



Human-, GRM-, and LLM-Generated School-Improvement Leadership Scores from Missouri: Structural Bias and Concurrent and Predictive Validity

Seth B. Hunter
George Mason University

David S. Woo
University of Utah

Katherine M. Bowser
National Council on
Teacher Quality

Syeda Sabrina Akter
George Mason University

Thomas Hairston
University of Missouri

Eter Mjavanadze
George Mason University

Antonios
Anastasopoulos
George Mason University

Sarah Hairston
Columbia Academy for
Learning and Enrichment

Head or assistant principal evaluation systems are widely deployed but rarely examined for validity, and AI scoring tools have entered K–12 workflows ahead of the evidence. We examine human supervisor ratings, Graded Response Model (GRM) rescaling, and large language model (LLM) scoring of 1,385 yearly leader-authored, standards-based, retrospective accounts of school improvement efforts and validate each against five school outcomes. Some structural school characteristics predict human-based and LLM scores, but often in different directions. As predictors, human and GRM scores often converge but predict few outcomes, though GRM may recover aspects of instructional leadership captured by student achievement scores. LLM scores diverge and uniquely predict staffing stability—teacher retention—a signal that survives adjustment for surface text features. Higher human-based scores predict worse staffing stability in the highest-racial-minority and highest-FRPL enrollment schools, while LLM scores show no such inversion. Scoring method choice may be equity-consequential and capture different aspects of school-improvement leadership.

VERSION: October 2026

Suggested citation: Hunter, Seth B., David S. Woo, Katherine M. Bowser, Syeda Sabrina Akter, Thomas Hairston, Eter Mjavanadze, Antonios Anastasopoulos, and Sarah Hairston. (2026). Human-, GRM-, and LLM-Generated School-Improvement Leadership Scores from Missouri: Structural Bias and Concurrent and Predictive Validity. (EdWorkingPaper: 26-1603). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/bv8c-s810>

**Human-, GRM-, and LLM-Generated School-Improvement Leadership Scores from
Missouri: Structural Bias and Concurrent and Predictive Validity**

Seth B Hunter, Corresponding Author
ORCiD 0000-0002-3051-872X
George Mason University
Associate Professor of Education Leadership
4400 University Dr, MS4C2
Fairfax, VA 22030

David S Woo
ORCiD 0000-0001-5728-6562
University of Utah
Assistant Professor
1721 Campus Center Drive
Salt Lake City, UT 84112

Katherine M Bowser
ORCiD 0009-0004-4771-2272
National Council on Teacher Quality
1032 15th Street NW #242
Washington D.C., 20005

Syeda Sabrina Akter
George Mason University
PhD Student
4400 University Dr, Research Hall 405
Fairfax, VA 22030

Thomas Hairston
ORCiD 0009-0007-5378-2041
University of Missouri
Director, Network for Educator Effectiveness
2800 Maguire Boulevard
Columbia, MO 65211

Eter Mjavanadze
ORCiD 0000-0001-9524-1285
George Mason University
PhD Student
4400 University Dr, MS4C2
Fairfax, VA 22030

Antonios Anastasopoulos
ORCiD 0000-0002-8544-246X
George Mason University

Associate Professor of Computer Science
4400 University Dr, Research Hall 409
Fairfax, VA 22030

Sarah Hairston

Columbia Academy for Learning and Enrichment
Executive Director of Secondary Programs
2201 W Nifong Boulevard
Columbia, MO 65203

Abstract

Head or assistant principal evaluation systems are widely deployed but rarely examined for validity, and AI scoring tools have entered K–12 workflows ahead of the evidence. We examine human supervisor ratings, Graded Response Model (GRM) rescaling, and large language model (LLM) scoring of 1,385 yearly leader-authored, standards-based, retrospective accounts of school improvement efforts and validate each against five school outcomes. Some structural school characteristics predict human-based and LLM scores, but often in different directions. As predictors, human and GRM scores often converge but predict few outcomes, though GRM may recover aspects of instructional leadership captured by student achievement scores. LLM scores diverge and uniquely predict staffing stability—teacher retention—a signal that survives adjustment for surface text features. Higher human-based scores predict worse staffing stability in the highest-racial-minority and highest-FRPL enrollment schools, while LLM scores show no such inversion. Scoring method choice may be equity-consequential and capture different aspects of school-improvement leadership.

School improvement is fundamentally a leadership endeavor. Principals, second only to teachers among school influences on student outcomes, primarily influence school improvement through the organizational conditions leaders create (Leithwood, Harris, & Hopkins, 2008; Grissom, Egalite, & Lindsay, 2021). How we measure the qualities of such leadership affects principal development and personnel decisions, and ultimately the students affected by those decisions. We examine scores derived from one Missouri-based leadership artifact: the Building Improvement Plan Process Organizer (BIP-PO), a retrospective, self-authored document in which a leader (i.e., assistant or head principal) describes how they led school improvement processes over one year. The BIP-PO is not a summative evaluation of leadership nor a judgment of whether improvement efforts succeeded; it is a structured account of how the leader managed the systems and collaborative processes undergirding school improvement, functions closely aligned with Missouri's Department of Elementary and Secondary Education (DESE) Leader Standards, and scored within the Network for Educator Effectiveness (NEE) evaluation framework (Hunter & Bowser, 2026). Whether the scores assigned to these documents exhibit concurrent and predictive validity is a psychometric and practical question.

Field-based educator evaluation systems now routinely score leadership documents by rubric, a process consuming considerable time and money. Nearly every state reformed leadership evaluation after Race to the Top, though components, processes, and consequences vary widely (Donaldson et al., 2021). The resulting scores have nonetheless been subject to remarkably little criterion validation, yet in many systems they remain a primary quantitative indicator of how well leadership functions. We ask whether BIP-PO scores exhibit criterion validity when examined against a battery of school outcomes, and whether the answer changes depending on who, or what, generates the scores. We compare three scoring methods: human

supervisor ratings, psychometric rescaling via the Graded Response Model (GRM), and ratings generated by a large language model (LLM). The closest precedent for our work is Grissom, Blissett, and Mitani (GBM) (2018), who validated Tennessee supervisor ratings of principals against school and principal value-added, student achievement growth, and teachers' survey ratings of school leadership; theirs remains the lone large-scale criterion-validity examination of principal performance ratings, and it relied on a single scoring method.

Human ratings of leadership have well-documented limitations: ceiling compression, plausible score inflation, and confounding with school context, specifically student demographics (GBM, 2018). Single-rater subjectivity is not a K-12 anomaly; organizational psychology has long documented how cognitive biases, limited observational access, and contextual cues distort ratings across sectors (DeNisi & Murphy, 2017). Our question is not whether human ratings are imperfect but whether alternative scoring methods offer different, and potentially more valid, signals.

Two contemporary developments shape our questions. The first is psychometric: item response theory, including the Graded Response Model (Samejima, 1969), estimates a latent trait that adjusts for differential item functioning and can sharpen discrimination among leaders, mitigating ceiling compression. Because the GRM operates on human-generated item responses, however, it adds no new information about the performance and inherits many biases from the underlying human scores. The second is technological: LLMs can read the same documents that human supervisors evaluate and can generate rubric-aligned scores from the text itself, independent of human judgment, and the marketplace is already positioning them as cost-effective alternatives to human scoring in educational assessment (Kaufman et al., 2025; Edthena, 2025). Whether the LLM restates the measurement problem or helps resolve it is an

empirical question. We treat GRM and LLM scoring as two additional vantage points from which to examine the same construct human scores aim to capture.

The equity stakes of LLM scoring are not trivial. Automated essay scoring (AES) research has documented that LLMs can encode demographic characteristics through surface linguistic features, training data imbalances, and sensitivity to dialect variation (Loukina et al., 2019; Lee et al., 2024; Bender et al., 2021). Deane (2013) articulated the foundational concern: AES may capture text-production skills rather than the substantive quality of ideas. Whether those patterns transfer to rubric-prompted LLM scoring of principal-authored improvement documents remains unsettled.

To our knowledge, this is the first study to apply human, psychometric, and LLM scoring to the same set of leadership documents and validate each method against a common outcome battery at scale. We make three contributions. First, we are the first to examine leader-authored retrospective school improvement documents as a measure of school-improvement leadership and to validate the quality scores assigned to those documents against a multi-outcome battery; prior work on structurally similar documents extracted descriptive content categories but did not examine criterion validity (Sun et al., 2019). Second, it extends GBM (2018) from a single human-rating stream to a formal three-method criterion-validity comparison, a contribution that travels beyond the BIP-PO to any performance domain in which human and AES coexist. Third, it centers equity as a constitutive analytic dimension: by jointly examining predictive validity, structural bias, and text-feature mediation across all three scoring methods, we show that the equity profile of evaluation scores depends on who or what does the scoring, a finding with direct implications for the design and oversight of educator evaluation systems that incorporate AI. The gaps we fill matter for research, practice, and policy, as education agencies consider

integrating AI into evaluation systems with little evidence on how automated scores of documents perform relative to the human judgments they would supplement or replace.

Our analytic design follows argument-based validation logic (Kane, 2013), treating each scoring method as generating claims about school-improvement leadership quality tested against external criteria. Notably, we assess the validity of scores derived from BIP-POs, not necessarily the quality of what leaders enacted in their buildings. The BIP-PO is a leader's account of the improvement process; our outcome battery tests whether variation in the quality of that account, as scored by different methods, corresponds to variation in measurable school performance.

Four research questions guide the analysis. First, to what extent do Human, GRM, and LLM BIP-PO scores show equivalent patterns of concurrent and predictive validity across five school outcomes? We estimate relationships using various fixed effects (FE), structural control variables, and use seemingly unrelated estimation to test whether validity coefficients differ across scoring methods. Second, to what extent does the validity of each scoring method vary across observable school composition contexts? Our first question treats the scores as given and asks what they predict; the second asks what the scores encode. Third, AES research has long investigated whether LLM scores capture the construct of interest or instead reflect surface properties of the text—length, lexical sophistication, syntactic complexity—that correlate with but do not constitute writing quality (Page, 1966; Shermis & Burstein, 2013; Mizumoto & Eguchi, 2023). We therefore ask to what extent LLM scores encode school demographic characteristics differently than Human and GRM scores. Finally, we ask the extent to which surface text features account for the LLM's validity profile, and whether it persists after surface-feature adjustment.

Literature Review

School-Improvement Leadership as a Multidimensional Construct

The mechanisms through which principals exert influence are not uniform. Grissom, Egalite, and Lindsay (2021) identify four broad channels through which principals affect schools: engaging in instructionally focused interactions with teachers, building a productive organizational climate, facilitating strategic personnel decisions, and managing the external environment. Day et al. (2016), drawing on a mixed-methods study of effective and improving schools in England, reached a complementary conclusion: successful principals combined instructional and transformational practices in ways sensitive to school context and the phase of improvement. The core insight is that leadership effectiveness is not unidimensional, and its multidimensionality has direct measurement implications. If leaders shape schools through multiple channels, any single outcome measure may capture only part of the leader's influence. Thus, a credible validity study of any leadership scoring instrument must deploy a battery of outcomes spanning theoretically relevant dimensions of influence. Different scoring methods may also privilege different elements; thus, variation across score-generators in the predicted outcome domains is informative.

Five-Outcome Battery Justification

The five outcomes we employ map to a theoretically or empirically grounded dimension of school-improvement leadership. Together, the battery spans student learning (achievement growth and value-added), what we characterize as “organizational health” (staffing stability), teacher perceptions about leadership effectiveness (teacher surveys), and student perceptions about teaching (student surveys). Achievement and school-level value-added to achievement relate to the instructional channel of school-improvement leadership (Branch et al., 2013; Coelli & Green, 2012; Dhuey & Smith, 2014). We include both achievement growth and a value-added

measure because they capture conceptually distinct features of the learning environment: growth reflects the cumulative conditions of a school, including prior achievement and student composition, whereas school-level value-added attempts to isolate within-year within-school growth net of student baseline performance (Reardon & Raudenbush, 2009; Ehlert et al., 2014). We adopt school-level rather than principal-level value-added for two reasons. The measurement reason is that recent work casts doubt on the validity and reliability of the principal-level estimator (Bartanen et al., 2024; Bartanen & Husain, 2022). The conceptual reason is that the BIP-PO asks principals to account for how they led a school's improvement work, not for their contribution net of teachers, students, and context; a criterion that partials out schoolwide mechanisms would partial out what the instrument documents.

Staffing stability represents the organizational-management dimension. Boyd et al. (2011) demonstrated that working conditions, shaped substantially by principals, are among the strongest predictors of teacher turnover, and Kraft, Marinell, and Yee (2016) found that principal leadership quality was the single most important school-level factor in teachers' decisions to stay in schools. Bêteille, Kalogrides, and Loeb (2012) traced the chain further, finding that principals who effectively manage talent retain higher-performing teachers and counsel out lower-performing ones, improving the teaching corps composition. Teacher retention thus serves as a downstream proxy for the organizational conditions that principals create and sustain.

Teacher and student surveys capture perceptions of leadership and instruction. The NEE Teacher Survey asks teachers to rate each leader's effectiveness; Jones et al. (2024) documented its reliability and dimensionality. Student surveys concern teaching. We favor student reports on instruction rather than the classroom observation score generated by the leaders under examination, which would confound criterion and construct.

What We Know About the Validity of Scoring Leadership Documents

Empirical evidence on the criterion validity of scored leadership documents remains thin, making GBM foundational. Using data from Tennessee’s statewide principal evaluation system, GBM examined whether rubric scores assigned during supervisor “site visits” predicted student achievement gains. GBM linked these scores to school and principal value-added estimates, student achievement growth, and two independent surveys of school leadership and climate, testing whether the scores exhibited criterion validity across multiple outcomes. Their findings affirmed concurrent validity while raising questions about other measurement properties; their scores correlated negatively with school poverty, with the same principal receiving lower ratings in years spent in higher-poverty schools, a pattern other performance measures did not explain. Our study extends GBM’s framework, uses a different class of scored documents, and employs a broader set of score generators. Our design lets us ask not only whether BIP-PO scores have criterion validity, but also whether the score generation method affects validity patterns.

Scoring School-Improvement Documents: Plans, Process, and the BIP-PO

The BIP-PO occupies an unusual position among scored educational performance documents. School improvement plans are prospective documents articulating goals and strategies for the coming year; the BIP-PO is a retrospective account of the schoolwide processes leaders led. These three axes of difference (temporal orientation, authorship, and content focus) distinguish the BIP-PO from both the plan-quality literature and the observation-based evaluation literature.

To better understand BIP-PO's purposes, we situate it within related but distinct work. Meyers and VanGronigen (2019) analyzed 364 short-cycle prospective school improvement plans from low-performing schools and found that most exhibited satisficing behavior: they met minimum procedural requirements without articulating the specific, actionable leadership processes that research links to school improvement. VanGronigen and Meyers (2020) corroborated this pattern, documenting weak overall prospective plan quality that improved only nominally over time. These studies focus on plan quality as the construct of interest; the BIP-PO differs because it focuses on leadership process quality.

The closest contemporary analog to our documents and methodology is research by Min Sun and colleagues. Sun et al. (2013) and Sun et al. (2019) used latent Dirichlet allocation to extract topic distributions from instructional improvement plans, treating the resulting topic measures as descriptive content variables. Sun et al. (2019) found that vaguely written tasks predicted worse student achievement, while specific, data-focused tasks predicted better outcomes. Sun and colleagues interpreted these text-derived measures as capturing “specificity and concreteness of writing,” providing early evidence that surface textual properties of leadership documents covary with school outcomes. More recently, Liu and Sun (2025) used AI to rate the quality of instructional improvement documents. Theirs is the closest methodological precedent for LLM-based scoring of leadership documents, although their focus on plan quality rather than process quality preserves the distinction we emphasize.

Psychometric Rescaling: IRT and the Graded Response Model

GBM (2018) and cross-sector research on performance appraisal document substantial ceiling compression in human job performance ratings (DeNisi and Murphy, 2017). One response to ceiling compression is to apply item response theory (IRT) models to recover a latent

performance continuum that the original ordinal scale compresses. The GRM is the IRT framework most commonly applied to ordinal ratings generated by rubric-based evaluations (Samejima, 1969). The GRM treats each rubric item as providing probabilistic information about an underlying latent trait, estimating item difficulty (threshold) and discrimination parameters that together define the measurement precision of each rubric dimension at different points along the latent continuum (see Online Appendix A for more technical details). We follow recent work applying GRM to teacher performance ratings for similar reasons, though we are unaware of other GRM applications to leader performance ratings (Edwards & Soland, 2024; Jones & Bergin, 2019; Kraft et al., 2020; Wind & Jones, 2019).

Automated and AI Scoring: From AES to LLMs

Page (1966) demonstrated that simple text features (e.g., word count, sentence length, and vocabulary diversity) predicted human essay scores with correlations rivaling inter-rater reliability. The AES tradition that followed expanded the feature set and modeling sophistication, while retaining a core reliance on surface-level text properties as proxies for writing quality (Shermis & Burstein, 2013). Perelman (2014) critiqued this reliance on surface properties by showing that AES systems could be deceived by texts that were long, syntactically complex, and lexically diverse but semantically incoherent, exposing the gap between what AES measures and what human scorers intend to reward. The critique crystallized a concern that persists into the LLM era: how much automated scoring captures substance (argumentation, reasoning, domain knowledge) versus surface (fluency, length, mechanical correctness). The emergence of LLMs represents a categorical shift in AES. Traditional AES relies on engineered features and statistical models trained on human-scored exemplars; LLMs can instead be prompted with

rubric criteria and asked to apply them to novel texts in a manner that at least superficially resembles human rubric application (Mizumoto & Eguchi, 2023; Cai & Yan, 2024).

Equity and Fairness in Scoring

Structural bias in performance ratings is a central threat to the consequential validity of any scoring system. Messick (1989) argued that the social consequences of score use are an integral component of validity; a scoring method whose criterion validity varies systematically across demographic groups fails validity tests. From this consequential-validity perspective, the choice of scoring method shapes which leaders are identified as effective in which school contexts, with downstream implications for career trajectories, resource allocation, and, ultimately, the students those leaders serve. Although equity-related evidence specific to leader ratings is limited, GBM suggests school factors outside the leader's control, including school poverty, may systematically bias ratings. Thus, the human ratings serving as the statistical baseline in our study are not bias-free. If LLM scores diverge from human scores, the divergence may reflect the removal of human bias, the introduction of algorithmic bias, or both.

Algorithmic fairness research raises a parallel set of concerns. Obermeyer et al. (2019) documented racial bias in a widely used healthcare algorithm that used cost as a proxy for health need, showing how seemingly neutral design choices can encode structural inequality. Loukina et al. (2019) and Sap et al. (2019) provided empirical evidence that AES performs differently across racial and dialectal groups, raising concerns about measurement invariance when textual style covaries with demographics. Education-specific evidence on algorithmic bias in scoring is emerging, suggesting that concerns about equity, fairness, and bias are warranted (Kwako and Ormerod, 2024; Lee et al., 2024; Warr et al., 2025). Ultimately, neither human nor algorithmic

scoring can be presumed fair *a priori*; both require empirical investigation of differential validity across the demographic groups and school contexts that structure educational inequality.

Study Setting

We use administrative data from NEE, a nonprofit principal and teacher evaluation and development system serving predominantly rural Missouri school districts (Hunter & Bowser, 2026). Districts adopt the system voluntarily, and by the study window (2015–16 through 2021–22), NEE served more than 200 districts statewide. Missouri requires each district to maintain a Continuous School Improvement Plan and, through it, school-level Building Improvement Plans (BIPs), but the state regulates neither the BIP’s form, documentation, nor its evaluation. To fill that gap, NEE supplies the BIP Process Organizer (BIP-PO); adoption is at district discretion.

The BIP-PO is not the plan. The BIP is a prospective, locally structured planning document oriented to the school’s improvement work; the BIP-PO is a retrospective, evaluatee-authored (i.e., leader-authored) account, addressed to the leader’s supervisor, in which the leader answers seven open-text prompts about how they led that plan’s development, implementation, monitoring, and communication over the year just completed or nearly completed. It is therefore oriented toward school-wide improvement rather than leaders’ individual contributions net of teachers, students, and context. Leaders enter their responses directly into the NEE Data Tool, and a single leader supervisor, typically the superintendent or another district leader designated by the local board, reviews and scores each element against the BIP-PO rubric.

The rubric elements assess: (1) the principal leads development of the BIP; (2) the BIP is collaboratively developed; (3) BIP objectives are aligned with CSIP goals; (4) the BIP includes measurable objectives with baseline data; (5) the BIP includes research-based implementation strategies; (6) plans for monitoring the BIP; and (7) plan for and evidence of sharing the

monitored BIP results. Each element is rated on a 0-to-4 rubric, with written descriptors anchored at 0 (does not meet standard), 2, and 4 (exceeds standard); supervisors may also assign the intermediate values of 1 and 3. However, the analytic sample contains no element-level zeros (Table 1 and Online Appendix A). Element-level ratings are the inputs to all three scoring methods we examine: the human composite (mean of element-level supervisor ratings), the GRM latent trait estimated from the same element-level ratings, and the LLM scores generated from the same narrative text the supervisor reads. Online Appendix A details Missouri’s school-improvement regulations, NEE’s origins and standards alignment, and the training and development cycle surrounding the BIP-PO.

Data

Our analytic dataset draws on four datasets: (1) NEE administrative records, including the BIP-PO corpus and supervisor-assigned element-level scores; (2) DESE linkages providing school, teacher, and student characteristics and student achievement data; (3) the NEE Teacher Survey, which captures teacher ratings of leader-specific effectiveness; and (4) the NEE Student Survey, which captures student ratings of their specific teacher’s practices.

Sample Construction

Our panel spans the 2015–16 through 2021–22 school years (hereafter 2016–2022), excluding 2019–20 because spring-2020 COVID-19 closures disrupted both the NEE evaluation cycle and standardized assessments. The BIP-PO universe comprises 1,733 principal-year observations from 731 unique principals, 444 schools, and 107 districts; the regression-eligible analytic sample is 1,385 principal-year observations from 103 districts after restricting to complete predictor data and nonmissing outcomes.

