-out firm-year while other observations from that firm are usually represented in training; persistent firm structure can therefore improve the target-specific predictive density. Grouped-firm validation asks for a benchmark for an unseen firm, for which the fitted firm history is unavailable. The same candidate catalogue consequently generates different normal-accrual mixtures when the reference frame changes.

The simulations test whether this pattern is the mechanical consequence of using different folds or instead follows the proposed heterogeneity mechanism. In the static simulation with no firm-level structure, the average row-minus-grouped Firm-RE premium is -0.0033 ; in the temporal-dependence block the corresponding anchor is 0.0132 . Once persistent firm heterogeneity is introduced, the premium becomes large: 0.8622 and 0.9565 at the two positive heterogeneity levels in the static block, and 0.8336 and 0.9503 in the temporal block. Thus changing the held-out unit does not by itself create a systematic Firm-RE advantage. The advantage appears when the data-generating process contains same-firm structure that row-level validation can reuse. Temporal dependence can matter in high-persistence boundary cells, but it does not reproduce the empirical reallocation in the zero-heterogeneity anchor. The full simulation grid and Bayesian confirmation diagnostics are reported in Supplementary Table S2.

Two restricted comparisons probe the strongest remaining alternatives. First, the baseline grouped score evaluates Firm-RE candidates at the population level because the

Table 4 Abnormal-accrual overlap across validation targets

Space	Measure	N	Top n	Common	Jaccard	Spearman
No-lookahead	$ DA^{raw} $	4,494	225	184	0.6917	0.8291
No-lookahead	$ DA^{z,est} $	4,494	225	133	0.4196	0.7086
Ex-post	$ DA^{raw} $	3,936	197	155	0.6485	0.8307
Ex-post	$ DA^{z,est} $	3,936	197	108	0.3776	0.7110

Notes: Top-tail comparisons use matched top-five-percent sets. Common is the number of firm-years appearing in both row-level and grouped-firm top tails. Lower Jaccard values indicate greater divergence in tail composition.

held-out firm’s estimated random effect is unavailable. A fully marginal new-firm predictive density is stricter because it integrates over a newly drawn firm effect. Under that rule, the leading grouped benchmark is unchanged in both information spaces, while the already small grouped Firm-RE share falls from 0.0368 to approximately zero in the primary no-lookahead space and from 0.0405 to approximately zero ex post. Second, removing all Firm-RE candidates almost eliminates the row/grouped family reallocation: the pooled-only L1 shifts are 0.0151 no lookahead and 0.0141 ex post, and the same pooled family leads under both targets within each space. These tests reject a broad interpretation in which the result merely reflects unstable choice among accounting specifications. The evidence instead localizes the contrast in whether the measurement benchmark is permitted to exploit persistent firm-specific structure. Supplementary Table S3 reports both checks.

4.2 Abnormal-accrual rankings and tail overlap

The next stage asks whether the benchmark reallocation is consequential for the generated residual. Table 4 separates global ordering from extreme-case classification and again reports the primary no-lookahead space first. For raw abnormal accruals, row- and grouped-target rankings remain strongly related: the Spearman correlation is 0.8291 in the no-lookahead space and 0.8307 ex post. A global diagnostic alone would therefore suggest substantial agreement. Tail membership gives a different picture. Among the 225 observations in each no-lookahead top-five-percent set, 184 are common to both targets; among the 197 observations in each ex-post set, 155 are common. The corresponding Jaccard similarities are 0.6917 and 0.6485.

This pattern identifies where target dependence matters. The validation choice does not induce a wholesale reordering of firm-years. It changes enough benchmark predictions near the screening boundary to alter which observations are classified as extreme. That distinction is substantively important for empirical designs that use abnormal-accrual tails as sample-selection rules, treatment-like partitions, risk screens, or descriptive indicators. For such uses, a rank correlation near 0.83 does not imply that the underlying measurement decision is invariant.