Four sample and measurement concerns arise from this administrative structure. First, district attrition from the NEE system produces an unbalanced panel. The second is multi-principal buildings, where roughly 34% of observations come from school years with two or more evaluated principals who share school-level outcomes, which may dilute relationships with school-level outcomes. The third is the 2019–20 COVID-19 exclusion, and the fourth is selective survey coverage. Online Appendix A documents each concern, reports diagnostic evidence, and presents the robustness analyses that address them (Tables A1 through A8); none alters the substantive conclusions reported below.

Measures

We standardize human, GRM, and LLM scores within school-year to enable cross-method comparisons on a common metric. Online Appendix A details the construction of each predictor; we summarize the operational features here. The mean pre-standardized human score across available elements for each leader-year in our analytic sample is 3.66 ($SD = 0.50$), indicating substantial ceiling compression.¹ GRM theta scores provide a wider range ($M = 0.01$, $SD = 0.83$; range -2.63 to 0.91) than raw human ratings, partially decompressing scores at the ceiling. The correlation between Human and GRM scores is $r = .92$, reflecting their shared dependence on supervisor judgment.

We score the same narrative text against the same rubric criteria using GPT-4o-mini under four prompting configurations that vary along two dimensions: provision of scored exemplars (zero-shot; few-shot from a fixed exemplar pool; or few-shot with a principal-linked sampling frame) and, for the latter, which sampling frame was used (within principal-year across

¹ A complete-case variant requiring all seven elements to be scored within each BIP-PO yields substantively unchanged results (Online Appendix A, Table A8).

elements; within principal-element across years).² The four cross-configuration correlations range from $r = .84$ to $.91$ (Online Appendix Table A9). We average the four within-year-standardized configuration scores to form a composite LLM Index ($M = 0.01$, $SD = 0.95$), which serves as our primary LLM predictor. This approach reduces configuration-specific noise and avoids selecting a single prompting strategy.³

Text Features

Any AI-based scoring system raises the question of whether it captures substantive rubric content or surface text properties. We extract a battery of 12 text features across four dimensions widely recognized in the AES literature as the operative axes of surface text variation: production/length, lexical sophistication, readability, syntactic complexity, and topic-content alignment (Page, 1966; Shermis & Burstein, 2013; Crossley & McNamara, 2012). Full definitions, extraction procedures, and parameter choices appear in Online Appendix A. We reduce the 12-dimensional feature space via principal-factor extraction and varimax rotation. We retain three factors: Volume (loading on word count, sentence count, and unique word count), Sophistication (loading on average word length, content-word ratio, and parse depth), and Rubric Overlap (loading on lexical overlap measures with the rubric text) (see Online Appendix A and Table A11).

School Outcome Measures

Each outcome is constructed as a within-year standardized composite of its constituent indicators, thereby reducing the multiple hypotheses testing (MHT) burden and improving

² Online Appendix D reports verbatim prompt text. Configuration details and exemplar-source definitions are in Online Appendix A.

³ The LLM 7-element case variant correlates with the less-than-7 element cases at $r = .991$ (Online Appendix A, Table A9).

reliability by averaging out component-specific measurement error. First, a standardized mean of school-average math and reading proficiency in tested grades, reflecting the composite academic performance levels of students in each leader's school. Second, a standardized composite mean of unshrunk school-level value-added estimates in math and reading, derived from student-level models controlling for prior achievement and student demographics (Reardon & Raudenbush, 2009; Ehler et al., 2014).⁴ Third, a standardized composite mean of two retention indicators: the proportion of teachers and the proportion of leaders (head and assistant principals) who remain in the school (Boyd et al., 2011). Fourth, the leader-year mean of teacher ratings of their specific principal's leadership effectiveness. Teacher survey measures are available for 1,014 of 1,385 analytic observations, with a modest overrepresentation of high-poverty and high-minority schools.⁵ Finally, the school-year average of student ratings of their teacher's practices. Each survey pertains to a single teacher, and students cannot be linked across surveys, so the index reflects average teacher quality in the building rather than a direct measure of principal effectiveness; we treat it as a distal validity criterion. Student survey measures are available for 978 of 1,385 analytic observations (71%).⁶ For analyses linking current-year scores to following-year outcomes, we link each leader-year observation to the same five outcomes measured in the following year ($t+1$). Future-year analytic samples are 13 to 48% smaller than same-year samples (mean and median reduction $\approx 33\%$).

⁴ We use unshrunk rather than Empirical Bayes (EB) estimates as EB shrinkage disproportionately attenuates estimates for smaller schools, which predominate in this rural sample; Online Appendix C (Table C5) reports an EB-shrunk sensitivity test.

⁵ Normalized differences are below the meaningful-difference threshold; see the Teacher Survey row of Online Appendix Table A7.

⁶ See the Student Survey row in Online Appendix Table A7 for composition comparisons.

Control Variables

Our preferred specification includes a lean set of school-level controls following GBM (2018). The lean set comprises five structural school-level characteristics: the proportion of students receiving free or reduced-price lunch (FRPL), the proportion of nonwhite students, the number of teachers, a high school indicator, and a rural indicator. In our preferred specification (District FE + Year FE), we control for lagged achievement only for the achievement outcome; a separate Year FE specification includes no lagged-achievement control for any outcome. We exclude lagged achievement from school value-added models because of the mechanical relationship between prior achievement and value-added estimation. In our preferred specification, we follow GBM (2018) by also excluding it from staffing stability, teacher survey, and student survey, where including it may overcontrol for an indirect pathway.

The Full Controls specification includes lagged achievement for staffing stability and survey outcomes and adds eight school-year staff demographic characteristics: average teacher experience, the proportion of teachers with advanced degrees (master's degree or higher), the proportions of nonwhite and male teachers, and the corresponding administrator characteristics (average experience, proportion with advanced degrees, proportion nonwhite, and proportion male). Table 1 reports descriptive statistics for all analytic variables.

Methods

Concurrent Validity

Our primary analysis estimates the association between each scoring approach and concurrent school outcomes. For each outcome Y and each scoring approach $S \in \{\text{Human, GRM, LLM}\}$, we estimate:

$$Y_{ist} = \beta_1 S_{it} + X_{it}\gamma + \alpha_d + \delta_t + \varepsilon_{ist} \quad (1)$$

where Y_{ist} is the outcome for principal i in school s and year t ; S_{it} is the within-year-standardized predictor score; X_{it} is a vector of lean controls; α_d represents district fixed effects; δ_t represents year fixed effects; and ε_{ist} is the error term. Standard errors are clustered at the district level. The coefficient β_1 is interpretable as the within-district, across-year association between a one-standard-deviation increase in the predictor and the outcome, net of X_{it} . We exclude staff demographic characteristics from the lean specification due to concerns about overcontrolling. Teacher and leader composition is plausibly endogenous to leader effectiveness (Boyd et al., 2011), so conditioning on these characteristics risks absorbing the pathways through which principal quality operates and attenuating coefficients toward zero. Conditioning on post-treatment variables may also introduce spurious associations. To assess sensitivity to this decision, we estimate the Full Controls specification, replacing Lean with Full Controls.

Because the three score-generators are applied to the same documents for the same principals, their errors share unmeasured common causes; testing coefficient equality across methods therefore requires a framework that models that correlation explicitly. We estimate Equation 1 separately for each predictor-outcome combination, then stack the three predictor-specific models for each outcome and estimate them jointly via stacked seemingly unrelated regressions to permit Wald tests of coefficient equality across scoring methods. Because Human and GRM scores are highly collinear, the joint model version of Equation 1 enters only GRM and LLM predictors simultaneously; the joint F-test indicates if the two approaches collectively explain incremental variation conditional on one another.

Our estimation approaches bear on the samples we report. We fit the Year FE + District FE models with Stata's *reghdfe*, which drops singleton observations (i.e., districts contributing a single principal-year). Stacked estimation via Stata's *suest* does not accept *reghdfe* output, so we

refit each model as OLS with district and year indicators on a common sample that retains the singletons. We report the *reghdfe* estimation sample in the tables and the slightly larger *suest* sample in the table notes; the two differ by 9 to 17 observations, depending on the outcome.

Predictive Validity

Predictive validity analyses replicate the concurrent framework using future-year ($t+1$) outcomes as the dependent variable, estimating:

$$Y_{is,t+1} = \beta_1 S_{it} + X_{it}\gamma + \alpha_d + \delta_t + \varepsilon_{is,t+1} \quad (2)$$

where the notation follows Equation 1, although the outcome is measured in the subsequent year. This specification tests whether current-year evaluation scores contain information about future school performance, a stronger validity criterion than concurrent association, because it rules out the possibility that scores merely reflect contemporaneous conditions observable to both evaluators and outcome measures. Predictive analytic samples range from 712 to 908 leader-year observations, depending on the outcome.

Heterogeneity Analyses

We conduct heterogeneity analyses on two moderator sets. The main-text analyses use two tercile moderators: school poverty (proportion FRPL) and racial composition (proportion nonwhite students). The appendix analyses use two binary moderators: school level (high school versus non-high school) and rurality (rural versus non-rural). These tests ask whether scoring methods differ across school contexts in ways that bear on the fairness concerns developed in the Literature Review's equity-and-fairness treatment.

For the tercile-moderator analyses, we estimate:

$$Y_{ist} = \beta_1 S_{it} + \beta_2(S_{it} \times T2) + \beta_3(S_{it} \times T3) + T2\lambda + T3\mu + X_{it}\gamma + \alpha_d + \delta_t + \varepsilon_{ist} \quad (3)$$

where T2 and T3 are indicators for the middle and highest terciles of the moderating variable (with the lowest tercile as the reference category). The interaction coefficients β_2 and β_3 capture how the predictor-outcome association changes across school contexts. For the binary-moderator analyses, we estimate a parallel interaction model in which each binary moderator (school level; rurality) is entered as a single indicator.

Structural Bias

GBM (2018) documents that supervisor ratings of principals correlate negatively with school poverty in ways that go beyond what alternative performance measures explain. Student racial composition, by contrast, did not consistently predict ratings in their models. We ask whether the three scoring methods differ in their demographic encoding and, if so, in which direction. For each scoring approach $S \in \{\text{Human, GRM, LLM}\}$, we estimate:

$$S_{it} = Z_{it}\pi + \delta_t + \varepsilon_{it} \quad (4)$$

where S_{it} is the standardized evaluation score; Z_{it} is the full 13-covariate set; δ_t represents year fixed effects; and ε_{it} is the error term. Lagged achievement is not part of the structural-bias control set. We omit District FE to preserve between-district variation in school demographics, which is central to the structural-bias question. We use Wald tests from stacked seemingly unrelated regressions to compare coefficients across scoring methods, with a joint omnibus test as the overall summary of demographic sensitivity.

Text Feature Analysis

If LLM scores diverge from Human and GRM scores in their concurrent or predictive associations, what are LLM scores reading in the text that the human-anchored measures are not, and how much of that reading is surface writing properties? We investigate this in two complementary analyses, both using the three text factors extracted described in the Data section.

First, we add the three text factors as right-hand-side controls to the validity and structural-bias models: for concurrent validity, we re-estimate Equation 1 with the factors added only in the LLM model; for structural bias, we re-estimate Equation 4 with the factors as additional covariates. Under the artifactual-LLM hypothesis, LLM coefficients should attenuate once we absorb surface features. Second, we construct a residualized LLM score by regressing the LLM composite on the three text factors and retaining the residual, isolating the component of LLM variance orthogonal to surface features. We re-estimate concurrent validity using only the residualized LLM score. The residualized approach mirrors the right-hand-side approach: right-hand-side analyses ask whether surface-feature absorption attenuates the LLM signal; residualization analyses ask whether the surface-feature-orthogonal portion retains validity.⁷

Robustness and Sensitivity Analyses

We organize robustness and sensitivity checks into three groups: specification, sampling, and measurement. We report sample-level diagnostics in Online Appendix A alongside the data concerns they address; specification and measurement robustness tables appear in Online Appendix C. For specification robustness, we re-estimate all concurrent and predictive models under the Year FE + Lean Controls or District FE + Year FE + Full Controls specification. Appendices A and C report sampling and measurement robustness test results. Sampling tests examine attrition, apply inverse probability weighting to address differential survey availability, restrict to a common five-outcome sample, restrict to single-leader school-year observations, and report distributional diagnostics. Measurement tests apply Empirical Bayes shrinkage to school-value-added, probe the 2019-20 COVID skip via complementary exclusion and interaction

⁷ This residualization analysis, the first-stage R^2 from regressing the LLM index on the three text factors, and outcome models re-estimated using the residualized score, is reported in Table B2 of Online Appendix B.

approaches, report the seven-element complete-case predictor variant, and decompose the staffing stability index into teacher- and administrator-retention components. In the main text, because MHTs are conducted within each regression table, we apply Benjamini-Hochberg false-discovery-rate correction at $q = .10$ (Benjamini & Hochberg, 1995); Online Appendix B defines the test families and reports the adjusted p-values.⁸

Findings

Table 1 reports the Human supervisor rating on its raw observed scale and the GRM theta and LLM index on the standardized metric; when standardizing within year, the Human score has comparable variance to the other two ($M = 0.03$, $SD = 0.95$). The GRM theta has a somewhat narrower range on the standardized metric ($M = 0.01$, $SD = 0.83$; min -2.63 , max 0.91). The LLM index has the widest range and largest standard deviation in the sample ($M = 0.01$, $SD = 0.95$; min -3.84 , max 1.76). The bivariate correlations among the predictor scores reveal the empirical patterns motivating the remainder of this section (see Table A12). Human and GRM scores are near-interchangeable ($r = .92$). LLM scores, however, correlate only weakly with Human ($r = .12$) and with GRM ($r = .07$).

Concurrent Validity

Table 2 shows that Human and GRM scores converge in their associations with same-year school outcomes, whereas the LLM index diverges substantively. In separate-predictor models (Panel A), GRM predicts achievement (.038), and both Human and GRM predict teacher-survey ratings of principal effectiveness (.039 and .045, respectively). GRM is marginally significant for the student survey (.010). The LLM index predicts staffing stability

⁸ See main text Table 2, 3, and 4 notes, and Appendix B for details.

(.060), is marginal for the teacher survey (.032), and is null for achievement and the student survey. School value-added associations are null for all three predictors.

Cross-method Wald tests formalize the divergence. Human and GRM differ on achievement ($p < .05$) and school value-added ($p < .01$), with GRM the larger estimate in each case. LLM (L) differs sharply from both Human (H) and GRM (G) on staffing stability ($H=L, p < .01$; $G=L, p < .01$) and from GRM on the student survey ($p < .05$). The pattern is asymmetric. Human and GRM are statistically interchangeable on stability, teacher survey, and student survey; LLM is statistically distinguishable from both human-based methods on stability and from GRM on student survey. Furthermore, the LLM stability association operates through teacher retention ($\beta = .088, p < .001$; Online Appendix Table C6). Of the six nominally significant cross-method Wald tests in Table 2, five survive MHT correction; the lone exception is $H = L$ on the student survey.

The joint model (Panel B) remains predictive even after the two scores compete for shared variance, but the divergence of relationships from Panel A persists: the joint F is significant for achievement ($p < .05$), stability ($p < .01$), and teacher survey ($p < .01$), and marginal for student survey ($p < .10$), while VAM remains null. In every case but one, only a single predictor carries the significant coefficient: GRM alone for achievement, teacher survey, and student survey; LLM alone for stability.

Predictive Validity

The predictive validity results in Table 3 show that concurrently distinguishable predictors become statistically equivalent for next-year outcomes. With the dependent variable shifted to $t+1$, Panel A reports a near-uniform predictive null; no score significantly predicts future achievement, future value-added, future stability, or future student survey. The sole

survivor is the future teacher survey, predicted by Human (.024) and GRM (.025). The LLM stability signal that was significant concurrently is no longer significant predictively, although the point estimate (.058) is of similar magnitude, and the standard error is roughly 67% larger in the smaller sample.

Cross-method Wald tests are weaker than their concurrent counterparts, although two teacher-survey tests become marginally significant. Five of fifteen Panel A tests reach marginal significance, predominantly comparing LLM to Human or GRM on stability and teacher survey, although none survives MHT correction. The joint model (Panel B) shows limited incremental predictive validity. Two design features bear on interpretation. First, predictive samples are 12.8% to 48.2% smaller than concurrent samples, increasing standard errors for most outcomes; notably, the teacher survey has the smallest reduction and is the one outcome retaining a significant predictive association, consistent with a power-based account. Notably, Online Table A7's predictive-outcome columns document that no outcome-specific normalized difference exceeds the Imbens and Rubin (2015) threshold of $|0.25|$. Second, restricting to single-leader schools sharpens several predictive coefficients in the direction that within-school measurement dilution would predict: the Human achievement association moves from .003 to .055 and the LLM stability association from .058 to .068 (Online Appendix Table A5). Because the restriction removes a third of the sample and changes its composition, we read this as consistent with measurement-error dilution.

Heterogeneity

The binary moderators (school level and rurality) produce no heterogeneity surviving MHT correction, so we do not discuss them here (see Online Table B1 for details). In Table 4 Panel A, the FRPL-tercile interactions are nominally significant for five predictor-outcome cells,

including i) the Human $\times$ poverty interaction for stability (-.076) and ii) the teacher survey (.052); iii) GRM $\times$ poverty for the teacher survey (.065); iv) marginally for LLM $\times$ poverty and school value-added (.101), and v) significantly for the teacher survey (.082). Substantively, the high-poverty (T3) cells drive this heterogeneity. The Human and GRM teacher-survey associations are larger in high-poverty schools than in low-poverty schools, so in high-poverty contexts, human-based scores are more informative about how teachers experience leadership effectiveness. Simultaneously, the Human stability association becomes negative in the high-poverty tercile (-.076) relative to a positive, insignificant slope in the low-poverty tercile. Yet, MHT correction removes all statistical significance. We treat the FRPL patterns as suggestive.

Some Panel B race-tercile results survive MHT correction. Substantively, in the highest-nonwhite tercile, the Human and GRM associations with staffing stability are strongly negative (-.120 and -.167, respectively). The LLM stability slope is not significant, though the associations are positive across all three terciles. Notably, the LLM avoids the inversion on staffing stability. The LLM is nominally associated with achievement and student surveys in the middle terciles, but these do not survive MHT correction. In the schools serving the highest shares of students of color, leaders scored as more effective by humans (and GRM) lead less stable staffing; LLM scores do not exhibit this inversion.

Structural Bias

School Context. Table 5 Panel A reports structural-bias regressions; odd-numbered columns omit text-feature controls and even-numbered columns add the three text factors. The proportion of FRPL students produces the clearest divergence: Human and GRM scores are strongly negatively associated with school poverty (columns 1-4) while the LLM association is positive (columns 5-6), and baseline cross-method Wald tests show separation between the LLM

and both human-based scores, with the latter associations not differing from one another. The marginally significant baseline LLM coefficient with FRPL does not survive MHT correction (.648), though the text-feature-adjusted one does marginally (.259). School size divides the methods the same way, with Human and GRM scores rising with the number of teachers (both .006) while the LLM association is precise and near zero (about -.001). The remaining three context variables show no divergence, and the proportion of nonwhite students is a precise null in every model, with no race test surviving correction.

Staff Composition. Relationships with the eight staff-composition characteristics are more complex, and most depend on whether text features are absorbed. Average teacher experience is the only LLM-specific association in columns 5 and 6: the LLM is negatively associated with it (−.047 baseline, −.022 text-adjusted), while both human-based scores are null, and the cross-method tests separate human-based vs. LLM throughout. The proportion of male administrators runs the other way. Human and GRM are negatively associated with it in both columns (−.26 to −.30), while the LLM is not; but the cross-method tests reach significance only once text features enter and the LLM coefficient moves to zero. Two results appear only in column 5. Administrators holding advanced degrees is an LLM-specific association at baseline (.200) that disappears under text adjustment, and the proportion of non-white teachers, a precise null for all three methods at baseline, becomes a large negative LLM association after text adjustment (−.770) while the human-based scores remain null. The remaining four characteristics show no significant association for any method. After MHT correction, two relationships survive — the percentage of male leaders for both human-based scores, and the LLM’s text-adjusted association with non-white teachers — as do the Wald divergences on teacher experience and on

male leaders, and the Human-versus-GRM divergence on leader experience, a covariate on which neither method has a detectable slope.⁹

Variance Explained. Table 5 reports R^2 : Human R^2 rises by .22 and GRM by .10 once text factors are added, but LLM rises by .644, with all three text-factor LLM coefficients large and significant. After absorbing text-feature variance, the LLM FRPL coefficient attenuates by roughly 60% but remains significant. Human and GRM scores are not orthogonal to surface-text features (joint F tests $p < .01$ and $p < .05$) but at magnitudes of order smaller than LLM.

Within-District Estimates. Because the structural-bias models omit District FE by design, the estimates above use between- and within-district variation. Online Appendix C re-estimates the Full Controls model with District FE, as in the GBM (2018) approach. The human-based poverty coefficients become insignificant while the LLM relationship is largely unchanged: supervisors in higher-poverty districts assign lower scores than supervisors in lower-poverty districts, and within a district the gradient disappears. This diverges from GBM, in which the poverty association persists within districts and indeed within leaders. Two features of the school-size result survive the restriction: the Human and GRM associations remain marginally positive, and the LLM association becomes negative and significant, strengthening the cross-method divergences. The administrator-gender association likewise holds within district.¹⁰

⁹ Coefficients and Wald tests are corrected as separate families throughout, following the convention applied to Tables 2 through 4.

¹⁰ We do not interpret the within-district estimates for rural status or school level, which are identified from a small number of schools whose classification changes, and we treat the District FE models throughout as a decomposition of the main-text estimates rather than as a competing specification.

Text Feature Analysis

Table 6 asks whether the LLM's concurrent validity profile from Table 2 survives surface-text adjustment. We re-estimate Equation 1, including the three text factors as additional covariates, only in the LLM model. The three factors are jointly a precise null as predictors of school outcomes conditional on the LLM index. The LLM stability signal increases under text-feature controls, with the standard error inflated by roughly 92%, and the cross-method divergence on stability is preserved. The cross-method test on the teacher survey, null in Table 2, reaches conventional significance under text-feature controls ($G=L + TF, p < .05$).

Were LLM scores driven by surface-text features, their concurrent validity should collapse when those features are absorbed. On stability, it does not, and the divergence from Human and GRM is preserved. On the teacher survey, the LLM's coefficient collapses, and the divergence widens, a mechanical consequence of comparing a shrinking LLM coefficient to a GRM coefficient held fixed at its Table 2 value. Online Appendix B suggestively corroborates this result: residualization removes roughly half the standard deviation of the original LLM, and the stability estimate retains its sign and magnitude (.060 $\rightarrow$.069) without reaching significance ($p = .16$). The residualized teacher-survey coefficient is precisely null (−.004), consistent with that signal being driven by surface text.¹¹ LLM scores are substantially driven by surface-text features, as the structural-bias and text-feature analyses show, yet the LLM signal that distinguishes high-stability from low-stability leader BIP-POs is not a surface-text artifact.

¹¹ The school-value-added association strengthens on the residualized LLM score to a marginal $\beta = .110$ ($p < .10$), where the original LLM coefficient was null, a suppression-style pattern that we treat as suggestive rather than confirmed because it does not survive MHT correction.

Robustness and Sensitivity

Across the Year FE + Lean Controls and District FE + Year FE + Full Controls models, the results largely replicate. When they do not replicate, it is in the Year FE + Lean Controls, indicating that between-district confounding is the source, which our main specification is built to absorb (Online Appendix C, Tables C1 and C2). The remaining battery corroborates our key inferences. Inverse probability weighting reproduces the concurrent and predictive patterns, although several heterogeneity slopes lose conventional significance under wider weighted standard errors (Online Appendix A, Tables A2 through A4). Concurrent results are largely unchanged or sharper when the sample is restricted to single-principal school-years (Table A5), under the seven-element predictor variant (Table A8), and under the common five-outcome restriction, though this restriction loses some statistical significance (Table C3). Neither the exclusion nor the interaction approach to the COVID-era skip logic changes our conclusions (Table A6). Empirical Bayes shrinkage leaves the value-added null intact (Table C5), and Table C4 rules out range restriction as an explanation for the differential slopes in Table 4. One result qualifies rather than corroborates. The single-principal restriction sharpens several predictive validity coefficients in the direction within-school dilution would predict, with Human achievement moving from .003 to .055 and LLM stability from .058 to .068, both reaching significance; the predictive null may partly reflect school-years in which two evaluated principals share the same outcomes.

Discussion and Conclusion

Cross-Method Validity Divergence Between Human-Based and LLM Scores

The central construct-validity finding is that three scoring methods applied to identical documents produce markedly different portraits of the relationships between school outcomes

and BIP-PO scores, which aim to score school-improvement leadership. The human-based ratings converge strongly. The LLM often diverges sharply from both human-based methods. LLM scores uniquely and consistently track staffing stability, with the association concentrated in the teacher-retention component.

The value-added null for all three scoring methods, and the absence of any Human score association with achievement, are consistent with prior work on the limits of achievement test-based leader metrics (Bartanen et al., 2024). The achievement result is not uniformly null, however: GRM scores predict achievement across all specifications we estimate, and Human and GRM differ significantly on that outcome. Thus, GRM rescaling may recover an instructional leadership signal as captured by achievement scores not detected in the raw human ratings.

Concurrent and predictive validity differ in these data. The association between BIP-PO scores and same-year outcomes is substantive and method-specific. Predictive validity, however, collapses toward near-equivalence across methods, with most associations attenuating and losing statistical significance; the exception is the future teacher survey, where both human-based scores remain significant in the main specification. NEE districts are typically small. It is plausible that supervisors and teachers share some persistent opinions of school-based leadership (i.e., assessments of school-improvement leadership or perceptions of general leadership effectiveness). The concurrent-predictive asymmetry suggests that the concurrent discriminations each method captures may reflect different facets of within-year leader-school alignment rather than forward-looking leadership capacity, aside from some persistent, shared supervisor-teacher opinions. Ultimately, concurrent measurement offers the richest information about the dimensional structure of BIP-PO quality.

We offer one speculative interpretation. The divergence between human and LLM scoring may reflect differences in attention. Human supervisors, who read BIP-POs within established institutional relationships and professional norms, may weight compliance language, procedural completeness, and presentational fluency more heavily. LLMs, with no such contextual priors, may attend more to substantive improvement content, goal specificity, and the internal coherence of logic models. This interpretation is consistent with the evidence, although it cannot be confirmed without direct analysis of scoring processes.

The Text-Feature Paradox

The three surface-level text factors represent about 72% of the variance in LLM scores, appearing to confirm the concern that machine-generated scores primarily represent surface writing properties. The evidence, however, does not support this interpretation. When surface text features are absorbed as controls, LLM validity for staffing stability holds or strengthens. The residualized LLM score, purged of the variance attributable to the text factors, retains the sign and approximate magnitude of the stability association, although at that reduced scale it no longer reaches statistical significance; we treat this as corroborating rather than confirming. This result reframes the meaning of text-adjusted associations. Rather than compromising validity, surface-text properties in BIP-PO documents may carry substantive information about leadership. Leaders who produce more detailed, linguistically sophisticated, and rubric-aligned documents may also lead more organized schools, communicate more effectively with staff, and allocate more deliberate attention to talent management. Sun et al. (2019) reached a compatible conclusion in the plan-quality context: the textual properties of plan content were not noise, but a signal to be decoded.

Incremental Validity and Multidimensionality

The joint GRM + LLM models provide the most direct evidence for multidimensionality in BIP-PO measurement. When modeling GRM and LLM simultaneously, each contributes incremental validity conditional on the other. GRM retains its association with teacher-survey outcomes after controlling for LLM, and LLM retains its association with staffing stability after controlling for GRM. Thus, each seemingly captures a dimension of BIP-PO quality that the other misses. This finding constitutes direct empirical evidence for the proposition that leadership effectiveness is inherently multidimensional (Grissom et al., 2021).

The incremental validity finding also points toward the complementarity model advanced by Holstein and Aleven (2022), who argued that human and algorithmic judgment should be combined rather than pitted against each other. The goal may not be to replace human scoring with algorithms, or vice versa, but to exploit the dimensional specificity of each method.

Heterogeneity and Equity

The finding with the greatest consequential-validity significance concerns heterogeneity across school demographic contexts. We detect no association between human-based scores and staffing stability in low or moderate poverty- or racial-minority-enrollment schools but find significantly negative associations in schools with high poverty and minority enrollment, whereas the LLM association with stability remains positive and setting-invariant. The pattern is not explained by attenuation or noise. If human supervisor ratings of BIP-POs are used for high-stakes decisions, they may systematically disadvantage leaders in the schools that most need stable, effective leadership. The LLM, scoring documents without direct access to information about school demographics, enrollment composition, or neighborhood characteristics, produces associations with staffing stability that do not reverse across terciles. Its context stability suggests

that scoring based on document content alone is less susceptible to school-context confounding than scoring embedded in supervisory relationships.

Our evidence is consistent with the possibility that the context dependence GBM (2018) documented in human supervisor ratings reflects the rating method rather than an inherent feature of leadership document evaluation. GBM attributed the context dependence they observed to the evaluation process broadly, without the cross-method design needed to isolate method-specific contributions. Our finding that LLM validity patterns do not exhibit the same context dependence as human-based scores suggests that the bias gradient may originate in the human-based rating process. If confirmed in other settings, this would imply that cross-method designs are needed to separate method-specific bias from construct-level associations.

We temper this interpretation with an important caveat. A substantial and growing literature documents that LLMs can reproduce and amplify societal biases present in their training data (Bender et al., 2021; Warr et al., 2025). LLM context stability in our data, when scoring this document type, does not support the general claim that LLM scoring is free from demographic bias. Our own models supply a local instance: after surface-text adjustment, the LLM is the only method associated with the proportion of non-white teachers, though the cross-method tests do not separate it from the human-based scores. The BIP-PO documents in our sample do not include explicit demographic identifiers, and we prompted the LLM with a rubric-referenced scoring protocol that directs attention to plan content rather than author characteristics. Whether LLM scoring maintains context stability across different document types, prompting strategies, or language models remains an open question.

Limitations

Seven limitations bound the interpretation of our findings. First, the study draws on a single state's leader evaluation system; we do not assume generalizability to other systems. Second, our design relies on a single rater for each scoring method, precluding decomposition of construct-relevant variance from rater-idiosyncratic variance. A multi-rater design would allow estimation of rater-specific components and provide a cleaner test of method-level divergence. Third, teacher survey coverage in our analytic sample reaches 73%, with a modest overrepresentation of high-poverty schools; unobserved factors correlated with both survey nonresponse and school outcomes may introduce residual bias. The survey-based findings should be interpreted with this coverage limitation in mind. Fourth, LLM reproducibility is a concern for any study relying on proprietary LLM infrastructure. We generated our scores using a specific version of GPT-4o-mini, and subsequent model updates, changes to API behavior, or differences across model providers could yield different scores for the same documents. Fifth, all validity evidence in this study is correlational. Sixth, the text-feature analysis relies on three broad dimensions, Volume, Sophistication, and Rubric Overlap, constructed from twelve underlying features. These dimensions are intentionally coarse, designed to capture the major axes of surface text variation. More granular text analysis, incorporating measures of argumentative structure, causal reasoning, or specificity of action steps, could more precisely explain how text properties relate to LLM scoring and outcome validity. Seventh, BIP-PO documents are retrospective, leader-authored accounts of school improvement planning and implementation. Scores derived from these documents reflect the quality of the documented narrative rather than direct observation of leadership practice. This scope constraint bounds the construct under study; we are validating scores on a written account of leadership, not leadership behavior itself.

Implications

Policymakers. The choice of scoring method is itself an equity-consequential decision. Human-based ratings generate validity evidence that varies systematically across certain school demographic contexts, with sign reversals in the highest-racial-minority and -poverty terciles; scores positively associated with desirable outcomes in some schools and negatively associated in others do not simply lack information; they may provide misleading information. Nor does any single method suffice. GRM and LLM scores seemingly capture different dimensions of BIP-PO quality, each exhibiting validity for distinct outcome domains, so policymakers designing or revising school leader evaluation frameworks might consider multi-method architectures. The LLM cannot be dismissed as capturing “just surface text,” though our evidence does not support uncritical adoption. And if the context dependence GBM (2018) documented is a property of the human rating method rather than of the leadership construct, addressing it may require changing the measurement method, not merely retraining raters or recalibrating scales.

Practice. For district leaders responsible for school leader evaluation, our findings suggest a practical path forward through multi-method feedback. Districts could provide school leaders with both supervisor judgment and a complementary LLM-based signal as complementary lenses on the same documented practice: a supervisor rating might highlight instructional and contextual dimensions of leadership quality, while an LLM score surfaces content dimensions seemingly captures an aspect of organizational health—staffing stability.

Research. The design we employ, comparing multiple scoring methods applied to identical artifacts across multiple outcome domains, may apply to any field in which constructed-response assessments are scored by both human and automated methods. Our results also caution

against validating a single scoring method against a narrow set of outcomes. All three methods show null associations with value-added and two of the three with achievement, a pattern that would lead to blanket rejection of BIP-PO scores were those the only outcomes examined; expanding the outcome space to staffing stability and teacher-survey domains reveals method-specific validity that a narrow validation design would miss entirely. Two extensions follow. Whether surface text carries substantive information in some document genres but not others would inform both AES design and validity argumentation, for which Sun et al.'s (2019) topic-prevalence measures in school improvement plans provide one natural extension. Our results also speak to the rescaling literature: Kraft et al. (2020) demonstrated the utility of GRM-based rescaling for the reliability and distributional properties of human ratings, and we find that rescaling may capture aspects of leadership that raw scores do not and preserve not only the judgmental signal but the structural-bias gradient. Yet, improved reliability does not attenuate school-context confounding when that confounding originates upstream in the rater rather than in the measurement model.

Cautionary Close

The stakes of this work extend beyond validation studies. Measurement choices effectively define what counts as effective school-improvement leadership. They shape which leaders are identified for recognition, development, or removal. They determine which school communities receive accurate performance information and which receive information distorted by the method used to generate it. In high-poverty and high-minority-enrollment schools, where leadership stability matters most and where our evidence shows the greatest measurement disparity, getting the measurement architecture right is an equity imperative.

The path forward may be to build systems that exploit the complementarity Holstein and Aleven (2022) described. The evidence presented here suggests that such systems are not only theoretically appealing but may be empirically necessary, and that the cost of ignoring cross-method divergence is borne disproportionately by the schools that can least afford it.

References

- Bartanen, B., & Husain, A. N. (2022). Connected networks in principal value-added models. *Economics of Education Review*, 90, 102292.
<https://doi.org/10.1016/j.econedurev.2022.102292>
- Bartanen, B., Husain, A. N., & Liebowitz, D. D. (2024). Rethinking principal effects on student outcomes. *Journal of Public Economics*, 234, 105115.
<https://doi.org/10.1016/j.jpubeco.2024.105115>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.
- Béteille, T., Kalogrides, D., & Loeb, S. (2012). Stepping stones: Principal career paths and school outcomes. *Social Science Research*, 41(4), 904–919.
- Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, 48(2), 303–333.
- Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2013). School leaders matter: Measuring the impact of effective principals. *Education Next*, 13(1), 62–69.
- Cai, H., & Yan, X. (2024). Triangulating natural language processing (NLP)-based analysis of rater comments and many-facet Rasch measurement (MFRM): An innovative approach to

- investigating raters' application of rating scales in writing assessment. *Language Testing*, 41(2), 384–411. <https://doi.org/10.1177/02655322231210231>
- Coelli, M., & Green, D. A. (2012). Leadership effects: School principals and student outcomes. *Economics of Education Review*, 31(1), 92–109.
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135.
- Day, C., Gu, Q., & Sammons, P. (2016). The impact of leadership on student outcomes: How successful school leaders use transformational and instructional strategies to make a difference. *Educational Administration Quarterly*, 52(2), 221–258.
- Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24.
<https://doi.org/10.1016/j.asw.2012.10.002>
- DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, 102(3), 421–433.
- Dhuey, E., & Smith, J. (2014). How important are school principals in the production of student achievement? *Canadian Journal of Economics*, 47(2), 634–663.
- Donaldson, M., Mavrogordato, M., Dougherty, S. M., Al Ghanem, R., & Youngs, P. (2021). Principal evaluation under the Elementary and Secondary Every Student Succeeds Act: A comprehensive policy review. *Education Finance and Policy*, 16(2), 347–361.
https://doi.org/10.1162/edfp_a_00332
- Edthena. (2025, October). *Observation Copilot: AI-assisted classroom observation feedback* [Product announcement].

- Edwards, K. D., & Soland, J. (2024). How scoring approaches impact estimates of growth in the presence of survey item ceiling effects. *Applied Psychological Measurement*, 48(3), 147–164. <https://doi.org/10.1177/01466216241238749>
- Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy*, 1(1), 19–27.
- Grissom, J. A., Blissett, R. S. L., & Mitani, H. (2018). Evaluating school principals: Supervisor ratings of principal practice and principal job performance. *Educational Evaluation and Policy Analysis*, 40(3), 446–472.
- Grissom, J. A., Egalite, A. J., & Lindsay, C. A. (2021). *How principals affect students and schools: A systematic synthesis of two decades of research*. Wallace Foundation.
- Holstein, K., & Aleven, V. (2022). Designing for human–AI complementarity in K-12 education. *AI Magazine*, 43(2), 239–248.
- Holstein, K., De-Arteaga, M., Tumati, L., & Cheng, Y. (2023). Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), Article 152. <https://doi.org/10.1145/3579628>
- Hunter, S. B., & Bowser, K. M. (2026). Reformed teacher evaluation in rural Missouri: Main and moderated relationships with student achievement and relationships-to-expenditure ratios. *Education Policy Analysis Archives*, 34. <https://doi.org/10.14507/epaa.34.9248>
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

- Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment, 24*(2), 91–118. <https://doi.org/10.1080/10627197.2018.1564272>
- Jones, E., Tsai, C.-L., & Bergin, C. (2024). Teachers can differentiate between standards-based leadership skills in principal evaluation surveys: A multilevel analysis. *School Effectiveness and School Improvement, 35*(4), 387–412. <https://doi.org/10.1080/09243453.2024.2369061>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.
- Kaufman, J. H., Woo, A., Eagan, J., Lee, S., & Kassan, E. B. (2025). *Uneven adoption of artificial intelligence tools among U.S. teachers and principals in the 2023–2024 school year* (RR-A134-25). RAND Corporation. <https://doi.org/10.7249/RRA134-25>
- Kraft, M. A., Marinell, W. H., & Yee, D. S.-W. (2016). School organizational contexts, teacher turnover, and student achievement. *American Educational Research Journal, 53*(5), 1411–1449. <https://doi.org/10.3102/0002831216667478>
- Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management, 39*(2), 315–347. <https://doi.org/10.1002/pam.22193>
- Kwako, A., & Ormerod, C. (2024). Can language models guess your identity? Analyzing demographic biases in AI essay scoring. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 78–86). Association for Computational Linguistics.

- Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, 55(5), 1982–2002. <https://doi.org/10.1111/bjet.13505>
- Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School Leadership & Management*, 28(1), 27–42. <https://doi.org/10.1080/13632430701800060>
- Liu, A., & Sun, M. (2025). From voices to validity: Leveraging large language models (LLMs) for textual analysis of policy stakeholder interviews. *AERA Open*, 11, Article 23328584251374595. <https://doi.org/10.1177/23328584251374595>
- Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–10).
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
- Meyers, C. V., & VanGronigen, B. A. (2019). A lack of authentic school improvement plan development: Evidence of principal satisficing behavior. *Journal of Educational Administration*, 57(3), 261–278. <https://doi.org/10.1108/JEA-09-2018-0154>
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5), 238–243.

- Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21, 104–111. <https://doi.org/10.1016/j.asw.2014.05.001>
- Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, 4(4), 492–519.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 1–97.
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1163>
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.
- Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes: A deep exploration of school improvement strategies. *Educational Evaluation and Policy Analysis*, 41(4), 510–536. <https://doi.org/10.3102/0162373719869318>
- Sun, M., Penuel, W. R., Frank, K. A., Gallagher, H. A., & Youngs, P. (2013). Shaping professional development to promote the diffusion of instructional expertise among teachers. *Educational Evaluation and Policy Analysis*, 35(3), 344–369. <https://doi.org/10.3102/0162373713482763>
- VanGronigen, B. A., & Meyers, C. V. (2020). Short-cycle school improvement planning as a lever to launch school turnaround: A descriptive analysis of plans. *Teachers College Record*, 122(5), 1–40. <https://doi.org/10.1177/016146812012200501>

- Warr, M., Oster, N. J., & Isaac, R. (2025). Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education*, 57(6), 1324–1349. <https://doi.org/10.1080/15391523.2024.2395295>
- Wind, S. A., & Jones, E. (2019). Not just generalizability: A case for multifaceted latent trait models in teacher observation systems. *Educational Researcher*, 48(8), 521–533. <https://doi.org/10.3102/0013189X19874084>

Table 1
Descriptive Statistics for Regression-Eligible Sample

Variable	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max
Panel A: Evaluation Score Predictors					
Human Supervisor Rating	1,385	3.660	0.498	1.000	4.000
GRM Theta Score	1,385	0.007	0.828	−2.627	0.907
LLM Index	1,385	0.006	0.949	−3.838	1.758
Panel B: Concurrent Outcomes (Year t)					
Achievement Index	1,385	0.017	0.926	−3.497	1.919
VAM Index	1,348	0.012	0.845	−4.609	4.612
Staffing Stability Index	1,385	−0.028	0.797	−4.091	0.541
Teacher Survey	1,014	2.401	0.271	0.691	3.000
Student Survey	978	2.141	0.163	1.329	2.995
Panel C: Predictive Outcomes (Year t+1)					
Future Achievement Index	935	0.064	0.938	−3.497	1.859
Future VAM Index	746	0.045	0.771	−3.390	3.367
Future Stability Index	723	−0.032	0.842	−4.777	0.832
Future Teacher Survey	878	2.430	0.256	0.902	3.000
Future Student Survey	727	2.157	0.164	1.329	2.995
Panel D: Control Variables					
Proportion Students FRPL	1,385	0.503	0.282	0.000	1.000
Proportion Students Nonwhite	1,385	0.257	0.293	0.000	1.000
Number of Teachers	1,385	42.074	22.624	5.000	161.000
High School	1,385	0.244	0.430	0.000	1.000
Rural	1,385	0.548	0.498	0.000	1.000
Avg. Teacher Experience (years)	1,385	12.872	2.554	2.941	22.800
Prop. Teachers Advanced Degree	1,385	0.567	0.162	0.038	0.962
Prop. Teachers Nonwhite	1,385	0.040	0.077	0.000	0.620
Prop. Teachers Male	1,385	0.214	0.145	0.000	0.658
Avg. Admin Experience (years)	1,385	17.525	4.874	2.000	37.000
Prop. Admin Nonwhite	1,385	0.059	0.177	0.000	1.000
Prop. Admin Male	1,385	0.434	0.389	0.000	1.000
Prop. Admin Advanced Degree	1,385	0.642	0.376	0.000	1.000
Lagged Achievement Index	1,364	0.010	0.937	−3.801	2.033

Note. *N* = 1,385 principal-year observations from 103 districts, 2016–2022 (excluding 2020). The regression-eligible sample requires non-missing values on all three predictors and 13 control variables. Outcome-specific *N*s reflect additional missingness on individual outcomes. Human Rating is the raw mean across 7 BIP-PO rubric elements (0–4 rubric; 1–4 observed); GRM Theta and LLM Index are standardized within year. Achievement, VAM, and Lagged Achievement indices are standardized composites of math and reading components; Stability is a standardized composite of staff retention. Teacher Surveys are principal-level means, while Student Survey scores are school-level means, both on original scales.