The divergence is larger for the estimation-uncertainty-scaled object. Its rank correlations fall to 0.7086 in the no-lookahead space and 0.7110 ex post. Only 133 of 225 no-lookahead tail observations and 108 of 197 ex-post tail observations are common

across targets, producing Jaccard similarities of 0.4196 and 0.3776. The scaled result is informative but cannot be interpreted as a simple strengthening of the raw-residual result. The denominator is itself a target-dependent posterior uncertainty estimate. Floors and caps move some scaled-tail overlaps toward the raw-score values, whereas predictive-uncertainty scaling creates yet another classification object. The defensible inference is therefore layered: validation-target dependence is present in the raw residual numerator; uncertainty scaling can amplify or alter that dependence because it adds a second generated component. Supplementary Table S4 reports the denominator variants.

4.3 Boundary checks and tail composition

Fold-local preprocessing addresses a separate design threat. The baseline transformations are estimated within each declared measurement space, so winsorization and standardization constants can reflect observations later assigned to validation folds. Re-estimating those transformations inside each fold does not eliminate the focal contrast. The Firm-RE row-minus-grouped shift remains positive in both spaces, and the raw-score Jaccard moves from 0.6917 to 0.6854 in the primary no-lookahead space and from 0.6485 to 0.6148 ex post. The uncertainty-scaled Jaccards move to 0.4331 and 0.4173, respectively. The exact leading no-lookahead model is more preprocessing-sensitive than the aggregate heterogeneity arm, which is why the paper’s claim concerns reallocation between firm-specific and population-level structure rather than invariance of a particular top cell. Supplementary Table S4 provides the fold-local comparison.

The tail-composition diagnostics ask a different question: do the observations that move between target-specific tails leave distinguishable subsequent accounting traces? For the raw residual, firm-years common to both tails are followed by lower earnings and ROA, and row-only tail observations show the same directional pattern. Grouped-only tail observations instead have positive coefficients for subsequent CFO, earnings, and ROA in the corresponding descriptive regressions. For example, the future-ROA coefficients are -0.0220 for common-tail observations, -0.0249 for row-only observations, and 0.0396 for grouped-only observations. The separation weakens for the estimation-uncertainty-scaled object. These regressions are not construct validation for managerial intent and are not used to declare one target economically correct. They provide a narrower discriminating observable: changing the validation reference frame selects tail observations with different subsequent accounting composition rather than merely assigning different labels to otherwise exchangeable cases. Supplementary Table S5 reports the corresponding coefficient patterns; the accompanying source exports retain standard errors, p -values, multiplicity-adjusted q -values, and related diagnostics.

The empirical chain is therefore directional and bounded. The held-out unit first changes the value assigned to firm-specific heterogeneity; the resulting predictive weights change the normal-accrual benchmark; the benchmark change propagates into the residual; and the residual change affects extreme-case membership. Each link has an observable counterpart, and the mechanism checks distinguish this chain

from pooled-family instability, preprocessing alone, or a generic consequence of changing folds. This is the sense in which the validation target operates as part of the measurement mapping in this application.

5 Discussion

The methodological problem examined here arises whenever a latent notion of “normal” behavior is operationalized by a predictively selected benchmark and abnormality is defined as the residual from that benchmark. In such settings, validation has two roles. It evaluates predictive performance, but it can also determine the reference model from which the construct proxy is generated. The empirical application shows that changing only the held-out unit is sufficient to change that reference model: observed-firm interpolation rewards firm-specific structure, whereas out-of-firm prediction favors a population benchmark. Simulation and restricted-model evidence locate the difference primarily in persistent firm heterogeneity rather than arbitrary substitution among pooled accounting families.

This result extends, rather than replaces, the structured-validation literature. The established point is that row-level and group-level validation estimate different predictive tasks in clustered data. The additional measurement implication is conditional: when those task-specific scores are used to select or weight the benchmark that defines a residual, the predictive estimand can propagate into construct operationalization. The held-out unit is then not a neutral computational setting. It helps determine what reference frame the empirical measure calls “normal.” This implication is relevant beyond accruals to residual-generated measures of abnormal investment, abnormal expenditure, excess utilization, value added, or other deviations from predicted behavior, provided that benchmark selection is itself performance-based.

The classification results show why this distinction can remain hidden in conventional diagnostics. Raw abnormal-accrual rankings retain correlations around 0.83 across targets, yet the top-five-percent sets overlap much less completely. A researcher interested only in average rank agreement could therefore conclude that the measurement choice is minor, while a design built around extreme observations would analyze a materially different set of cases. Uncertainty scaling sharpens the warning because it introduces an estimated, target-dependent denominator; scaled residuals should be regarded as separate generated measurement objects, not as innocuous rescalings of the raw proxy.

The findings also define their own boundary conditions. Nothing here implies that grouped-firm validation is universally superior, that row-level validation is biased whenever repeated observations exist, or that abnormal accruals directly reveal managerial intent. A within-firm reference frame is coherent when benchmark performance should reward prediction using a represented firm’s history. An out-of-firm reference frame is coherent when benchmark performance should be evaluated with the focal firm’s history unavailable. In this design, those validation-specific performance weights are then applied to common full-sample refits when constructing the residual measure. The methodological requirement is correspondence between the intended construct mapping and the validation estimand. For residual-based proxies in clustered longitudinal data, researchers should therefore report and justify the held-out unit together

with the benchmark family, heterogeneity structure, information timing, uncertainty definition, and classification rule. When validation selects the benchmark from which a proxy is generated, validation design is part of measurement design.

Data availability

The firm-year financial-statement data were obtained from FiinPro/FiinPro-X under a commercial licence and cannot be redistributed publicly. Access to the underlying source data requires an appropriate licence from the data provider. Non-proprietary aggregate outputs used to assemble the reported tables can be supplied to the editors or reviewers subject to the journal’s double-anonymous procedure and the applicable data-licence restrictions.

Code availability

The analysis code is maintained in a version-controlled repository. To preserve double-anonymous review, identifying repository links are omitted from the blinded manuscript and supplementary material. An anonymized code snapshot can be supplied through the journal submission system upon request; public repository information can be added to the final accepted version.

References

- Adcock R, Collier D (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95(3):529–546. <https://doi.org/10.1017/S0003055401003100>
- Adin A, Krainski ET, Lenzi A, et al (2024) Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics* 62:100843. <https://doi.org/10.1016/j.spasta.2024.100843>
- Ball R, Shivakumar L (2006) The role of accruals in asymmetrically timely gain and loss recognition. *Journal of Accounting Research* 44(2):207–242. <https://doi.org/10.1111/j.1475-679X.2006.00198.x>
- Bürkner PC, Gabry J, Vehtari A (2020) Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation* 90(14):2499–2523. <https://doi.org/10.1080/00949655.2020.1783262>
- Bates S, Hastie T, Tibshirani R (2024) Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association* 119(546):1434–1445. <https://doi.org/10.1080/01621459.2023.2197686>
- Breuer M, Schütt HH (2023) Accounting for uncertainty: an application of Bayesian methods to accruals models. *Review of Accounting Studies* 28(2):726–768. <https://doi.org/10.1007/s11142-021-09654-0>

- Byzalov D, Basu S (2024) The misuse of regression-based x-scores as dependent variables. *Journal of Accounting and Economics* 77(2–3):101643. <https://doi.org/10.1016/j.jacceco.2023.101643>
- Chen W, Hribar P, Melessa S (2018) Incorrect inferences when using residuals as dependent variables. *Journal of Accounting Research* 56(3):751–796. <https://doi.org/10.1111/1475-679X.12195>
- Chen W, Hribar P, Melessa S (2023) Standard error biases when using generated regressors in accounting research. *Journal of Accounting Research* 61(2):531–569. <https://doi.org/10.1111/1475-679X.12470>
- Christodoulou D, Ma L, Vasnev A (2018) Inference-in-residuals as an estimation method for earnings management. *Abacus* 54(2):154–180. <https://doi.org/10.1111/abac.12121>
- Colas B, Brousseau C (2025) Industry classification misfits: identification, consequences and guidance. *Review of Accounting Studies* 30(4):3295–3343. <https://doi.org/10.1007/s11142-025-09886-4>
- Collins DW, Pungaliya RS, Vih AM (2017) The effects of firm growth and model specification choices on tests of earnings management in quarterly settings. *The Accounting Review* 92(2):69–100. <https://doi.org/10.2308/accr-51551>
- Dechow P, Ge W, Schrand C (2010) Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics* 50(2):344–401. <https://doi.org/10.1016/j.jacceco.2010.09.001>
- Dechow PM, Dichev ID (2002) The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review* 77(s-1):35–59. <https://doi.org/10.2308/accr.2002.77.s-1.35>
- Dechow PM, Sloan RG, Sweeney AP (1995) Detecting earnings management. *The Accounting Review* 70(2):193–225. URL <http://www.jstor.org/stable/248303>
- Dichev I, Owens E (2025) Accrual duration. *Review of Accounting Studies* 30(1):611–649. <https://doi.org/10.1007/s11142-024-09826-8>
- Ecker F, Francis J, Olsson P, et al (2013) Estimation sample selection for discretionary accruals models. *Journal of Accounting and Economics* 56(2):190–211. <https://doi.org/10.1016/j.jacceco.2013.07.001>
- Fiechter P, Wallmeier M (2026) Estimating discretionary accruals in the cross-section of firms: A reinterpretation of the jones model and its variants. *Accounting Open* 2:100006. <https://doi.org/10.1016/j.accop.2026.100006>