Table 2*Concurrent Validity: Associations Between Principal Evaluation Scores and Same-Year School Outcomes*

<i>Panel A: District + Year FE, Lean Controls</i>					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
Human Supervisor (H)	.020 (.013)	−.011 (.042)	−.029 (.024)	.039** (.012)	.006 (.004)
GRM Theta (G)	.038* (.015)	.038 (.048)	−.037 (.036)	.045** (.015)	.010† (.006)
LLM Index (L)	.008 (.016)	.022 (.029)	.060* (.024)	.032† (.016)	−.007 (.006)
χ^2 : H = G	5.42*	9.06**	0.29	0.86	1.53
χ^2 : H = L	0.46	0.43	8.85**	0.25	3.69†
χ^2 : G = L	2.26	0.09	7.74**	0.85	5.17*
<i>N</i>	1,354	1,335	1,374	997	965
<i>Panel B: GRM + LLM Joint Model —District + Year FE, Lean Controls</i>					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
GRM Theta (G)	.037* (.015)	.034 (.050)	−.054 (.036)	.039** (.013)	.012* (.006)
LLM Index (L)	.003 (.017)	.017 (.030)	.068** (.024)	.025 (.015)	−.009 (.006)
<i>F</i> (GRM + LLM = 0)	3.20*	0.54	5.05**	4.97**	2.59†
<i>N</i>	1,354	1,335	1,374	997	965

Note. Entries are standardized regression coefficients. Panel A reports separate predictor models; each cell is from an independent regression of the concurrent outcome on a single evaluation score, with standard errors adjusted for cross-equation correlation. Panel B reports a joint model with GRM and LLM scores entered simultaneously. District-clustered standard errors in parentheses. Wald χ^2 tests compare cross-equation coefficient equality. F-test is the joint significance test for GRM and LLM in the combined model. Lagged achievement included for Achievement outcome only. Outcome-specific maximal samples within the regression-eligible sample. District and year fixed effects and lean controls. † $p < .10$. * $p < .05$. ** $p < .01$.

Table 3*Predictive Validity: Associations Between Principal Evaluation Scores and Future (t+1) School Outcomes*

<i>Panel A: District + Year FE, Lean Controls</i>					
	Achievement (t+1)	VAM (t+1)	Stability (t+1)	Teacher Survey (t+1)	Student Survey (t+1)
Human Supervisor (H)	.003 (.021)	.026 (.045)	−.017 (.032)	.024** (.009)	.002 (.009)
GRM Theta (G)	.016 (.026)	.082 (.061)	−.012 (.037)	.025* (.012)	.003 (.011)
LLM Index (L)	−.001 (.019)	.017 (.036)	.058 (.040)	.000 (.012)	.003 (.006)
χ^2 : H = G	1.55	3.82†	0.06	0.05	0.04
χ^2 : H = L	0.02	0.04	2.75†	3.52†	0.02
χ^2 : G = L	0.35	1.35	3.66†	3.48†	0.00
N	908	733	712	869	714
<i>Panel B: GRM + LLM Joint Model —District + Year FE, Lean Controls</i>					
	Achievement (t+1)	VAM (t+1)	Stability (t+1)	Teacher Survey (t+1)	Student Survey (t+1)
GRM Theta (G)	.017 (.026)	.080 (.060)	−.028 (.035)	.027* (.012)	.003 (.010)
LLM Index (L)	−.003 (.019)	.006 (.036)	.062 (.039)	−.005 (.011)	.003 (.006)
F (GRM + LLM = 0)	0.21	0.88	1.87	2.52†	0.17
N	908	733	712	869	714

Note. Each cell is from an independent regression of the future outcome on a single evaluation score, with standard errors adjusted for cross-equation correlation. See Table 2 Notes for additional details. † $p < .10$. * $p < .05$. ** $p < .01$.

Table 4*Heterogeneous Associations by School Context: Total Predictor Slopes at Each Moderator Level*

<i>Panel A: Poverty Tercile</i>					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
Human Supervisor (H)					
Low poverty (T1)	.047 [−.014, .107]	−.002 [−.117, .113]	.038 [−.053, .129]	.009 [−.023, .041]	.007 [−.014, .027]
Medium poverty (T2)	.018 [−.017, .054]	.021 [−.075, .116]	.006 [−.047, .060]	.027 [−.014, .068]	.004 [−.009, .017]
High poverty (T3)	.005 [−.025, .036]	−.028 [−.118, .062]	−.076** [−.111, −.040]	.052** [.041, .063]	.006 [−.004, .015]
GRM Theta (G)					
Low poverty (T1)	.070* [.014, .127]	.025 [−.125, .175]	−.001 [−.127, .125]	.022 [−.010, .055]	.017 [−.009, .042]
Medium poverty (T2)	.029 [−.007, .066]	.043 [−.048, .134]	.014 [−.054, .083]	.025 [−.018, .068]	.004 [−.016, .023]
High poverty (T3)	.021 [−.014, .055]	.051 [−.084, .186]	−.092* [−.163, −.022]	.065** [.046, .084]	.010 [−.003, .024]
LLM Index (L)					
Low poverty (T1)	−.009 [−.036, .017]	−.038 [−.105, .029]	.043* [.008, .077]	.009 [−.011, .029]	−.020* [−.036, −.005]
Medium poverty (T2)	.016 [−.025, .058]	.101* [.022, .179]	.072 [−.002, .146]	−.003 [−.039, .033]	.003 [−.009, .016]
High poverty (T3)	.021 [−.050, .093]	.049 [−.043, .142]	.091* [.018, .163]	.082** [.034, .130]	−.005 [−.016, .007]
<i>N</i>	1,354	1,335	1,374	997	965
<i>Panel B: Racial Composition Tercile</i>					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
Human Supervisor (H)					
Low nonwhite (T1)	.008 [−.039, .055]	.075 [−.018, .169]	.017 [−.030, .063]	.036† [.001, .072]	.004 [−.009, .018]
Medium nonwhite (T2)	.012 [−.027, .052]	−.100† [−.192, −.008]	.037 [−.036, .110]	.033 [−.006, .071]	.012* [.002, .023]
High nonwhite (T3)	.028* [.009, .047]	−.027 [−.136, .083]	−.120** [−.162, −.078]	.043* [.015, .070]	.002 [−.007, .011]
GRM Theta (G)					
Low nonwhite (T1)	.020 [−.033, .073]	.115 [−.006, .236]	.017 [−.062, .097]	.045* [.010, .081]	.007 [−.016, .030]
Medium nonwhite (T2)	.036 [−.007, .080]	−.080 [−.178, .019]	.059 [−.027, .146]	.051† [.008, .093]	.013 [−.000, .027]
High nonwhite (T3)	.036* [.008, .064]	.059 [−.092, .211]	−.167** [−.215, −.119]	.037 [−.011, .085]	.010 [−.003, .024]
LLM Index (L)					
Low nonwhite (T1)	−.056† [−.106, −.005]	−.033 [−.136, .070]	.070 [−.026, .165]	.016 [−.027, .059]	.003 [−.018, .024]

Medium nonwhite (T2)	.025* [.007, .044]	.072 [−.006, .151]	.040 [−.010, .089]	.020 [−.018, .058]	−.031** [−.047, −.015]
High nonwhite (T3)	.041 [−.007, .089]	.016 [−.057, .089]	.078 [−.005, .162]	.056 [−.001, .113]	.005 [−.009, .019]
<i>N</i>	1,354	1,335	1,374	997	965

Note. Entries are total predictor slopes (β) at each moderator level with 90% confidence intervals. Each cell reports the total relationship between the predictor and the outcome for schools in that tercile. Lagged achievement is included for Achievement only; VAM excludes lagged achievement. District-clustered standard errors. † $p < .10$. * $p < .05$. ** $p < .01$.

Table 5*Structural Bias: School Context, Staff Composition, and Text Feature Associations With Predictor Scores*

Panel A: Full Controls

	Human (1)	Human + TF (2)	GRM (3)	GRM + TF (4)	LLM (5)	LLM + TF (6)
% Students FRPL	−.763** (.278)	−.815** (.289)	−.812** (.270)	−.841** (.273)	.648† (.324)	.259* (.107)
% Students Non-White	.169 (.455)	.181 (.440)	.249 (.450)	.256 (.435)	.028 (.307)	.009 (.101)
High School	−.214† (.115)	−.222† (.111)	−.205* (.099)	−.215* (.098)	−.185† (.111)	−.109† (.056)
Rural	−.051 (.212)	−.049 (.201)	−.010 (.210)	−.009 (.202)	−.021 (.190)	−.004 (.061)
N Teachers	.006** (.002)	.006** (.002)	.006** (.002)	.006** (.002)	−.001 (.002)	−.000 (.001)
Avg. Teacher Experience	.030 (.024)	.034 (.022)	.028 (.023)	.030 (.022)	−.047* (.023)	−.022* (.010)
% Teachers Advanced Degree	−.152 (.479)	−.207 (.470)	.075 (.437)	.046 (.425)	.268 (.433)	.015 (.169)
% Teachers Non- White	−.011 (.727)	−.239 (.734)	.099 (.588)	−.053 (.589)	−.020 (.532)	−.770** (.212)
% Teachers Male	−.104 (.332)	−.087 (.314)	−.067 (.287)	−.048 (.277)	−.034 (.330)	−.143 (.168)
Avg. Leader Experience	−.006 (.008)	−.008 (.008)	.000 (.007)	−.001 (.007)	.007 (.007)	.000 (.004)
% Leader Non-White	−.151 (.193)	−.128 (.198)	−.030 (.176)	−.009 (.179)	.081 (.222)	.120 (.135)
% Leader Male	−.297** (.095)	−.282** (.090)	−.265** (.076)	−.260** (.074)	−.122 (.101)	.007 (.050)
% Leader Advanced Degree	−.038 (.089)	−.075 (.091)	−.027 (.076)	−.051 (.079)	.200* (.091)	.030 (.042)
Factor 1 (Volume)		.087* (.036)		.044 (.029)		.495** (.018)
Factor 2 (Sophistication)		−.014 (.031)		−.018 (.025)		.364** (.016)
Factor 3 (Rubric Overlap)		.117**		.078**		.468**

		(.031)		(.029)		(.018)
R^2	.101	.123	.134	.146	.097	.741

Panel B: Cross-Equation Wald Tests (With Text Feature Models)

	χ^2 : H+TF = L+TF	χ^2 : G+TF = L+TF	χ^2 : H+TF = G+TF
% Students FRPL	13.25**	16.35**	0.09
% Students Non-White	0.14	0.31	0.73
High School	0.88	0.94	0.04
Rural	0.05	0.00	0.95
N Teachers	10.83**	10.80**	0.18
Avg. Teacher Experience	6.34*	5.79*	0.23
% Teachers Advanced Degree	0.20	0.01	3.38†
% Teachers Non-White	0.47	1.49	0.40
% Teachers Male	0.02	0.09	0.12
Avg. Administrator Experience	0.81	0.01	8.65**
% Administrators Non-White	2.01	0.98	2.99†
% Administrators Male	7.96**	9.15**	0.36
% Administrators Advanced Degree	1.19	0.88	0.69
Factor 1 (Volume)	96.53**	169.44**	—
Factor 2 (Sophistication)	114.57**	172.47**	—
Factor 3 (Rubric Overlap)	89.13**	126.23**	—
Joint: FRPL + Race	13.27**	16.35**	—
Joint: F1 + F2 + F3	$F = 5.10^{**}$	$F = 3.29^*$	$F = 876.92^{**}$

Note. $N = 1,385$ for all models. Unstandardized OLS coefficients; the dependent variable is the standardized predictor score (Human, GRM, or LLM). Standard errors (clustered by district) in parentheses. Columns 1, 3, and 5 report the Full Controls specification (year fixed effects, school demographics, and staff demographics). Columns 2, 4, and 6 add three EFA text-feature factor scores (Volume, Sophistication, Rubric Overlap; see Table A11) as additional predictors. All thirteen covariates in the specification are reported. Panel B reports cross-equation Wald χ^2 tests comparing coefficients across the with-text-feature models. Joint F-tests report within-equation tests of the joint significance of the three text factor scores ($F1 + F2 + F3 = 0$). District fixed effects are omitted to preserve the between-district variation in school demographics that is central to the structural-bias question; Table C7 reports the same models with district fixed effects added. TF = text features. † $p < .10$. * $p < .05$. ** $p < .01$.

Table 6*Concurrent Validity with Text Feature Controls: Associations Between Evaluation Scores and Same-Year School Outcomes*

<i>Panel A: Separate Predictor Models — District FE + Year FE, Lean Controls</i>					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
Human Supervisor (H)	.020 (.013)	−.011 (.042)	−.029 (.024)	.039** (.012)	.006 (.004)
GRM Theta (G)	.038* (.015)	.038 (.048)	−.037 (.036)	.045** (.015)	.010† (.006)
LLM Index + TF (L)	.029 (.025)	.108† (.057)	.077† (.046)	.002 (.022)	.007 (.014)
Factor 1 (Volume)	−.018 (.022)	−.068 (.065)	−.051 (.038)	.014 (.014)	−.007 (.013)
Factor 2 (Sophistication)	−.009 (.013)	−.038 (.032)	−.002 (.021)	.003 (.013)	−.007 (.010)
Factor 3 (Rubric Overlap)	−.013 (.014)	−.059* (.028)	.016 (.045)	.035* (.014)	−.013 (.011)
<i>F</i> (TF = 0)	0.37 (<i>p</i> = .776)	1.44 (<i>p</i> = .235)	0.60 (<i>p</i> = .618)	2.01 (<i>p</i> = .118)	0.57 (<i>p</i> = .637)
<i>N</i>	1,354	1,335	1,374	997	965
<i>Panel B: Cross-Equation Wald Tests</i>					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
χ^2 : H = G	5.42*	9.06**	0.29	0.86	1.53
χ^2 : H = L+TF	0.12	2.12	5.77*	3.81†	0.00
χ^2 : G = L+TF	0.09	0.66	7.74**	6.46*	0.09

Note. Entries are standardized regression coefficients from separate predictor models. The Human and GRM models are identical to those in Table 2 Panel A and are reprinted here to make the cross-method comparison legible within a single table. The LLM model adds three EFA text-feature factor scores (Volume, Sophistication, Rubric Overlap; see Table A11) as covariates alongside the LLM index. Standard errors (clustered by district) in parentheses. Wald χ^2 tests compare cross-equation coefficient equality. *F* (TF = 0) is the joint significance test for the three text factor scores in the LLM model. Lagged achievement is included for the Achievement outcome only. TF = text features. Table B2 reports an alternative residualization approach that purges text-surface variance from the LLM score before estimating concurrent validity. † *p* < .10. * *p* < .05. ** *p* < .01.

Online Appendix A: Data, Sample, and Measurement Foundations

This appendix provides the empirical foundation for the analyses in the main text. It documents four sample and measurement concerns arising from the study’s administrative data structure, describes how we address each concern, and reports diagnostic evidence on whether each concern threatens the main-text inferences. The concerns are: (1) district attrition from the NEE evaluation system, which produces an unbalanced panel and raises the possibility that the predictive sample differs systematically from the concurrent sample; (2) multiple leaders per building, whereby approximately 34% of leader-year observations come from school-years with two or more evaluated leaders sharing school-level outcomes, which may dilute estimates via measurement error; (3) COVID-19 era exclusions, specifically the absence of the 2019–20 school year data, which creates measurement gaps in the outcome linkage structure; and (4) selective survey coverage, because teacher and student survey outcomes are not available for all observations in the regression-eligible sample. This appendix also details the operational construction of the three predictor scores and the text feature battery.

Study Setting Detail

NEE was developed at the University of Missouri’s College of Education as a post-Race-to-the-Top evaluation system and launched in 2011–12 (Hunter & Bowser, 2026). Districts adopt the system voluntarily and pay a per-pupil fee that covers the digital observation protocol, rubric-aligned evaluator supports, secure data storage, and ongoing technical assistance.

In Missouri, leaders' school improvement planning follows a defined accountability process. Title 5 of the Missouri Code of State Regulations (5 CSR 20-100.125) establishes the Missouri School Improvement Program (MSIP), which requires each local board to adopt, monitor, and annually review a district-level Continuous School Improvement Plan (CSIP); the

CSIP in turn guides school-level Building Improvement Plans (BIPs), the structure and content of which the state leaves to local district and school discretion. NEE's leader evaluation tools, including the BIP-PO rubric we analyze, are situated in these regulations, processes, and standards. Furthermore, NEE instruments align with the Missouri Leadership Standards and Interstate Leaders Licensure Consortium standards.

The seven BIP-PO elements that we score sit within a broader NEE leader development and evaluation cycle. The BIP-PO is scored annually. NEE provides initial and ongoing training to both leaders and their supervisors, and supplies the rubric and exemplars used in scoring. Operationally, each BIP-PO is scored by a leader's supervisor, typically the leader's line supervisor (most often the district superintendent in these small rural districts). NEE offers asynchronous training options to help leaders complete the BIP-PO and supervisors score and provide feedback on the completed BIP-PO. Those training courses are voluntary for both leaders and supervisors.

Sample Construction Concerns

The regression-eligible analytic sample comprises 1,385 leader-year observations drawn from 731 unique leaders across 444 schools and 107 districts over the 2015–16 through 2021–22 school years (hereafter 2016–2022), excluding 2019–20. Constructing this sample from the full BIP universe of 1,733 observations requires several exclusion and linkage steps, each introducing a potential source of selection or measurement concern.

Concern 1: District Attrition from the NEE System

The 107-district BIP universe is narrower than NEE's full statewide footprint of more than 200 districts during the study window, reflecting that not every NEE district produces BIPs in every year and that some districts are excluded from the analytic file because of missing

administrative identifiers needed to link BIPs to DESE records or because the district participated in NEE for only part of the operational cycle in a given year.

Districts adopt NEE at different times, and a small number exited the system over the study window, producing flows into and out of the BIP-eligible universe and, by extension, into and out of our sample. If the districts or leaders who remain in the sample differ systematically from those who exit: i) the concurrent and predictive estimation samples could be compositionally nonequivalent, and ii) estimated predictor–outcome relationships could reflect selection. This opt-in/opt-out dynamic is the primary driver of the panel attrition documented in Table A1, alongside the within-district turnover of leaders across school years (see Panel A for changes in sample compositions).

We conduct a formal attrition analysis comparing covariate balance between predictive-eligible and concurrent-only observations across all five outcome domains. For each outcome, we compute normalized differences at the outcome level across all three predictor scores and the full set of 13 control covariates. Following Imbens and Rubin (2015), we flag normalized differences exceeding the conventional 0.25 threshold as indicative of meaningful compositional differences. We further assess whether observed imbalances affect estimated relationships by running inverse probability weighted (IPW) sensitivity analyses (reported below), which reweight the predictive sample to match the concurrent sample on observables.

Table A1 reports the results of the predictive sample attrition analysis. No outcome-level normalized difference exceeds the 0.25 threshold (see Panel B): achievement is closest at .238 ($p < .001$), followed by stability at .163, while VAM (.067), teacher survey (.089), and student survey (.017) are well below 0.25. However, covariate-level imbalance is more pronounced (see Panel D). Several school and staff characteristics show moderate-to-large normalized differences

for at least one outcome domain (see Panel D). FRPL share, rurality, the high-school indicator, share of teachers with advanced degrees, teacher count, and — the largest imbalances in the table, and notable given their bearing on the paper’s central equity comparisons — share of non-white students exceed the 0.25 severe-imbalance threshold for at least one outcome (non-white share reaches .382 for the teacher survey and .329 for the student survey). Average teacher experience is moderately imbalanced (up to .247 for stability) but falls just short of the 0.25 threshold for every outcome domain. These imbalances suggest that the predictive-eligible subsample overrepresents non-rural, lower-poverty, non-high-school buildings with more-credentialed teaching staff, consistent with the types of districts that maintain continuous NEE participation.

The covariate imbalances documented in Table A1 motivate the IPW sensitivity analyses reported in Tables A2, A3, and A4 (concurrent main, predictive main, and concurrent heterogeneous associations, respectively). Stabilized weights range narrowly (mean 1.08, $SD = 0.20$; range 0.83 to 2.03), indicating that the observable selection driving district attrition is modest and that only a small reweighting is needed to recover covariate balance against the concurrent sample. IPW-weighted concurrent estimates (Table A2) closely track the unweighted estimates in main-text Table 2: the LLM stability signal remains significant ($\beta = .061, p < .05$), Human and GRM continue to predict teacher survey ($\beta = .040$ and $.047$, both $p < .01$), and cross-method Wald tests confirm the stability divergence ($H = L \chi^2 = 7.76, p < .01$; $G = L \chi^2 = 7.35, p < .01$). IPW-weighted predictive estimates (Table A3) replicate the predictive null, with only the Human and GRM teacher-survey associations surviving ($\beta = .023$ and $.025$, both $p < .05$). IPW-weighted heterogeneity estimates (Table A4) replicate the main-text patterns from Table 4, including the race-by-stability inversion for Human and GRM in the highest-nonwhite tercile.

The concordance of weighted and unweighted results, together with the noted weight distribution, confirms that the attrition-driven compositional differences documented in Table A1 do not drive the main-text validity findings. We conclude that district attrition, while producing nontrivial compositional shifts, does not threaten the primary concurrent or predictive validity inferences.

Concern 2: Multiple Leaders Per Building

Approximately 34% of leader-year observations in the analytic sample are from school-years in which two or more leaders were evaluated in the same building. In those cases, school-level outcome measures (achievement, value-added, stability, student survey) are identical across the leaders sharing a building-year, but each leader's predictor scores may differ because they are derived from that leader's own BIP narrative. This within-building clustering of shared outcomes and distinct predictors raises two analytic concerns: first, standard errors may understate uncertainty if the effective sample size is smaller than the observation count suggests; second, the predictor–outcome relationship could be attenuated if within-building predictor variation is mostly noise relative to between-building variation in outcomes.