- Frankel RM, Sun Y (2018) Predicting accruals based on cash-flow properties. *The Accounting Review* 93(5):165–186. <https://doi.org/10.2308/accr-52001>
- Hribar P, Collins DW (2002) Errors in estimating accruals: Implications for empirical research. *Journal of Accounting Research* 40(1):105–134. <https://doi.org/10.1111/1475-679X.00041>
- Jackson AB (2018) Discretionary accruals: Earnings management ... or not? *Abacus* 54(2):136–153. <https://doi.org/10.1111/abac.12117>
- Jones JJ (1991) Earnings management during import relief investigations. *Journal of Accounting Research* 29(2):193–228. <https://doi.org/10.2307/2491047>
- Kothari SP, Leone AJ, Wasley CE (2005) Performance matched discretionary accrual measures. *Journal of Accounting and Economics* 39(1):163–197. <https://doi.org/10.1016/j.jacceco.2004.11.002>
- Liu Z, Van Niekerk J, Rue H (2025) Leave-group-out cross-validation for latent Gaussian models. *SORT-Statistics and Operations Research Transactions* 49(1):121–146. <https://doi.org/10.57645/20.8080.02.25>
- McNichols MF (2000) Research design issues in earnings management studies. *Journal of Accounting and Public Policy* 19(4):313–345. [https://doi.org/10.1016/S0278-4254\(00\)00018-1](https://doi.org/10.1016/S0278-4254(00)00018-1)
- McNichols MF (2002) Discussion of the quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review* 77(s-1):61–69. <https://doi.org/10.2308/accr.2002.77.s-1.61>
- McNichols MF, Stubben SR (2018) Research design issues in studies using discretionary accruals. *Abacus* 54(2):227–246. <https://doi.org/10.1111/abac.12128>
- Merkle EC, Furr D, Rabe-Hesketh S (2019) Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika* 84(3):802–829. <https://doi.org/10.1007/s11336-019-09679-0>
- Murphy KM, Topel RH (1985) Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics* 3(4):370–379. <https://doi.org/10.1080/07350015.1985.10509471>
- Nezlobin AA, Sloan RG, Giedt JZ (2022) Construct validity in accruals quality research. *The Accounting Review* 97(5):377–398. <https://doi.org/10.2308/TAR-2019-0213>
- Owens EL, Wu JS, Zimmerman J (2017) Idiosyncratic shocks to firm underlying economics and abnormal accruals. *The Accounting Review* 92(2):183–219. <https://doi.org/10.2308/accr-51523>

- Pagan A (1984) Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* 25(1):221–247. <https://doi.org/10.2307/2648877>
- Qiu J, Lake DE, Chernyavskiy P, et al (2025) Fast leave-one-cluster-out cross-validation using clustered network information criterion. *Statistical Methods in Medical Research* 34(7):1413–1430. <https://doi.org/10.1177/09622802251345486>
- Rabinowicz A, Rosset S (2022) Cross-validation for correlated data. *Journal of the American Statistical Association* 117(538):718–731. <https://doi.org/10.1080/01621459.2020.1801451>
- Roberts DR, Bahn V, Ciuti S, et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8):913–929. <https://doi.org/10.1111/ecog.02881>
- Vaida F, Blanchard S (2005) Conditional akaike information for mixed-effects models. *Biometrika* 92(2):351–370. <https://doi.org/10.1093/biomet/92.2.351>
- Yao Y, Vehtari A, Simpson D, et al (2018) Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis* 13(3):917 – 1007. <https://doi.org/10.1214/17-BA1091>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.pdf](#)