We test sensitivity to this clustering by restricting the sample to single-leader school-years, which eliminates the shared-outcome concern entirely. The single-leader restriction analysis (Table A5; $N = 519$ to 867 across concurrent outcomes) finds substantively unchanged, and often more precise, coefficient estimates compared to the full sample. Several results strengthen under the restriction: GRM achievement reaches $\beta = .062$ ($p < .01$), and the LLM stability signal remains significant ($\beta = .066$, $p < .05$). Restricting to single-leader observations does, however, sharpen several predictive validity coefficients in the direction that within-school measurement error dilution would predict: Human achievement moves from $\beta = .003$ (Table 3)

to $\beta = .055$ ($p < .05$) and LLM stability from $\beta = .058$ to $\beta = .068$ ($p < .05$). We read this as suggestive that the predictive null partly reflects school-years in which two evaluated leaders share the same outcomes, rather than as independent confirmation of dilution.

The robustness of results under the single-leader restriction, combined with district-clustered standard errors in all main specifications, indicates that the multiple-leaders-per-building feature of our sample does not substantially distort the main-text findings.

Concern 3: COVID-19 Era Exclusions

Excluding the 2019–20 school year from the panel, necessitated by disruptions to both the NEE evaluation cycle and state assessment administration during spring 2020 COVID-19 closures, introduces two measurement issues. First, 2019 base-year observations are linked to 2021 ‘same-year’ outcomes rather than 2020 outcomes, creating a two-year rather than one-year lag for the lagged achievement control used in achievement models. If the lagged control’s predictive power decays over a longer horizon, the achievement models for the 2019 cohort may have weaker covariate adjustment than models for other cohorts. Second, future-year (predictive) linkages for the 2019 cohort likewise reflect a two-year rather than one-year window, potentially introducing heterogeneity in the predictive validity estimates if predictor–outcome relationships attenuate or strengthen over longer horizons.

We address this concern through two complementary sensitivity analyses (Table A6). The exclusion approach (Panels A, B, E, and F) drops the 2019 base-year cohort entirely and re-estimates all models using the remaining years where the one-year linkage structure holds uniformly. The interaction approach (Panels C, D, G, and H) retains all observations but includes a COVID-year indicator interacted with all right-hand-side terms, allowing the data to reveal whether predictor–outcome relationships differ for the two-year-lag cohort. Concurrent validity

panels report only achievement and VAM outcomes because these are the only models in which the lagged achievement control or VAM construction is mechanically affected by the two-year lag; predictive validity panels report all five outcomes.

Both approaches produce substantively unchanged coefficient estimates and inference patterns across all five outcome domains and all three predictor scores. The interaction approach is particularly informative: the COVID-year interaction terms for Human and GRM on achievement and VAM are large and positive, while the LLM interaction terms are smaller and mixed in sign. For predictive validity, the LLM interaction with the two-year-lag indicator is significantly negative for achievement ($\beta = -.101, p < .01$) and VAM ($\beta = -.322, p < .01$), suggesting that the LLM predictive signal, when present, attenuates over longer prediction windows. Critically, neither approach changes any qualitative conclusion, indicating that the necessary COVID-era exclusion does not threaten the main-text inferences.

Concern 4: Survey Coverage and Availability

The teacher survey and student survey outcomes are not universally available across the regression-eligible sample. The teacher survey covers 1,014 of 1,385 leader-year observations (73%), with a modest overrepresentation of high-poverty and high-minority schools (normalized differences below the 0.25 meaningful-difference threshold reported later in this section, see Table A7). The student survey covers 978 of 1,385 observations (71%). If the subset of observations with survey data differs systematically from the full sample, estimates for the survey outcomes could reflect a selected subpopulation rather than the broader NEE leader population. Unlike the three administrative outcomes (achievement, VAM, stability), which are available for nearly all observations, the survey outcomes require that districts administer the relevant instruments in a given year and achieve sufficient response rates.

Table A7 reports normalized differences between each outcome-specific estimation sample and the full regression-eligible base sample ($N = 1,385$) across 17 covariates. This comparison reveals the magnitude and direction of compositional differences arising from survey availability. For each of the five concurrent estimation samples, the five predictive estimation samples, and the single common (pooled) sample, we assess whether any normalized difference exceeds the 0.25 threshold.

Table A7 reports that no normalized difference exceeds $|0.25|$ across the 17 covariates and 11 estimation samples, indicating that all outcome-specific estimation samples, including the teacher and student survey subsamples, are compositionally similar to the base sample. The largest imbalances appear for the future teacher survey and future student survey samples, consistent with the expectation that predictive-sample selection compounds concurrent-sample selection, but even these remain below the conventional threshold.

The absence of meaningful compositional differences between the survey subsamples and the full sample provides assurance that the teacher and student survey validity estimates are not driven by selection into survey coverage. The survey-outcome findings reported in the main text can be interpreted as representative of the broader NEE leader population, subject to the usual caveats about external validity beyond this evaluation system.

Predictor Construction Detail

This section expands on the operational construction of the three scoring approaches summarized in the main text Measures subsection. All three scores are standardized within the school-year prior to analysis to place them on a common metric and facilitate cross-method comparisons. Table A12 reports the bivariate correlations that motivate the cross-method comparison: the two human-based scores correlate at $r = .92$, while the LLM correlates at $r = .12$

with Human and $r = .07$ with GRM, and the LLM's concurrent achievement correlation is negative ($-.17$), whereas both human-based scores are positive (.22 and .24).

Human Supervisor Ratings

Supervisors can assign ratings of 0 (does not meet standard) to 4 (exceeds standard) on each of the seven BIP elements after reviewing the leader's narrative responses, though we only observe a range of 1-4. We compute the mean rating across available elements for each leader-year to construct a continuous supervisor score.

Not all BIPs received human ratings for every element; in these cases, we average over the elements that were rated, whereas LLM scores are available for all seven elements in every BIP. As a robustness check for predictor construction, we re-estimate all analyses using a complete-case variant that requires all seven rubric elements to be scored within each BIP, thereby contracting the sample by roughly 10 to 15 percent but ensuring that predictor scores are constructed from identical element sets across observations (Table A8). The core concurrent validity pattern replicates under this restriction: the LLM stability signal reaches $\beta = .056$ ($p < .05$), Human and GRM predict teacher survey ($\beta = .043$ and $.051$, both $p < .01$), and cross-method divergence on stability is confirmed ($H = L \chi^2 = 7.10, p < .01$; $G = L \chi^2 = 6.84, p < .01$). The joint F for GRM + LLM is significant for achievement, stability, and teacher survey, and substantive conclusions are unchanged. In our analytic sample, the mean human rating is 3.66 ($SD = 0.50$) on the 1–4 observed scale, indicating substantial ceiling compression consistent with documented leniency bias in educator evaluation systems (Kraft & Gilmour, 2017; Weisberg et al., 2009).

GRM Theta Scores

We apply a Graded Response Model (GRM) to the seven element-level human ratings, producing a latent trait estimate for each leader-year (Samejima, 1969). As described in the main text, GRM rescaling refines the statistical treatment of human ratings without introducing new information about the documents. GRM scores provide a wider effective range ($M = 0.01$, $SD = 0.83$; range -2.63 to 0.91) than raw human ratings, partially decompressing scores at the ceiling. The correlation between Human and GRM scores is $r = .92$.

LLM Scores

We use a large language model (LLM), specifically GPT-4o-mini, to score the same narrative text that supervisors evaluate, producing ratings on the same rubric scale. Unlike Human and GRM scores, which derive from supervisor judgment, LLM scores derive from the text content of leader responses as processed by the language model. This distinction is conceptually important: the LLM scoring approach treats BIP evaluation as a text-comprehension and rubric-alignment task rather than a human judgment task, and the degree to which LLM scores converge with human scores (and predict outcomes similarly) speaks to the extent to which the underlying rubric-relevant information is encoded in the narrative text itself.

Because readers may be less familiar with LLM-based scoring than with traditional supervisor ratings or IRT models, we describe the LLM scoring procedure in detail here; Online Appendix D provides the actual prompt excerpts and the configuration summary table that operationalize what is described below. At a high level, the process involves providing the LLM with (a) the text of the rubric element, (b) the leader's narrative response, and, in some configurations, (c) five scored exemplar responses to anchor the model's rating scale. The model then returns a rating on the 0–4 scale for that element. This process is repeated independently for

each of the seven BIP elements in each leader year, yielding a seven-element rating profile that parallels the human supervisor's profile.

We tested four LLM configurations that varied along two dimensions: the provision of scored exemplars (zero-shot versus few-shot) and, for two of the three few-shot configurations, the leader-linked sampling frame used to select training exemplars; the third few-shot configuration draws instead from a fixed, randomly sampled exemplar pool (see Configuration (b) below). The four configurations are as follows. Configuration (a), zero-shot, provides only the rubric text and the leader's response with no scored examples; the model must infer the rating scale solely from the rubric description. Configuration (b), few-shot, provides five scored exemplar responses alongside the rubric, giving the model concrete anchors for each rating level (cf. Brown et al., 2020). Because these five exemplars are drawn from a fixed, randomly sampled pool rather than a leader-linked sampling frame, Configuration (b) varies only on the first dimension (exemplar provision) and does not correspond to a level of the sampling-frame dimension. Configuration (c), few-shot within leader-year across elements (WPYAE), draws text from the same leader and year across all elements, testing whether within-leader context improves scoring consistency by providing adjacent-element information. Configuration (d), few-shot within leader-element across years (WPEAY), draws exemplars from the same leader and rubric element but across different school years, testing whether within-element text improves calibration by exposing the model to within-rubric-element variation over time. Configurations (c) and (d) address the known sensitivity of LLM scoring to exemplar selection (Zhao et al., 2021).

Cross-configuration correlations range from $r = .84$ to $.91$ (Table A9 Panel A), indicating substantial convergence across prompting strategies. We average the four within-year

standardized configuration scores to form a composite LLM Index ($M = 0.01$, $SD = 0.95$), which serves as our primary LLM predictor. The composite approach reduces configuration-specific noise and avoids requiring selection of a single prompting strategy.

The complete-case predictor variant described in the Human Supervisor Ratings subsection is constructed in parallel for the LLM Index; the complete-case and primary predictors correlate at $r = 1.000$ for both Human and GRM, a mechanical result, and $r = .991$ for the LLM Index (Table A9, Panel B), and the results are substantively unchanged.

Text Feature Construction

Because any AI-based scoring system raises the question of whether it captures substantive rubric content or merely surface text properties, we extract a battery of 12 text features from the BIP narratives for use in the bias and validity analyses described in the main text Methods section. The features span four dimensions widely recognized in the automated essay scoring (AES) literature as the operative axes of surface text variation: production/length, lexical sophistication, readability and syntactic complexity, and topic-content alignment (Page, 1966; Shermis & Burstein, 2013; Crossley & McNamara, 2012). Table A10 provides a reference table that maps each feature to its AES category and definition.

We reduce dimensionality via exploratory factor analysis with principal-factor extraction and varimax rotation. Parallel analysis (paran, 500 iterations) returned a missing retained-factor count on this sample; we retain three factors, which produce an interpretable and stable rotation and which we label Volume (loading on word count, sentence count, and unique word count), Sophistication (loading on average word length, content-word ratio, and parse depth), and Rubric Overlap (loading on lexical overlap measures with the rubric text). Table A11 reports the full factor-loading matrix (Panel A) and the correlations between the factor scores and the three

predictor scores (Panel B). The three-factor solution accounts for a substantial share of the total variation in the text feature battery and provides interpretable dimensions for the text-feature bias analyses in the main text (Panel C).

Appendix A References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135.
- Hunter, S. B., & Bowser, K. M. (2026). Reformed teacher evaluation in rural Missouri: Main and moderated relationships with student achievement and relationships-to-expenditure ratios. *Education Policy Analysis Archives*, 34. <https://doi.org/10.14507/epaa.34.9248>
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.
- Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 6300–6308). <https://doi.org/10.24963/ijcai.2019/879>
- Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8-75). Chief of Naval Technical Training, Naval Air Station Memphis.
- Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5), 234–249.

- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5), 238–243.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 1–97.
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project.
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 139, pp. 12697–12706). PMLR.

Table A1

Predictive Sample Attrition Analysis: Covariate Balance Between Predictive-Eligible and Concurrent-Only Observations

	Achievement	VAM	Stability	Teacher Survey	Student Survey
Panel A: Sample Composition					
Concurrent-eligible (base $N = 1,385$)					
Predictive-eligible ($\text{in_pred} = 1$)	816	654	723	811	657
Attrited: concurrent-only ($\text{in_pred} = 0$)	417	579	510	422	576
Panel B: Concurrent Outcome Level — Mean (SD) by Attrition Group					
Predictive-eligible mean	.086	.040	−.081	2.407	2.139
Attrited mean	−.138	−.017	.049	2.380	2.136
Normalized difference	.238**	.067	.163**	.089	.017
p-value (t-test)	.000	.254	.004	.395	.818
Panel C: Predictor Score Balance — Mean (predictive-eligible vs. attrited)					
Human Supervisor (H)					
Predictive-eligible mean	.083	.106	.103	.024	.043
Attrited mean	−.083	−.062	−.081	.033	.009
Normalized difference	.173*	.178*	.192*	.009	.035
GRM Theta (G)					
Predictive-eligible mean	.058	.063	.053	−.005	.010
Attrited mean	−.099	−.060	−.064	.024	−.001

Normalized difference	.187*	.147*	.138*	.034	.014
LLM Index (L)					
Predictive-eligible mean	.040	.063	.064	.087	.039
Attrited mean	-.075	-.068	-.087	-.163	-.041
Normalized difference	.121*	.139*	.160*	.258**	.085

Panel D: Full Covariate Balance (all 13 controls)

% Students FRPL	.257**	.252**	.184*	.354**	.108
% Non-white students	.015	.067	.054	.382**	.329**
Teacher count	.157*	.057	.050	.060	.266**
High school	.263**	.217*	.199*	.054	.094
Rural school	.342**	.263**	.288**	.019	.207*
Avg. teacher experience	.227*	.246*	.247*	.213*	.060
% Teachers adv. degree	.269**	.244*	.215*	.150*	.165*
% Teachers non-white	.099	.096	.122	.138	.075
% Teachers male	.184*	.160*	.151*	.075	.216*
Avg. admin experience	.030	.031	.044	.137	.139
% Admins non-white	.071	.049	.084	.169*	.094
% Admins male	.168*	.085	.084	.015	.076
% Admins adv. degree	.015	.032	.013	.045	.047
Outcome-level attrition flag?	No	No	No	No	No
Covariate imbalance (Panel D)?	Yes	Yes	Yes	Yes	Yes
Joint logit imbalance (Section 16)?	Yes	Yes	Yes	No	Yes

Note. Concurrent-eligible base sample: $N = 1,385$ (regression-eligible leaders). Predictive-eligible = obs with non-missing predictive outcome and controls; attrited = concurrent-eligible obs missing the predictive outcome. Panel A reports outcome-specific Ns. Panel B compares mean concurrent outcome levels across groups; normalized difference = $|(\mu_1 - \mu_0) / \sqrt{[(SD_1^2 + SD_0^2)/2]}|$. Panel C reports predictor score means by group for the achievement attrition flag (largest attrition). Panel D reports normalized differences for all 13 controls. ** = NormDiff $\geq .25$ (severe imbalance); * = NormDiff $\geq .15$ (moderate imbalance). Three distinct attrition flags are reported, because they answer different questions and do not agree. Outcome-level attrition flag = Yes if the concurrent outcome's own normalized difference reaches .25 (Panel B); no outcome reaches it, so all five read No. Covariate imbalance (Panel D) = Yes if the outcome has at least one severe (**) covariate imbalance; all five do — Achievement (FRPL, high school, rural, teachers w/ adv. degree), VAM (FRPL, rural), Stability (rural), Teacher Survey (FRPL, % non-white students), Student Survey (% non-white students, teacher count), 11 cells in total. Joint logit imbalance (Section 16) = Yes if a logit

of predictive-sample membership on all 16 covariates rejects joint balance at $p < .05$: Achievement $\chi^2(16) = 75.60$, VAM 71.37, Stability 84.66, and Student Survey 40.59 all reject ($p \leq .0006$), while Teacher Survey does not ($\chi^2(16) = 19.57, p = .240$). Teacher Survey nonetheless carries the two largest single imbalances in the table (% non-white students .382, FRPL .354).

Table A2

IPW Sensitivity: Concurrent Validity

	Achievement	VAM	Stability	Teacher Survey	Student Survey
Panel A: District + Year FE, Lean Controls					
Human Supervisor (H)	.020 (.015)	-.026 (.042)	-.033 (.025)	.040** (.010)	.005 (.005)
GRM Theta (G)	.036* (.016)	.019 (.052)	-.042 (.036)	.047** (.015)	.010† (.006)
LLM Index (L)	.004 (.017)	.017 (.029)	.061* (.027)	.031† (.016)	-.007 (.006)
χ^2 : H = G	4.20*	6.46*	0.33	1.21	1.89
χ^2 : H = L	0.80	0.81	7.76**	0.36	4.42*
χ^2 : G = L	2.77†	0.00	7.35**	1.19	6.09*
<i>N</i>	1,354	1,335	1,374	997	965
Panel B: GRM + LLM Joint Model					
GRM Theta (G)	.037* (.016)	.015 (.053)	-.061† (.037)	.041** (.014)	.012* (.006)
LLM Index (L)	-.001 (.017)	.014 (.029)	.071* (.028)	.023 (.015)	-.009 (.006)
<i>F</i> (GRM + LLM = 0)	2.65†	0.22	4.11*	5.58**	2.96†
<i>N</i>	1,354	1,335	1,374	997	965

Note. Entries are standardized regression coefficients from IPW-weighted concurrent validity models. Inverse probability weights reweight the predictive sample to match the concurrent sample's covariate distribution (stabilized weights: mean = 1.08, *SD* = 0.20, range 0.83–2.03). Panel A reports coefficients from separate reghdfe regressions on a common suest sample; chi-squared statistics test equality of predictor coefficients via suest. Panel B reports a joint reghdfe with GRM Theta and LLM Index entered

simultaneously; F tests examine their joint significance. All models include district and year fixed effects with lean controls. Standard errors (clustered by district) in parentheses. † $p < .10$. * $p < .05$. ** $p < .01$.

Table A3

IPW Sensitivity: Predictive Validity

	Achievement (t+1)	VAM (t+1)	Stability (t+1)	Teacher Survey (t+1)	Student Survey (t+1)
Panel A: District + Year FE					
Human Supervisor (H)	−.003 (.021)	.013 (.046)	−.025 (.035)	.023* (.009)	.003 (.009)
GRM Theta (G)	.012 (.027)	.080 (.064)	−.021 (.039)	.025* (.012)	.004 (.011)
LLM Index (L)	−.003 (.020)	.024 (.037)	.049 (.041)	.000 (.012)	.003 (.006)
χ^2 : H = G	1.86	4.63*	0.05	0.12	0.07
χ^2 : H = L	0.00	0.05	2.44	3.09†	0.00
χ^2 : G = L	0.28	1.00	2.82†	3.62†	0.01
<i>N</i>	908	733	712	869	714
Panel B: GRM + LLM Joint Model					
GRM Theta (G)	.013 (.027)	.076 (.062)	−.036 (.039)	.027* (.012)	.004 (.010)
LLM Index (L)	−.005 (.020)	.013 (.035)	.054 (.040)	−.004 (.011)	.002 (.006)
<i>F</i> (GRM + LLM = 0)	0.14	0.78	1.39	2.78†	0.13
<i>N</i>	908	733	712	869	714

Note. Entries are standardized regression coefficients from IPW-weighted predictive validity models. Inverse probability weights reweight the predictive sample to match the concurrent sample's covariate distribution (stabilized weights: mean = 1.08, *SD* = 0.20, range 0.83–2.03). Panel A reports coefficients from separate reghdfe regressions on a common suest sample; chi-squared statistics test equality of predictor coefficients via suest. Panel B reports a joint reghdfe with GRM Theta and LLM Index entered

simultaneously; F tests examine their joint significance. All models include district and year fixed effects with lean controls. Standard errors (clustered by district) in parentheses. † $p < .10$. * $p < .05$. ** $p < .01$.

Table A4

IPW Sensitivity: Concurrent, Heterogeneous Associations by School Context, Total Predictor Slopes at Each Moderator Level

	Achievement	VAM	Stability	Teacher Survey	Student Survey
Panel A: Poverty Tercile					
Human Supervisor (H)					
Low (T1)	.056 (.043)	-.006 (.075)	.060 (.060)	.010 (.019)	.005 (.012)
Medium (T2)	.023 (.064)	.010 (.122)	.006 (.091)	.032 (.038)	.005 (.021)
High (T3)	.002 (.067)	-.049 (.112)	-.086 (.087)	.049 [†] (.028)	.005 (.019)
GRM Theta (G)					
Low (T1)	.079 [†] (.040)	.018 (.098)	.022 (.079)	.024 (.020)	.015 (.016)
Medium (T2)	.034 (.059)	.028 (.150)	.013 (.114)	.029 (.040)	.005 (.027)
High (T3)	.013 (.064)	.021 (.155)	-.107 (.119)	.064* (.029)	.010 (.025)
LLM Index (L)					
Low (T1)	-.008 (.017)	-.036 (.044)	.045 [†] (.025)	.008 (.013)	-.021* (.009)
Medium (T2)	.016 (.036)	.113 (.074)	.064 (.051)	-.002 (.026)	.003 (.015)
High (T3)	.004 (.049)	.007 (.076)	.095 [†] (.057)	.072* (.031)	-.006 (.015)

<i>N</i>	1,354	1,335	1,374	997	965
Panel B: Racial Composition Tercile					
Human Supervisor (H)					
Low nonwhite (T1)	.004 (.032)	.051 (.058)	.021 (.030)	.042 [†] (.022)	.006 (.008)
Medium nonwhite (T2)	.015 (.052)	-.104 (.099)	.030 (.062)	.030 (.038)	.012 (.013)
High nonwhite (T3)	.030 (.046)	-.051 (.099)	-.131** (.046)	.041 (.036)	.002 (.012)
GRM Theta (G)					
Low nonwhite (T1)	.017 (.035)	.098 (.077)	.026 (.049)	.053* (.023)	.010 (.013)
Medium nonwhite (T2)	.035 (.058)	-.094 (.119)	.051 (.081)	.049 (.041)	.013 (.021)
High nonwhite (T3)	.035 (.051)	.024 (.138)	-.178* (.076)	.037 (.044)	.009 (.020)
LLM Index (L)					
Low nonwhite (T1)	-.059 [†] (.031)	-.032 (.062)	.066 (.056)	.018 (.028)	.004 (.011)
Medium nonwhite (T2)	.027 (.046)	.080 (.098)	.043 (.086)	.020 (.046)	-.030 [†] (.018)
High nonwhite (T3)	.038 (.052)	.003 (.099)	.080 (.101)	.052 (.051)	.002 (.019)
<i>N</i>	1,354	1,335	1,374	997	965

Note. Entries are total predictor slopes at each moderator level (tercile) from IPW-weighted heterogeneity models. Each panel adds poverty-tercile (Panel A) or racial-composition-tercile (Panel B) interaction terms to the concurrent IPW model; total slopes are computed at each tercile level. IPW weights as in Tables A2 and A3. Standard errors (clustered by district) in parentheses. † $p < .10$. * $p < .05$. ** $p < .01$.

Table A5

Robustness: Single-Leader Sample Replication

	Achievement	VAM	Stability	Tch Survey	Std Survey
Concurrent Outcomes ($N = 519\text{--}867$)					
Panel A: Separate Predictors					
Human Supervisor (H)	.054** (.016)	−.037 (.041)	.009 (.020)	.047** (.015)	.003 (.009)
GRM Theta (G)	.062** (.019)	−.009 (.049)	.011 (.032)	.060** (.019)	.007 (.011)
LLM Index (L)	.006 (.013)	.012 (.037)	.066* (.028)	.016 (.015)	−.011† (.007)
χ^2 : H=G	0.68	2.05	0.02	2.18	0.54
χ^2 : H=L	5.92*	1.07	4.79*	3.85*	1.74
χ^2 : G=L	6.95**	0.15	2.77†	6.66**	2.02
N	859	845	867	594	519
Panel B: GRM + LLM Joint Model					
GRM Theta (G)	.063** (.020)	−.013 (.050)	−.007 (.032)	.059** (.018)	.010 (.011)
LLM Index (L)	−.003 (.013)	.012 (.037)	.066* (.028)	.016 (.015)	−.011† (.006)
F (GRM + LLM = 0)	5.16**	0.08	2.80†	5.30**	1.70
N	859	845	867	594	519
Predictive Outcomes ($N = 350\text{--}550$)					
	Achievement	VAM	Stability	Tch Survey	Std Survey
Panel C: Separate Predictors					
Human Supervisor (H)	.055* (.024)	−.013 (.057)	−.024 (.038)	.023† (.014)	.008 (.011)

GRM Theta (G)	.057† (.033)	.021 (.064)	−.034 (.037)	.026 (.021)	.019 (.014)
LLM Index (L)	.012 (.021)	−.015 (.043)	.068* (.029)	−.006 (.012)	.005 (.007)
χ^2 : H=G	0.02	1.38	0.36	0.08	3.25†
χ^2 : H=L	1.88	0.00	3.08†	2.83†	0.09
χ^2 : G=L	1.34	0.28	4.76*	2.22	1.43
<i>N</i>	550	484	458	492	350

Panel D: GRM + LLM Joint Model

GRM Theta (G)	.056 (.035)	.027 (.066)	−.060 (.044)	.030 (.021)	.018 (.014)
LLM Index (L)	.012 (.021)	−.015 (.042)	.068* (.029)	−.006 (.012)	.005 (.007)
<i>F</i> (GRM + LLM = 0)	1.65	0.15	3.30*	1.28	0.92
<i>N</i>	550	484	458	492	350

Note. Entries are standardized regression coefficients from the single-leader sample (school-years with only one NEE-evaluated leader; $N = 519\text{--}867$ concurrent, 350–550 predictive). Panels A and C report coefficients from separate reghdfe regressions on a common suest sample; chi-squared statistics test equality of predictor coefficients via suest. Panels B and D report a joint reghdfe with GRM Theta and LLM Index entered simultaneously; *F* tests examine their joint significance. District and year fixed effects with lean controls (Year FE + District FE + Lean Controls). Lagged achievement was included as a control for Achievement outcomes. Standard errors (clustered by district) in parentheses. † $p < .10$. * $p < .05$. ** $p < .01$.

Table A6

COVID-Era Skip-Logic Sensitivity: Exclusion and Interaction Approaches for Concurrent and Predictive Validity

Concurrent Validity, Exclusion Approach (Drop springyear = 2021)		
	Achievement	VAM
Panel A: Separate Predictors		
Human Supervisor (H)	0.005 (0.017)	−0.021 (0.038)
GRM Theta (G)	0.022 (0.018)	0.027 (0.047)
LLM Index (L)	0.001 (0.015)	0.008 (0.030)
χ^2 : H = G	2.66	6.33*
χ^2 : H = L	0.08	0.34
χ^2 : G = L	1.08	0.10
<i>N</i>	1,203	1,198
Panel B: GRM + LLM Joint Model		
GRM Theta (G)	0.022 (0.018)	0.026 (0.052)
LLM Index (L)	−0.003 (0.016)	0.004 (0.033)
Joint <i>F</i> (GRM + LLM)	0.750	0.233
<i>N</i>	1,203	1,198

Concurrent Validity, Interaction Approach (year_2021 $\times$ All RHS)**Panel C: Separate Predictors**

Human Supervisor (H)	0.003 (0.015)	−0.026 (0.041)
Human Supervisor $\times$ 2021	+0.161 (0.033)	+0.170 (0.079)
GRM Theta (G)	0.016 (0.018)	0.017 (0.052)
GRM Theta $\times$ 2021	+0.182 (0.045)	+0.227 (0.089)
LLM Index (L)	0.002 (0.015)	0.017 (0.029)
LLM Index $\times$ 2021	+0.008 (0.045)	−0.047 (0.065)
Joint F (H $\times$ interactions), p	.000	.000
Joint F (G $\times$ interactions), p	.000	.000
Joint F (L $\times$ interactions), p	.000	.002
N	1,354	1,335

Panel D: GRM + LLM Joint Model

GRM Theta (G)	0.017 (0.018)	0.013 (0.056)
GRM Theta $\times$ 2021	+0.183	+0.228

	(0.045)	(0.089)
LLM Index (L)	−0.003	0.013
	(0.015)	(0.032)
LLM Index × 2021	+0.023	−0.025
	(0.036)	(0.064)
Joint F (GRM+LLM interactions), p	.000	.000
N	1,354	1,335

Predictive Validity, Exclusion Approach (Drop springyear = 2019)

	Achievement (t+1)	VAM (t+1)	Stability (t+1)	Tch Survey (t+1)	Std Survey (t+1)
Panel E: Separate Predictors					
Human Supervisor (H)	0.008	0.051	−0.020	0.025**	0.001
	(0.021)	(0.042)	(0.036)	(0.010)	(0.011)
GRM Theta (G)	0.021	0.094†	−0.024	0.024†	0.002
	(0.027)	(0.049)	(0.046)	(0.014)	(0.013)
LLM Index (L)	−0.004	0.032	0.027	0.003	0.005
	(0.020)	(0.040)	(0.039)	(0.015)	(0.007)
χ^2 : H = G	1.24	4.01*	0.04	0.03	0.03
χ^2 : H = L	0.28	0.15	1.10	2.69	0.20
χ^2 : G = L	0.86	1.28	1.45	1.97	0.07
N	818	657	592	709	606

Panel F: GRM + LLM Joint Model

GRM Theta (G)	0.023 (0.026)	0.090 (0.049)	−0.030 (0.044)	0.024† (0.014)	0.001 (0.013)
LLM Index (L)	−0.006 (0.020)	0.023 (0.040)	0.030 (0.036)	−0.002 (0.014)	0.005 (0.006)
Joint F (GRM + LLM)	0.433	1.898	0.759	1.656	0.328
N	818	657	592	709	606

Predictive Validity, Interaction Approach (two_year_lag × All RHS)

	Achievement (t+1)	VAM (t+1)	Stability (t+1)	Tch Survey (t+1)	Std Survey (t+1)
Panel G: Separate Predictors					
Human Supervisor (H)	−0.005 (0.021)	0.034 (0.037)	−0.009 (0.035)	0.024* (0.010)	0.003 (0.010)
Human Supervisor × 2019	+0.042 (0.088)	−0.290 (0.287)	−0.083 (0.123)	+0.001 (0.031)	−0.005 (0.019)
GRM Theta (G)	0.004 (0.028)	0.076 (0.049)	−0.016 (0.043)	0.023† (0.013)	0.004 (0.012)
GRM Theta × 2019	+0.060 (0.084)	−0.164 (0.273)	+0.011 (0.129)	+0.023 (0.032)	+0.004 (0.020)
LLM Index (L)	−0.001 (0.020)	0.043 (0.040)	0.051 (0.039)	0.003 (0.015)	0.007 (0.007)
LLM Index × 2019	−0.101** (0.033)	−0.322** (0.115)	+0.102 (0.078)	−0.030 (0.027)	−0.003 (0.021)

Joint F ($H \times$ interactions), p	.017	.000	.612	.031	.000
Joint F ($G \times$ interactions), p	.019	.000	.205	.020	.000
Joint F ($L \times$ interactions), p	.017	.000	.341	.018	.001
N	908	733	712	869	714
Panel H: GRM + LLM Joint Model					
GRM Theta (G)	0.008	0.079	−0.033	0.025†	0.003
	(0.027)	(0.049)	(0.041)	(0.013)	(0.012)
GRM Theta $\times$ 2019	+0.100	−0.041	+0.013	+0.022	+0.005
	(0.085)	(0.258)	(0.126)	(0.030)	(0.020)
LLM Index (L)	−0.002	0.035	0.055	−0.001	0.006
	(0.020)	(0.041)	(0.037)	(0.014)	(0.006)
LLM Index $\times$ 2019	−0.123**	−0.327*	+0.102	−0.029	−0.003
	(0.031)	(0.137)	(0.079)	(0.028)	(0.021)
Joint F (GRM+LLM interactions), p	.000	.000	.192	.023	.000
N	908	733	712	869	714

Note. All models use the preferred specification: District FE + Year FE + Lean Controls, and lagged achievement for achievement outcomes. Standard errors clustered by district in parentheses. Concurrent panels (A–D) report Achievement and VAM only. These are the two concurrent outcomes structurally tied to prior-year achievement — Achievement through its lagged-achievement control, VAM through its underlying value-added model, which is itself estimated from prior achievement — and are therefore the only concurrent outcomes this base-year skip logic can plausibly affect; Stability, Teacher Survey, and Student Survey do not depend on prior-year achievement and are not part of this sensitivity check. Predictive panels (E–H) report all five outcomes, since the $t+1$ lag structure being tested applies uniformly across outcome domains.

Table A7

Sample Comparison: Normalized Differences Between Outcome-Specific Estimation Samples and the Regression-Eligible Base

Sample

	<i>N</i>	Mean	<i>SD</i>	Ach	VAM	Stab	TchSvy	StdSvy	F.Ach	F.VAM	F.Stab	F.TchSvy	F.StdSvy	Common
Human	1385	0.026	0.953	0.000	0.000	0.000	0.048	−0.010	−0.042	−0.063	−0.086	0.026	0.007	−0.016
GRM	1385	0.007	0.828	0.000	0.000	0.000	0.067	0.001	−0.045	−0.053	−0.057	0.040	0.022	0.023
LLM	1385	0.006	0.949	0.000	0.006	0.000	−0.093	−0.038	−0.041	−0.056	−0.062	−0.114	−0.063	−0.140
% FRPL	1385	0.503	0.282	0.000	0.007	0.000	−0.123	−0.073	0.082	0.114	0.025	−0.150	−0.080	−0.089
% Non-White Students	1385	0.257	0.293	0.000	0.000	0.000	−0.093	−0.097	0.003	0.035	0.010	−0.119	−0.146	−0.177
Teacher Count	1385	42.074	22.624	0.000	−0.010	0.000	0.020	−0.053	−0.047	−0.022	−0.015	−0.011	−0.114	−0.040
High School	1385	0.244	0.430	0.000	−0.005	0.000	−0.028	−0.075	0.079	0.103	0.076	−0.028	−0.055	0.084
Rural	1385	0.548	0.498	0.000	0.008	0.000	−0.058	0.006	0.094	0.102	0.126	−0.042	0.064	0.078
Teacher Experience (yrs)	1385	12.872	2.554	0.000	0.005	0.000	0.087	0.114	−0.068	−0.105	−0.083	0.115	0.059	0.026
Teacher Adv. Degree	1385	0.567	0.162	0.000	−0.012	0.000	0.056	0.006	−0.081	−0.105	−0.075	0.087	−0.057	−0.094
Teacher Non-White	1385	0.040	0.077	0.000	0.000	0.000	−0.050	−0.052	0.041	0.054	0.041	−0.038	−0.028	−0.066
Teacher Male	1385	0.214	0.145	0.000	−0.007	0.000	−0.042	−0.103	0.070	0.085	0.057	−0.028	−0.105	0.007
Admin Experience (yrs)	1385	17.525	4.874	0.000	0.004	0.000	0.057	0.032	0.009	0.006	0.045	0.072	0.086	0.017
Admin Non-White	1385	0.059	0.177	0.000	−0.011	0.000	−0.054	−0.045	0.029	0.023	0.017	−0.055	−0.051	−0.090
Admin Male	1385	0.434	0.389	0.000	0.000	0.000	−0.008	−0.060	0.057	0.044	0.018	−0.008	−0.047	0.000
Admin Adv. Degree	1385	0.642	0.376	0.000	0.000	0.000	−0.011	0.003	−0.005	0.019	−0.027	−0.035	−0.003	−0.019
Lagged Achievement	1364	0.010	0.937	0.000	−0.007	0.000	0.081	0.076	−0.061	−0.113	−0.078	0.102	0.101	0.094

Notes. Columns show normalized differences between each outcome-specific estimation sample and the regression-eligible base sample ($N = 1,385$). Normalized difference = $(\text{mean_base} - \text{mean_sub}) / \sqrt{[(\text{SD}^2_base + \text{SD}^2_sub) / 2]}$. Values $> |0.25|$ (bolded) indicate meaningful compositional differences (Imbens and Rubin,

2015). Concurrent outcomes: Ach = achievement, VAM = value-added, Stab = stability, TchSvy = teacher survey, StdSvy = student survey. Predictive outcomes prefixed with F. Common = common concurrent sample (all five concurrent outcomes non-missing).

Table A8

Robustness: Re-estimation Using Scores Based on All Seven Evaluation Elements

	Achievement	VAM	Stability	Tch Survey	Std Survey
Concurrent Outcomes					
Panel A: Separate Predictors					
Human Supervisor	.035* (.015)	.023 (.044)	-.025 (.023)	.043** (.012)	.004 (.004)
GRM Theta	.058** (.017)	.070† (.048)	-.044 (.036)	.051** (.015)	.009 (.007)
LLM Index	.003 (.016)	.015 (.032)	.056* (.025)	.015 (.014)	-.011† (.007)
χ^2 : H=G	5.72*	9.39**	1.39	0.95	1.19
χ^2 : H=L	2.59	0.03	7.10**	6.41*	3.69†
χ^2 : G=L	6.33*	1.04	6.84**	9.00**	4.42*
<i>N</i>	1220	1204	1238	911	878
Panel B: GRM + LLM Joint Model					
GRM Theta	.059** (.017)	.070 (.049)	-.044 (.036)	.051** (.016)	.009 (.007)
LLM Index	.007 (.016)	.018 (.029)	.062** (.022)	.031† (.016)	-.009 (.007)
<i>F</i> (GRM + LLM = 0)	.004**	.355	.026*	.004**	.115
<i>N</i>	1207	1190	1225	892	863
Predictive Outcomes					
Panel C: Separate Predictors					
Human Supervisor	.003 (.021)	.008 (.045)	-.047 (.039)	.024** (.008)	-.002 (.010)
GRM Theta	.019 (.025)	.070† (.061)	-.041 (.046)	.026* (.011)	.000 (.012)
LLM Index	.004 (.022)	.007 (.040)	.064† (.037)	-.017† (.011)	.001 (.008)
χ^2 : H=G	1.90	3.97*	0.07	0.15	0.13
χ^2 : H=L	0.00	0.00	5.64*	14.05***	0.09
χ^2 : G=L	0.28	1.32	5.61*	20.18***	0.01
<i>N</i>	839	678	659	786	662
Panel D: GRM + LLM Joint Model					

GRM Theta	.019 (.026)	.070 (.062)	-.041 (.046)	.026* (.011)	.000 (.012)
LLM Index	.005 (.018)	.007 (.036)	.050 (.036)	-.000 (.011)	.003 (.006)
F (GRM + LLM = 0)	.756	.483	.056†	.000***	.991
N	826	666	646	775	650

Note. Standardized coefficients reported. Year FE + District FE + Lean Controls. Re-estimation using *7 predictor variants (listwise deletion of item-level missing data) instead of primary *miss variants (used records with any number of element scores).
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table A9

LLM Configuration Correlations and Primary vs. Robustness Predictor Correlations

Panel A: LLM Configuration Correlations				
(1) Zero-shot	—			
(2) Few-shot	.908**	—		
(3) Few-shot WPYAE	.852**	.874**	—	
(4) Few-shot WPEAY	.839**	.856**	.910**	—
<i>N</i>	1,385	1,385	1,385	1,385

Panel B: Primary (*miss) vs. Robustness (*7) Predictor Correlations		
Predictor	<i>r</i> (*miss, *7)	<i>N</i>
Human Supervisor	1.000**	1,238
GRM Theta	1.000**	1,238
LLM Index	.991**	1,385

Note. Panel A reports Pearson correlations among the four retained GPT-4o-mini scoring configurations on the primary (*miss) variables in the regression-eligible sample ($N = 1,385$). The *miss coding assigns the mean supervisor rating to leaders with missing evaluations in a given year; no minimum evaluation threshold is applied. The retained LLM composite is the z-standardized mean of these four configurations. Panel B reports the bivariate correlation between the primary (*miss) and robustness (*7) versions of each predictor. The *7 coding requires a minimum of seven scored evaluations per leader-year; observations not meeting this threshold are excluded ($N = 1,238$ for Human and GRM; $N = 1,385$ for LLM, which has no *7-induced missingness beyond the *miss baseline). For Human and GRM, $r = 1.000$ indicates perfect rank-order correspondence within the *7-eligible subsample: the 7-evaluation restriction reduces sample size but does not change the ordinal scoring. † $p < .10$. * $p < .05$. ** $p < .01$.

Table A10

Text Feature Reference: Twelve Surface Features, AES Category, Definition, and EFA Factor Assignment.

Feature Name	AES Category	Brief Definition	Factor
tf_word_count	Length-based	Total number of words in the BIP response	Volume
tf_sentence_count	Length-based	Total number of sentences	Volume
tf_avg_words_per_sentence	Length-based	Mean sentence length in words	Sophistication
tf_unique_word_count	Length-based	Count of distinct word types (stopword-filtered)	Volume
tf_ttr	Word-based	Type-token ratio: unique words divided by total words (stopword-filtered)	Volume*
tf_avg_word_length	Word-based	Mean word length in characters	Sophistication
tf_content_word_ratio	Word-based	Proportion of content-bearing words (nouns, verbs, adjectives, adverbs)	Sophistication
tf_flesch_kincaid	Readability	Flesch reading ease: sentence length and syllable count (Kincaid et al., 1975)	None†
tf_avg_parse_depth	Syntactic	Mean dependency tree depth per token (Ke & Ng, 2019)	Sophistication*
tf_overlap_count	Prompt-relevant	Raw count of rubric vocabulary words in response (stopword-filtered)	Rubric Overlap*
tf_overlap_prop_rubric	Prompt-relevant	Proportion of rubric vocabulary words in response (stopword-filtered)	Rubric Overlap*

Feature Name	AES Category	Brief Definition	Factor
tf_jaccard	Prompt-relevant	Jaccard similarity between response and rubric vocabulary (stopword-filtered)	Rubric Overlap

Note. AES categories follow the surface-feature taxonomy of Ke and Ng (2019); these five categories (Length-based, Word-based, Readability, Syntactic, Prompt-relevant) map onto the four broader dimensions described in the main text Methods section (production/length, lexical sophistication, readability and syntactic complexity, and topic-content alignment), with the Readability and Syntactic categories here jointly composing the readability-and-syntactic-complexity dimension. Factor column reports each feature's dominant EFA factor (largest absolute rotated loading); full numeric loadings for all three factors are in Table A11. * denotes a feature that also cross-loads at $|\lambda| \geq .30$ or above on a second factor (see Table A11 for values). † Flesch-Kincaid does not load on any factor (uniqueness = .97; see Table A11). Feature names prefixed tf_ correspond to variable names in the analytic dataset. TTR, unique word count, and all three Prompt-relevant features were computed on stopword-filtered text. Content word ratio derived from part-of-speech tags using spaCy. tf_flesch_kincaid: Flesch reading ease (Kincaid et al., 1975); higher values indicate simpler text. tf_avg_parse_depth: mean dependency tree depth per token (Ke & Ng, 2019).

Table A11

Principal-Factor EFA of Text Features: Rotated Loadings and Correlations with Predictor Scores

Panel A: Varimax-Rotated Factor Loadings

Text Feature	Factor 1 Volume	Factor 2 Sophistication	Factor 3 Rubric Overlap	Uniqueness
Word count	0.969			0.049
Sentence count	0.912			0.158
Avg words/sentence		0.485		0.669
Overlap count	0.673		0.713	0.022
Overlap proportion	0.687		0.700	0.024
Jaccard similarity			0.875	0.184
Unique word count	0.961			0.034
Type-token ratio	−0.560	0.502		0.418
Avg word length		0.848		0.217
Content word ratio		0.746		0.422
Avg parse depth	0.377	0.606	0.379	0.347
Flesch-Kincaid				0.969

Panel B: Correlations Among Factor Scores and Predictor Scores

	F1	F2	F3	H	G	L
Factor 1 (Volume)	1.000					
Factor 2 (Sophistication)	0.012	1.000				
Factor 3 (Rubric Overlap)	0.022	0.031	1.000			
Human (H)	0.098	−0.029	0.069	1.000		
GRM (G)	0.067	−0.039	0.035	0.924	1.000	
LLM (L)	0.548	0.420	0.528	0.116	0.074	1.000

Note. $N = 1,385$ (regression-eligible sample). Panel A reports varimax-rotated factor loadings from principal-factor EFA on 12 standardized (within-year) text features extracted from leaders' BIP-POs. Three factors were retained. Parallel analysis (paran, 500 iterations) returned a missing retained-factor count for this sample. A scree ("kink") criterion supports the three-factor solution: eigenvalue drops are large through the third factor (5.325 to 2.220 to 0.942, drops of 3.105 and 1.278) and flatten out afterward into a markedly shallower, comparatively stable rate of decline (0.593, 0.314, 0.032 for the fourth through sixth factors, drops of 0.349, 0.279, and 0.282), before falling

toward zero and turning negative from the seventh factor on. Loadings $< .30$ are suppressed. Factor 1 (Volume) loads on word count, sentence count, unique words, and overlap count/proportion (see Table A10 for definitions). Factor 2 (Sophistication) loads on average word length, content word ratio, and parse depth (see Table A10). Factor 3 (Rubric Overlap) loads on Jaccard similarity and overlap count/proportion (see Table A10). Four of the twelve features cross-load on more than one factor at $|\lambda| \geq .30$ or above rather than loading uniquely on the factor named above: overlap count and overlap proportion also load on Factor 3, as listed; type-token ratio, not named above, loads $-.560$ on Factor 1 and $.502$ on Factor 2; and average parse depth loads $.377$ on Factor 1 and $.379$ on Factor 3 in addition to its $.606$ on Factor 2, a comparable rather than a clearly dominant loading. Flesch-Kincaid readability (see Table A10) does not load on any factor (uniqueness = $.97$; its individual-item KMO = $.231$ is the weakest of the twelve, against an overall sample KMO of $.79$). Panel B reports Pearson correlations among within-year standardized factor scores and standardized predictor scores (H = Human supervisor, G = GRM theta, L = LLM index). Text factors correlate $r = .42-.55$ with L but only $r = .03-.10$ with H and G, confirming that text-surface features are substantially embedded in LLM scores but nearly orthogonal to supervisor-based scores.

Table A12*Bivariate Correlations Among Predictors and Between Predictors and Outcomes*

	Human	GRM	LLM	N
Panel A: Predictor Intercorrelations				
Human Supervisor Rating	—			1,385
GRM Theta Score	.924**	—		1,385
LLM Index	.116**	.074**	—	1,385
Panel B: Concurrent Outcomes (Year t)				
Achievement Index	.218**	.244**	-.173**	1,385
VAM Index	-.025	-.004	.019	1,348
Staffing Stability Index	.014	.017	-.014	1,385
Teacher Survey	.174**	.168**	.041	1,014
Student Survey	.066*	.069*	-.063*	978
Panel C: Predictive Outcomes (Year t+1)				
Future Achievement Index	.235**	.254**	-.154**	935
Future VAM Index	.004	.040	-.011	746
Future Stability Index	.120**	.123**	-.039	723
Future Teacher Survey	.147**	.148**	-.015	878
Future Student Survey	.088*	.082*	.017	727

Note. Panel A reports pairwise Pearson correlations among the three predictor indices in the regression-eligible sample ($N = 1,385$). Panels B and C report bivariate correlations between each predictor and each outcome, computed on outcome-specific maximal samples (listwise deletion within the regression-eligible sample). N varies across outcomes because teacher surveys, student surveys, and value-added measures are available at different rates. All predictor indices are standardized within year. Future outcomes (Panel C) are measured at $t+1$ with skip-2020 logic. $^{\dagger} p < .10$. $* p < .05$. $** p < .01$.

Online Appendix B

This appendix presents two substantive extensions of the main-text analyses and documents the multiple-hypothesis-test (MHT) families applied throughout the paper. Table B1 extends the heterogeneity analysis to binary moderators (school level and rurality), complementing the tercile-based analyses in the main-text Table 4. Table B2 presents the residualized LLM concurrent validity analysis, mirroring the right-hand-side text-feature controls in the main-text Table 6.

Findings

Table B1 reports heterogeneity analyses using two binary moderators: school level (high school versus non-high school) and rurality (rural versus non-rural). These analyses complement the FRPL-tercile and race-tercile moderators in Table 4 in the main text. None of the 30 binary-moderator interaction tests reaches nominal $p < .10$, and all MHT-adjusted p -values exceed conventional thresholds, indicating essentially null heterogeneity across school level and rurality.

Table B2 presents concurrent validity estimates using the residualized LLM score, constructed by regressing the LLM composite on the three text factors (Volume, Sophistication, Rubric Overlap) and retaining the residual. The first-stage R-squared of .724 confirms that text-surface features dominate the raw LLM composite; the residual retains a standard deviation of only 0.498, roughly half the original scale. Panel A reports separate-predictor regressions for each measure. The LLM-Residual produces a marginally significant coefficient for VAM ($\beta = .110, p < .10$) and an imprecisely estimated point estimate of .069 for staffing stability, though both are larger in magnitude than the corresponding original LLM coefficients (.022 and .060, respectively; see Panel B). By contrast, the teacher survey association, which was marginally significant for the original LLM ($\beta = .032, p < .10$), falls to essentially zero after residualization

($\beta = -.004$), suggesting that the original LLM's teacher survey validity was largely carried by text-surface features. Panel B summarizes the comparison directly: residualization amplifies the VAM and achievement point estimates while eliminating the teacher survey association, a pattern consistent with the interpretation that the original LLM composite's validity reflects a mixture of text-surface correlates and evaluative signal, though the right-hand-side controls approach in Table 6 provides a complementary test of this interpretation.

Panel C formalizes these comparisons with cross-equation Wald tests. The Human and GRM measures differ significantly on achievement ($\chi^2 = 5.42, p < .05$) and VAM ($\chi^2 = 9.06, p < .01$), reflecting the GRM's stronger association with student outcome measures. The LLM-Residual diverges from the Human score on staffing stability ($\chi^2 = 4.89, p < .05$) and teacher survey ($\chi^2 = 5.03, p < .05$), and diverges from the GRM on the same two outcomes ($\chi^2 = 6.72, p < .01$ and $\chi^2 = 8.14, p < .01$, respectively). These patterns indicate that the surface text-orthogonal component of LLM scores carries validity information for student outcomes (VAM) while losing its association with teacher-reported perceptions, a profile consistent with, though not exclusively established by, the interpretation that the LLM captures substantive evaluative signal beyond surface text characteristics; the right-hand-side controls approach in Table 6 provides a complementary test.

Multiple Hypothesis Test Families

Families are defined within each table and are never pooled across tables; individual coefficients and cross-equation Wald tests are corrected as separate families. The heterogeneity families comprise 30 tests each (2 moderators $\times$ 5 outcomes $\times$ 3 predictors), one for the tercile moderators reported in Table 4 and one for the binary moderators reported in Online Appendix B. The cross-equation Wald families for the validity tables comprise 15 tests each (3 predictor

pairs $\times$ 5 outcomes). The structural-bias table reports thirteen covariates across six models, so its families are larger: 87 coefficients in Panel A and 47 cross-equation Wald tests in Panel B. Robustness and sensitivity tables are excluded from correction. The procedure is conservative in this setting because the dependence induced by seemingly unrelated estimation reduces the effective number of independent tests below the nominal family size. Tables 2, 3, and 4 report cross-method Wald tests, and Tables B1 and B2 report the corresponding appendix tables, alongside Benjamini-Hochberg false-discovery-rate-adjusted p -values at $q = .10$ within each test family. Of the six nominally significant cross-method Wald tests in Table 2, five survive BH correction at $q = .10$; the lone exception is $H = L$ on the student survey.

Table B1

Heterogeneous Associations by School Level and Rurality: Total Predictor Slopes at Each Moderator Level

	Achievement	VAM	Stability	Tch Survey	Std Survey
Panel A: High School vs. Non-High School					
Human Supervisor (H)					
Non-high school	.016[−.004, .036]	.014[−.043, .071]	−.013[−.054, .028]	.032*[.011, .052]	.007[−.002, .015]
High school	.031[−.031, .093]	−.066[−.213, .081]	−.063[−.174, .048]	.056*[.018, .093]	.004[−.012, .019]
GRM Theta (G)					
Non-high school	.029*[.003, .055]	.057[−.008, .123]	−.023[−.085, .039]	.033*[.006, .060]	.010[−.001, .022]
High school	.062[−.004, .128]	−.011[−.190, .169]	−.071[−.204, .062]	.076**[.036, .116]	.009[−.012, .030]
LLM Index (L)					
Non-high school	.010[−.011, .031]	.014[−.037, .065]	.077**[.031, .123]	.041*[.013, .068]	−.011*[−.021, −.001]
High school	.002[−.066, .070]	.046[−.073, .165]	.005[−.073, .083]	.008[−.035, .051]	.002[−.018, .022]
<i>N</i>	1,348	1,348	1,385	1,014	978
Panel B: Rurality (Urban/Suburban vs. Rural)					
Human Supervisor (H)					
Urban/Suburban	.032*[.008, .056]	−.054[−.140, .032]	−.066[−.142, .011]	.045**[.029, .062]	.003[−.002, .009]
Rural	.012[−.021, .045]	.025[−.050, .101]	.001[−.038, .039]	.034*[.005, .062]	.008[−.003, .019]
GRM Theta (G)					
Urban/Suburban	.063**[.029, .097]	.024[−.104, .152]	−.087[−.203, .028]	.041*[.002, .080]	.009[−.003, .021]
Rural	.019[−.015, .053]	.049[−.048, .146]	.003[−.059, .065]	.048**[.018, .079]	.011[−.004, .025]
LLM Index (L)					
Urban/Suburban	.024[−.006, .054]	.012[−.051, .075]	.054*[.009, .099]	.054*[.005, .102]	−.004[−.022, .014]
Rural	−.014[−.051, .023]	.035[−.042, .112]	.068[−.004, .140]	.009[−.021, .039]	−.011[−.029, .008]
<i>N</i>	1,348	1,348	1,385	1,014	978

Note. Entries are total predictor slopes (β) at each moderator level with 90% confidence intervals. Panel A reports concurrent outcomes for high school vs. non-high school; Panel B reports concurrent outcomes for urban/suburban vs. rural. District and year fixed effects with lean controls. Lagged achievement included for achievement outcomes only. Sample: regression-eligible principals; Ns reflect the full interaction-model sample (outcome-specific). Stars denote statistical significance of individual predictor slopes at each moderator level, not significance of the interaction test (i.e., whether slopes differ across moderator levels). † $p < .10$. * $p < .05$. ** $p < .01$.

Table B2

Concurrent Validity With Residualized LLM Scores: Text-Surface Variance Removed

Panel A: Separate Predictor Models — District + Year FE, Lean Controls

	Achievement	VAM	Stability	Teacher Survey	Student Survey
Human Supervisor (H)	.020 (.013)	-.011 (.042)	-.029 (.024)	.039** (.012)	.006 (.004)
GRM Theta (G)	.038* (.015)	.038 (.048)	-.037 (.036)	.045** (.015)	.010† (.006)
LLM-Residual ($\hat{L}$)	.029 (.024)	.110† (.057)	.069 (.047)	-.004 (.021)	.008 (.014)
<i>N</i>	1,354	1,335	1,374	997	965

Panel B: Comparison — Original LLM vs. LLM-Residual

	Ach	VAM	Stabl	TchSvy	StuSvy
Original LLM (β)	.008	.022	.060*	.032†	-.007
LLM-Residual (β)	.029	.110†	.069	-.004	.008

Panel C: Cross-Equation Wald Tests

	Achievement	VAM	Stability	Teacher Survey	Student Survey
χ^2 : H = G	5.42*	9.06**	0.30	0.86	1.53
χ^2 : H = $\hat{L}$	0.12	2.13	4.89*	5.03*	0.04
χ^2 : G = $\hat{L}$	0.09	0.68	6.72**	8.14**	0.02

Note. Entries are standardized regression coefficients. The LLM-Residual ($\hat{L}$) is the residual from a first-stage regression of the LLM index on the three EFA text-feature factor scores (Volume, Sophistication, Rubric Overlap; see Table A11), estimated on the regression-eligible sample ($N = 1,385$). First-stage $R^2 = .724$; the residual has mean = 0, $SD = 0.498$, and correlations with all three factor scores of exactly 0.000 by construction. Human and GRM models are identical to Table 2 Panel A. Standard errors (clustered by district) in parentheses. Wald χ^2 tests compare cross-equation coefficient equality via *suest*. District and year fixed effects and lean controls (% FRPL, % Non-White, high school, rurality, N teachers). Lagged achievement included for the Achievement outcome only. † $p < .10$. * $p < .05$. ** $p < .01$.

Online Appendix C: Robustness and Sensitivity Battery

This appendix contains the specification and measurement robustness battery referenced in the Findings section. The tables below are organized into four thematic groups: specification robustness (C1 and C2), sampling robustness (C3 and C4), measurement robustness (C5 and C6), and structural-bias robustness (C7).

Specification Robustness

Tables C1 and C2 re-estimate all concurrent and predictive models under alternative specifications: Year FE + Lean Controls (Panels A–B) and District + Year FE with Full Controls including staff demographics (Panels C–D). The preferred District + Year FE with Lean Controls specification appears in the main-text Tables 2 and 3. We include the Year FE panels as a descriptive baseline. They pool within- and between-district variation and absorb none of the district-level sorting, resource, and policy differences that plausibly drive both scores and outcomes, so we report them to show where the district-fixed-effects estimates depart from an unadjusted benchmark. Substantive inference rests on the preferred District FE + Year FE + Lean Controls specification.

For concurrent validity (Table C1), the core findings replicate across specifications. The LLM stability signal strengthens across the specification gradient: the LLM coefficient on staffing stability is positive and nonsignificant in Year FE only ($\beta = .051$, ns), significant under the preferred specification ($\beta = .060$, $p < .05$; Table 2), and remains significant under full controls ($\beta = .063$, $p < .05$). Human and GRM associations with teacher survey are robust across all specifications, and the GRM achievement advantage over Human persists (χ^2 tests significant in both Year FE and Full Controls panels). The joint F for GRM + LLM reaches significance for teacher survey across all specifications ($p < .01$ in all panels).

For predictive validity (Table C2), the predictive teacher-survey signal is the only predictive association that survives the addition of district fixed effects: Human and GRM predict the future teacher survey under Year FE ($\beta = .031$ and $.038$, both $p < .01$) and under the preferred specification ($\beta = .024$, $p < .01$; $\beta = .025$, $p < .05$; Table 3). Under the Full Controls model, the coefficients attenuate to $\beta = .011$ and $.013$ and are no longer significant, so we treat the predictive teacher-survey result as holding in sign, but sensitive to the staff demographic covariates. The predictive null for LLM scores is consistent across specifications. The GRM stability signal under Year FE only ($\beta = .084$, $p < .01$) attenuates under the preferred specification and under the Full Controls model, consistent with between-district confounding.

Sampling Robustness

Tables C3 and C4 assess sensitivity to sample composition. Table C3 restricts estimation to a common five-outcome sample in which all observations have non-missing values for every predictor and outcome. Table C4 reports predictor score distributions by school context tercile. Online Appendix A (Tables A2 through A6) reports IPW sensitivity analyses, single-principal replication, and COVID-era skip-logic tests, integrated with the sample concerns they address.

The common-sample restriction (Table C3; $N = 462$ concurrent, $N = 216$ – 219 predictive) reduces power substantially but preserves the core concurrent validity patterns: the LLM stability signal maintains sign and magnitude ($\beta = .074$, ns in Panel A; $\beta = .082$ in the joint model), GRM predicts achievement ($\beta = .050$) and teacher survey ($\beta = .045$, $p < .05$), and the joint F is significant for teacher survey ($F = 5.71$, $p < .01$). The single-principal replication strengthens several results (see Online Appendix A, Table A5; $N = 519$ – 867 concurrent).

Table C4 reports predictor score distributions stratified by school context. Human and GRM scores show significantly greater variance in high-poverty schools (SDs of 1.06 and 0.86 in T3 vs. 0.76 and 0.68 in T1; Levene W50 $p < .001$ for both), ruling out range restriction as a mechanical explanation for differential slope attenuation in the main-text Table 4. LLM scores show the opposite pattern: variance compresses in high-poverty schools ($SD = 0.86$ in T3 vs. 1.03 in T1) while means increase, consistent with LLM scores rewarding longer documents produced in higher-poverty settings. Across racial composition terciles, the picture differs: GRM and LLM variances are statistically indistinguishable (Levene W50 $p = .627$ and $.623$), while the Human variance difference is marginal (W50 = 2.38, $p = .093$) and does not reach conventional significance.

Measurement Robustness

Tables C5 and C6 probe sensitivity to measurement decisions. Table C5 replaces the unshrunk school value-added (VAM) estimates used in the main text with Empirical Bayes (EB) shrinkage estimates. Table C6 decomposes the staffing stability index into its teacher-retention and principal-retention components. Online Appendix A (Tables A6 and A8) reports the COVID-era skip-logic sensitivity and seven-element predictor variant, integrating them with the sample and measurement concerns they address.

The school-value-added null in the main-text Tables 2 and 3 holds under EB shrinkage (Table C5): no scoring method significantly predicts either concurrent or predictive EB-VAM, confirming that the VAM null is substantive rather than a measurement artifact from noisy small-school estimates. The GRM advantage over Human on concurrent VAM detected via Wald test ($\chi^2 = 14.06$, $p < .01$) reflects differential sensitivity to EB shrinkage rather than a substantive validity difference, as neither individual coefficient reaches significance.

The stability index decomposition (Table C6) reveals that the LLM stability advantage operates through teacher retention ($\beta = .088, p < .001$) rather than school leader retention ($\beta = .032, ns$). The LLM teacher-retention signal strengthens in the joint model ($\beta = .094, p < .001$) and persists predictively ($\beta = .105, p < .05$). Human and GRM scores are null for both retention components, and the cross-method divergence on teacher retention is significant ($H = L \chi^2 = 8.12, p < .01$; $G = L \chi^2 = 4.47, p < .05$).

Structural-Bias Robustness

Table C7 re-estimates the main-text structural-bias models (Table 5, columns 1, 3, and 5) with district fixed effects added to the same Full Controls specification. The main-text models omit district fixed effects by design to preserve the between-district variation in school demographics central to the structural-bias question; this table reports the within-district counterpart, so the two together decompose each association into its between- and within-district parts. It is also the specification closest to Grissom, Blissett, and Mitani (2018), whose analogous analysis of supervisor ratings includes district fixed effects. The human-based poverty associations do not survive the restriction (Human $-.763$ to $.059$; GRM $-.812$ to $.135$, both nonsignificant), while the LLM association is nearly unchanged ($.648$ to $.626$). The school-size divergence strengthens within district, and the association between scores and the proportion of male administrators holds for both human-based methods. Within-district identification for school characteristics that rarely change—rural status and school level in particular—rests on a small number of switching observations, so those estimates are fragile and not interpreted.

Table C1

Concurrent Validity Sensitivity: Alternative Specifications

Panel A: Year FE Only, Lean Controls					
	Achievement	VAM	Stability	Teacher Survey	Student Survey
Human Supervisor (H)	.043 (.029)	−.015 (.036)	.014 (.024)	.039** (.011)	.008 (.007)
GRM Theta (G)	.060 (.038)	.008 (.046)	.024 (.033)	.046** (.013)	.008 (.009)
LLM Index (L)	−.017 (.023)	−.006 (.032)	.051 (.032)	.031* (.015)	.002 (.007)
χ^2 : H = G	1.70	1.96	0.52	1.42	0.00
χ^2 : H = L	4.95*	0.05	0.93	0.23	0.36
χ^2 : G = L	5.62*	0.08	0.39	0.69	0.23
<i>N</i>	1,385	1,348	1,385	1,014	978
Panel B: GRM + LLM Joint Model —Year FE Only, Lean Controls					
GRM Theta (G)	.064† (.037)	.009 (.045)	.016 (.034)	.042** (.013)	.007 (.009)
LLM Index (L)	−.024 (.022)	−.007 (.032)	.049 (.032)	.026† (.015)	.002 (.007)
<i>F</i> (GRM + LLM = 0)	3.01†	0.05	1.43	7.09**	0.39
<i>N</i>	1,385	1,348	1,385	1,014	978
Panel C: District + Year FE, Full Controls					
Human Supervisor (H)	.015 (.012)	−.017 (.046)	−.010 (.020)	.029** (.010)	.007 (.004)
GRM Theta (G)	.034*	.033	−.021	.035**	.011†

	(.015)	(.051)	(.033)	(.013)	(.006)
LLM Index (L)	.009	.029	.063*	.026†	−.008
	(.016)	(.028)	(.027)	(.015)	(.006)
χ^2 : H = G	5.94*	9.79**	0.50	0.71	1.33
χ^2 : H = L	0.10	0.84	7.34**	0.05	4.66*
χ^2 : G = L	1.42	0.01	7.19**	0.35	5.73*
<i>N</i>	1,354	1,335	1,354	981	954
Panel D: GRM + LLM Joint Model —District + Year FE, Full Controls					
GRM Theta (G)	.033*	.027	−.037	.029*	.013†
	(.015)	(.053)	(.032)	(.012)	(.006)
LLM Index (L)	.005	.025	.068*	.021	−.009
	(.017)	(.029)	(.027)	(.015)	(.006)
<i>F</i> (GRM + LLM = 0)	2.70†	0.66	4.48*	3.96*	2.82†
<i>N</i>	1,354	1,335	1,354	981	954

Note. Panels A–B report the Year FE specification (year fixed effects only, lean controls), included as a descriptive baseline. Panels C–D report Full Controls (district and year fixed effects plus staff demographic covariates). See Table 2 for District + Year FE with lean controls, the preferred specification. Entries are standardized regression coefficients. Panels A and C report separate predictor models estimated via *suest*; Panels B and D report joint *reghdfe* models with GRM and LLM entered simultaneously. Standard errors (clustered by district) in parentheses. Wald χ^2 tests compare cross-equation coefficient equality via *suest*. F-test is the joint significance of GRM and LLM in the combined Panel B/D model. Lagged achievement included for achievement outcome only (Specs A, B). † $p < .10$. * $p < .05$. ** $p < .01$.

Table C2

Predictive Validity Sensitivity: Alternative Specifications

Panel A: Year FE Only, Lean Controls					
	Achievement (t+1)	VAM (t+1)	Stability (t+1)	Teacher Survey (t+1)	Student Survey (t+1)
Human Supervisor (H)	.052 (.036)	.005 (.040)	.065* (.030)	.031** (.011)	.011 (.009)
GRM Theta (G)	.067 (.047)	.046 (.051)	.084** (.032)	.038** (.014)	.011 (.012)
LLM Index (L)	−.016 (.023)	−.019 (.032)	.039 (.050)	.015 (.012)	.014† (.008)
χ^2 : H = G	0.91	4.22*	2.31	1.38	0.00
χ^2 : H = L	3.73†	0.26	0.20	1.20	0.07
χ^2 : G = L	3.36†	1.27	0.60	2.10	0.04
<i>N</i>	935	746	723	878	727
Panel B: GRM + LLM Joint Model —Year FE Only, Lean Controls					
GRM Theta (G)	.071 (.046)	.049 (.052)	.079* (.034)	.037* (.014)	.009 (.012)
LLM Index (L)	−.025 (.023)	−.024 (.033)	.032 (.052)	.010 (.011)	.013 (.008)
<i>F</i> (GRM + LLM = 0)	1.76	0.64	3.54*	4.09*	1.70
<i>N</i>	935	746	723	878	727
Panel C: District + Year FE, Full Controls					
Human Supervisor (H)	−.001 (.022)	.031 (.046)	−.001 (.032)	.011 (.011)	.002 (.007)
GRM Theta (G)	.013	.082	−.003	.013	.003

	(.027)	(.060)	(.036)	(.015)	(.009)
LLM Index (L)	−.002	.033	.057†	−.006	.003
	(.018)	(.040)	(.034)	(.010)	(.006)
χ^2 : H = G	1.60	3.13†	0.01	0.15	0.01
χ^2 : H = L	0.00	0.00	1.74	1.59	0.02
χ^2 : G = L	0.22	0.79	2.82†	1.80	0.00
<i>N</i>	908	733	703	856	709
Panel D: GRM + LLM Joint Model —District + Year FE, Full Controls					
GRM Theta (G)	.013	.076	−.018	.016	.002
	(.028)	(.060)	(.035)	(.015)	(.009)
LLM Index (L)	−.004	.023	.059†	−.009	.003
	(.020)	(.040)	(.033)	(.010)	(.006)
<i>F</i> (GRM + LLM = 0)	0.12	0.90	1.85	0.95	0.14
<i>N</i>	908	733	703	856	709

Note. Panels A–B report the Year FE specification (year fixed effects only, lean controls), included as a descriptive baseline. Panels C–D report Full Controls (district and year fixed effects plus staff demographic covariates). See Table 3 for District + Year FE with lean controls, the preferred specification. Entries are standardized regression coefficients. Panels A and C report separate predictor models estimated via *suest*; Panels B and D report joint *reghdfe* models with GRM and LLM entered simultaneously. Standard errors (clustered by district) in parentheses. Wald χ^2 tests compare cross-equation coefficient equality via *suest*. F-test is the joint significance of GRM and LLM in the combined Panel B/D model. Future outcomes measured at t+1; standardized against outcome-year distributions. Lagged achievement included for Achievement outcome only (Specs A, B); VAM outcomes always exclude lagged achievement. † $p < .10$. * $p < .05$. ** $p < .01$.

Table C3

Common-Sample Robustness: Concurrent and Predictive Validity

	Achievement	VAM	Stability	Teacher Survey	Student Survey
Concurrent Outcomes					
Panel A: Separate Predictors					
Human Supervisor (H)	.001 (.033)	.051 (.100)	−.067* (.033)	.038** (.011)	.011 (.007)
GRM Theta (G)	.050 (.044)	.193 (.121)	−.067 (.042)	.045* (.017)	.018 (.012)
LLM Index (L)	.005 (.024)	−.010 (.038)	.074 (.071)	.032* (.014)	−.019 (.012)
χ^2 : H = G	11.29**	22.32**	0.00	0.78	0.95
χ^2 : H = L	0.01	0.34	5.71*	0.11	5.72*
χ^2 : G = L	0.71	2.64	4.11*	0.49	5.36*
<i>N</i>	462	462	462	462	462
Panel B: GRM + LLM Joint Model					
GRM Theta (G)	.050 (.046)	.198 (.122)	−.080† (.042)	.041* (.016)	.021† (.012)
LLM Index (L)	.000 (.026)	−.030 (.041)	.082 (.074)	.027* (.013)	−.021† (.012)
Joint <i>F</i> (GRM + LLM)	0.680	1.430	2.353	5.708**	2.509†
<i>N</i>	462	462	462	462	462
Predictive Outcomes					
Panel A: Separate Predictors					
Human Supervisor (H)	.085** (.027)	.043 (.066)	.056 (.096)	.044** (.015)	.020 (.012)
GRM Theta (G)	.094** (.032)	.120† (.062)	.004 (.101)	.063* (.025)	.024† (.012)
LLM Index (L)	.007 (.026)	−.013 (.051)	.071 (.065)	.036 (.025)	.001 (.012)
χ^2 : H = G	0.42	9.44**	1.70	2.14	0.11
χ^2 : H = L	5.29*	1.33	0.02	0.07	1.33
χ^2 : G = L	4.74*	7.07**	0.36	0.59	2.48
<i>N</i>	216	219	219	219	219
Panel B: GRM + LLM Joint Model					

GRM Theta (G)	.096** (.033)	.127* (.057)	−.011 (.105)	.058* (.026)	.025* (.012)
LLM Index (L)	−.006 (.027)	−.032 (.051)	.073 (.067)	.027 (.025)	−.003 (.012)
Joint F (GRM + LLM)	4.410*	3.450*	0.602	5.489**	2.079
N	216	219	219	219	219

Note. Panel A estimated via *suest*; cross-predictor equality tests via Wald χ^2 ($H = G$, $H = L$, $G = L$). Panel B estimated via *reghdfe* with district and year fixed effects; lean controls; lagged achievement index for achievement outcome only. Common-sample restriction: only observations with non-missing values for all predictors (Human, GRM, LLM) included simultaneously; $N = 462$ concurrent, $N = 216$ – 219 predictive. Common-sample N s are substantially smaller than the full-sample analysis (Tables 2–3) due to the simultaneous non-missing requirement across all three scoring methods. † $p < .10$. * $p < .05$. ** $p < .01$.

Table C4: Predictor Score Distributions by School Context Tercile**Panel A: By Poverty Tercile**

Predictor	Tercile	Mean	<i>SD</i>	Min	Max	<i>N</i>
Human Supervisor	T1 (Low)	0.303	0.757	-6.072	0.754	462
	T2 (Mid)	-0.055	0.960	-4.221	0.754	462
	T3 (High)	-0.169	1.055	-4.423	0.754	461
GRM Theta	T1	0.303	0.682	-2.627	0.907	462
	T2	-0.088	0.846	-2.414	0.907	462
	T3	-0.196	0.862	-2.556	0.907	461
LLM Index	T1	-0.176	1.025	-3.150	1.476	462
	T2	-0.036	0.913	-3.838	1.386	462
	T3	0.230	0.858	-2.865	1.758	461

Panel B: By Racial Composition

Predictor	Tercile	Mean	<i>SD</i>	Min	Max	<i>N</i>
Human Supervisor	T1 (Low)	-0.097	1.007	-6.072	0.754	462
	T2 (Mid)	0.106	0.872	-4.192	0.754	462
	T3 (High)	0.069	0.965	-4.423	0.754	461
GRM Theta	T1	-0.147	0.813	-2.627	0.907	462
	T2	0.105	0.807	-2.121	0.907	462
	T3	0.062	0.845	-2.556	0.907	461
LLM Index	T1	-0.091	0.935	-3.838	1.476	462
	T2	-0.025	0.949	-3.150	1.498	462
	T3	0.134	0.951	-3.838	1.758	461

Panel C: Levene Variance Equality Tests

Test	Predictor	W50 Statistic	<i>p-value</i>
By Poverty Tercile	Human Supervisor	28.85	< .001
	GRM Theta	37.30	< .001
	LLM Index	6.39	.002

By Racial Composition Tercile	Human Supervisor	2.38	.093
	GRM Theta	0.47	.627
	LLM Index	0.47	.623

Note. Entries are summary statistics for standardized predictor scores within the regression-eligible sample ($N = 1,385$), stratified by school-level poverty (Panel A) and racial composition (Panel B) terciles computed within that same sample, so that the tercile definitions match the moderators used in Table 4. Panel C reports Brown-Forsythe (W50) test statistics for equality of variance across terciles. See Table 4 for heterogeneous slope estimates.

Table C5

Robustness: Empirical Bayes VAM as Alternative Value-Added Outcome

	EB-VAM (Concurrent)	EB-VAM (Predictive)
Panel A: Separate Predictors		
Human Supervisor (H)	-.024 (.039)	.027 (.053)
GRM Theta (G)	.040 (.047)	.068 (.066)
LLM Index (L)	.009 (.030)	.008 (.041)
χ^2 : H = G	14.06**	2.30
χ^2 : H = L	0.61	0.14
χ^2 : G = L	0.37	0.96
<i>N</i>	1,345	745
Panel B: GRM + LLM Joint Model		
GRM Theta (G)	.039 (.048)	.068 (.066)
LLM Index (L)	.004 (.030)	-.002 (.040)
<i>F</i> (GRM + LLM = 0)	0.38	0.55
<i>N</i>	1,332	732

Note. Standard errors (clustered by district) in parentheses. EB-VAM applies empirical Bayes shrinkage to the unshrunk VAM estimates used in Table 2 (concurrent) and Table 3 (predictive). District and year FE and lean controls. [†] $p < .10$. * $p < .05$. ** $p < .01$.

Table C6

Stability Index Decomposition: Teacher Retention vs. Principal Retention

	Teacher Retention	Principal Retention	Future Teacher Retention	Future Principal Retention
Panel A: Separate Predictor Equations				
Human Supervisor (H)	-.002 (.030)	-.056 (.047)	-.006 (.039)	-.014 (.076)
GRM Theta (G)	-.017 (.048)	-.056 (.059)	-.047 (.043)	.034 (.092)
LLM Index (L)	.088*** (.021)	.032 (.047)	.105* (.041)	.024 (.058)
χ^2 : H = G	0.47	0.00	9.41**	1.63
χ^2 : H = L	8.12**	2.06	2.98†	0.36
χ^2 : G = L	4.47*	1.73	5.08*	0.02
<i>N</i>	1,385	1,385	723	715
Panel B: Joint GRM + LLM Model				
GRM Theta (G)	-.041 (.051)	-.066 (.060)	-.076 (.051)	.029 (.088)
LLM Index (L)	.094*** (.021)	.041 (.048)	.115* (.044)	.021 (.049)
<i>F</i> (GRM + LLM = 0)	9.91***	0.84	3.48*	0.11
<i>N</i>	1,374	1,374	712	703

Note. All models include district and year fixed effects and lean controls. Standard errors clustered by district in parentheses. Wald χ^2 tests assess cross-equation predictor equality via the suest variance-covariance matrix. F-test is the joint significance of GRM and LLM. Future outcomes are measured at t+1 with skip-2020 logic. The composite Stability Index used in Tables 2 and 3 is the standardized average of Teacher Retention and Principal Retention. † $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table C7

Structural-Bias Robustness: Adding District Fixed Effects

<i>Panel A: Full Controls, With and Without District Fixed Effects</i>						
	Human (1)	Human (2)	GRM (3)	GRM (4)	LLM (5)	LLM (6)
	Year FE	+ District FE	Year FE	+ District FE	Year FE	+ District FE
% Students FRPL	−.763** (.278)	.059 (.209)	−.812** (.270)	.135 (.205)	.648† (.324)	.626† (.318)
% Students Non-White	.169 (.455)	.464 (.638)	.249 (.450)	.469 (.518)	.028 (.307)	−.266 (.306)
High School	−.214† (.115)	−.069 (.088)	−.205* (.099)	−.045 (.074)	−.185† (.111)	−.125 (.118)
Rural	−.051 (.212)	.370 (.223)	−.010 (.210)	.802** (.201)	−.021 (.190)	−1.966** (.300)
N Teachers	.006** (.002)	.002† (.001)	.006** (.002)	.002† (.001)	−.001 (.002)	−.004** (.002)
Avg. Teacher Experience	.030 (.024)	.018 (.013)	.028 (.023)	.011 (.011)	−.047* (.023)	−.025 (.021)
% Teachers Advanced Degree	−.152 (.479)	−.562† (.294)	.075 (.437)	−.239 (.255)	.268 (.433)	−.002 (.388)
% Teachers Non-White	−.011 (.727)	−.868 (.671)	.099 (.588)	−.494 (.405)	−.020 (.532)	−.101 (.382)
% Teachers Male	−.104 (.332)	.307 (.274)	−.067 (.287)	.352 (.243)	−.034 (.330)	.120 (.298)
Avg. Administrator Experience	−.006 (.008)	−.005 (.005)	.000 (.007)	−.000 (.004)	.007 (.007)	.006 (.006)
% Administrators Non-White	−.151 (.193)	−.222 (.132)	−.030 (.176)	−.115 (.109)	.081 (.222)	.034 (.178)
% Administrators Male	−.297** (.095)	−.234** (.077)	−.265** (.076)	−.216** (.060)	−.122 (.101)	−.097 (.090)
% Administrators Advanced Degree	−.038 (.089)	.007 (.086)	−.027 (.076)	−.003 (.069)	.200* (.091)	.179† (.089)
R^2	.101	.457	.134	.506	.097	.346
N	1,385	1,385	1,385	1,385	1,385	1,385

Panel B: Cross-Equation Wald Tests (District Fixed Effects Models)

	χ^2 : H = L	χ^2 : G = L	χ^2 : H = G
% Students FRPL	2.01	1.59	0.77
% Students Non-White	1.12	1.60	0.00
High School	0.21	0.43	0.36
Rural	48.25**	100.55**	19.73**
N Teachers	17.74**	17.77**	0.01
Avg. Teacher Experience	3.48†	2.48	1.19
% Teachers Advanced Degree	1.63	0.34	4.91*
% Teachers Non-White	1.21	0.59	1.43
% Teachers Male	0.33	0.59	0.16
Avg. Administrator Experience	3.01†	1.33	4.48*
% Administrators Non-White	1.58	0.58	3.11†
% Administrators Male	2.03	1.91	0.30
% Administrators Advanced Degree	2.76†	3.36†	0.10
Joint: FRPL + Race	2.27	2.16	—

Note. N = 1,385 for all models. Unstandardized OLS coefficients; the dependent variable is the standardized predictor score (Human, GRM, or LLM). Standard errors (clustered by district) in parentheses. Odd-numbered columns reproduce Table 5 columns 1, 3, and 5 (year fixed effects only) and are shown for comparison; even-numbered columns add district fixed effects to the same Full Controls specification. Panel B reports suest cross-equation Wald χ^2 tests for the district fixed effects models. Rural status and school level rarely change within a district, so the corresponding district-fixed-effects estimates are identified from a small number of switching observations; we report them for completeness and do not interpret them. † p < .10. * p < .05. ** p < .01.

Online Appendix D: LLM Prompt Excerpts and Configuration Details

This appendix provides representative excerpts from the GPT-4o-mini prompts used to generate LLM scores for BIP responses, together with a structural summary of the four prompting configurations. All four configurations share the same scoring objective: assigning a rubric-aligned score between 0 and 4 to each BIP element response, with the model returning only a numeric score and no additional commentary. The complete prompt text for all four configurations, along with full design rationale and validation results, is available in Hunter, Bowser, et al. (in preparation). The four configurations are: (1) Zero-Shot, (2) Five-Shot, (3) Few-Shot WPEAY (Within Principal, Elements, across Years), and (4) Few-Shot WPYAE (Within Principal, Year, across Elements).

System Prompt (Shared Across All Configurations)

All inference calls use GPT-4o-mini with the following system message: *You are a helpful assistant.* All outputs were extracted as raw string responses from the model’s first completion token.

Prompt Structure by Configuration

Zero-Shot

Each BIP response was scored independently using only the element-specific rubric and the response text. No demonstration examples and no prior principal scoring history were provided.

```
Task: You will be scoring a Principal on the development of the BIP based on the
      given input. The scores will be on a range from 0 to 4 and will follow a rubric
      given below. Study the text carefully and assign what you think is the correct
      score based on the rubric. You will respond with the score only.
Rubric: [<rubric>. You can assign a score of 1 and 3 if necessary].
Input: [<text>].
```

Question: Provide a score for the input based on the rubric. Do not add any other text. The numeric value of the score is all we expect. Score:

The phrase “You can assign a score of 1 and 3 if necessary” explicitly licenses odd integer values on the 0–4 scale. The instruction “Do not add any other text” suppresses justifications, which were a common failure mode in earlier testing.

Five-Shot

Identical to Zero-Shot, except each prompt was augmented with five labeled demonstration examples drawn from a fixed pool (9–11 instances per element, randomly sampled with a fixed seed). Examples mixed supervisor-scored and annotator-rated instances to avoid anchoring the model to either end of the human rating distribution.

Task: You will be scoring a Principal on the development of the BIP based on the given input. The scores will be on a range from 0 to 4 and will follow a rubric given below. I will also be including some examples for you to analyze and understand what the scores should look like for the given text. Analyze the data but do not copy paste anything in your response. Study the text carefully and assign what you think is the correct score based on the rubric and your knowledge from the examples. You will respond with the score only.

Rubric: [<rubric>. You can assign a score of 1 and 3 if necessary.

Example: [<examples>]

Input: [<text>].

Question: Provide a score for the input based on the rubric. Do not add any other text. The numeric value of the score is all we expect.

Score:

Each example was formatted as: [Example i Text: <text> Score: <score>]. Examples were embedded inside the rubric brackets, immediately following the rubric text. Five-shot was selected over eight-shot after a direct comparison: five examples produced lower RMSE and MAE against both reference score types (results available upon request).

Few-Shot WPEAY (Within Principal, Elements, across Years)

Records were grouped by the triple (PersonId, BuildingID, DistrictID) and element. All available school years for a given principal-element pair were scored in a single cumulative conversation, allowing the model to condition on the principal’s longitudinal scoring history on that criterion. The rubric was constant across all turns; only the BIP response text changed from year to year.

The conversation was initialized once per principal-element group with:

Task: You will be scoring a Principal on the development of the BIP based on the given input. The scores will be on a range from 0 to 4 and will follow a rubric given below. Study the text carefully and assign what you think is the correct score based on the rubric. You will respond with the score only, no other text necessary.

Rubric: [<rubric> You can assign a score of 1 and 3 if necessary.]

For each successive school year t , the following block was appended before each API call:

Here are some example data for you to understand the task.

Example: [<examples>]

Input: [<text_ t >].

Question: Provide a score for the input based on the rubric. Do not add any other text. The numeric value of the score is all we expect.

Score:

After each prediction, the assigned score was appended to the growing conversation with a bridging instruction: <score_ t >. Now, for the same principal, score this new text according to the new rubric.

Few-Shot WPYAE (Within Principal, Year, across Elements)

Records were grouped by the four-tuple (PersonId, BuildingID, DistrictID, SchoolYear). All rubric elements for a given principal-year were scored in a single cumulative conversation, typically seven turns (one per element), with each previously assigned score visible when the next element was evaluated.

The conversation was initiated once per principal-year group, using the same task preamble as in WPEAY. For each element i , the following block was appended:

```
Rubric: [<rubric_i> You can assign a score of 1 and 3 if necessary].  
Here are some example data for you to understand the task.  
Example: [<examples_i>]  
Input: [<text_i>]. Question: Provide a score for the input based on the rubric. Do  
not add any other text. The  
numeric value of the score is all we expect.  
Score:
```

Unlike WPEAY, the rubric changed at every turn as new elements were introduced. After each prediction, the bridging continuation $<score_i>$. Now, for the same principal, score this new text according to the new rubric. was appended, enabling the model to condition its prediction on all previously assigned element scores within that year.

Table D1: Structural Comparison of the Four GPT-4o-mini Scoring Configurations.

	Zero-Shot	Five-Shot	WPEAY	WPYAE
Grouping key	None	None	PersonId + Building + District + Element	PersonId + Building + District + Year
Fixed dimension within group	N/A	N/A	Element (rubric constant)	Year (temporal context fixed)
Varying dimension within group	N/A	N/A	School year	Rubric element
Examples per turn	0	5	5 (resampled per element)	5 (resampled per element)
Prior scores in context	No	No	Yes (prior years, same element)	Yes (prior elements, same year)
Instances treated as i.i.d.	Yes	Yes	No	No
Context grows within group	No	No	Yes	Yes

Note. WPEAY = Within Principal, Elements, across Years; WPYAE = Within Principal, Year, across Elements. All four configurations used GPT-4o-mini. The example pool for each element contained 9–11 instances, drawn from a mix of supervisor-scored and annotator-rated BIP responses, and sampled with a fixed random seed for reproducibility.