Journal of Learning Theory and Methodology
Журнал теорії та методології навчання

ISSN 2708-7581 (Online)
ISSN-L 2708-7573

Original Scientific Article

Beyond Course Averages: A Generalized Bayesian Hierarchical Framework for Course-Level Learning Evaluation

Vicente E. Montano[1ABCD] and Archie G. Reyes[1ABCD]

[1]University of Mindanao

Authors' Contribution: A – Study design; B – Data collection; C – Statistical analysis; D – Manuscript Preparation; E – Funds Collection

DOI: 10.17309/jltm.2026.7.1.04

Abstract

Background. Course-level learning assessment in higher education is commonly based on comparisons of average performance indicators, implicitly assuming independence across courses and equal reliability of estimates. When enrollments are small and uneven, such approaches produce statistically unstable estimates and exaggerate extreme values, leading to potentially misleading interpretations.

Objectives. This study aims to develop a generalizable methodological framework for applying Bayesian hierarchical modeling (BHM) to course-level learning assessment, explicitly accounting for sampling uncertainty and unequal group sizes.

Materials and Methods. A Bayesian hierarchical model was specified in which student learning outcomes were modeled at the individual level while accounting for course membership. The model decomposes total variance into within-course and between-course components and estimates course-level effects using posterior distributions. Partial pooling was applied to stabilize estimates for courses with small enrollments. An empirical illustration was conducted using anonymized data from 279 students across 22 courses.

Results. Naïve comparisons based on course averages were found to systematically exaggerate extreme outcomes under small sample conditions, resulting in unstable and potentially misleading conclusions. The application of Bayesian hierarchical modeling substantially reduced artificial extremity while preserving statistically supported between-course differences. After pooling, most course effects were not distinguishable from the program average, while a limited number of courses showed consistent deviations.

Conclusions. Bayesian hierarchical modeling provides a statistically robust alternative to descriptive aggregation and course ranking. By incorporating uncertainty and stabilizing estimates, it enables more reliable interpretation of course-level performance and supports targeted, evidence-based academic evaluation.

Keywords: bayesian hierarchical modeling, multilevel analysis, course-level assessment, small sample instability, educational measurement.

Introduction

Performance indicators are instrumental in aligning with a program's expectations, thereby facilitating course delivery strategies and assessment procedures (Anwar et al., 2012; Cabrera et al., 2001). Essential steps are taken before developing performance indicators, including the measurement of student outcomes. Typically, these are communicated to students in the course description, which outlines the course's learning objectives and the faculty's expectations (Hristov et al., 2023). Student outcomes are intended to provide general information about the focus of student learning and are generally stated in terms of outcomes rather than being measurable (Lewis, 2021). Compared to other indicators, performance indicators are concrete, measurable standards that students must demonstrate as evidence of achievement. Performance indicators are derived from course outcomes (Kennedy, 2008).

Small-enrollment and uneven-enrollment courses pose critical statistical challenges that, in turn, affect the generalizability, reliability, and validity of research results. The fundamental issue arises from a small sample size (low statistical power) and the possibility of unreliable data (Button et al., 2013). There are several critical statistical issues, but low statistical power is the most significant due to a decreased ability to detect an actual effect or difference (Type II error) (Anderson & Maxwell, 2017). The situation suggests that a significant improvement in the course, such as a new teaching method, might be deemed "not statistically significant" given a very small sample size. Results from a

JLTM
LLC OVS

Corresponding Author: Vicente Montano,
e-mail: vicente_montano@umindanao.edu.ph

small, specific student group may not accurately reflect the traits of a larger, more diverse population. Compromised external validity makes it difficult to apply the conclusions to other contexts (Mascha & Vetter, 2018).

The Bayesian Hierarchical Modeling (BHM) framework is a robust statistical technique designed to overcome the challenges of small and uneven courses, leveraging data at multiple levels (e.g., students nested within courses, courses nested within departments) to produce more stable and reliable estimates (Berry et al., 2013). The BHM method is particularly applicable in educational environments where traditional statistical methods yield unreliable results with small sample sizes (Vandendijck et al., 2016). Bayesian statistics form the basis of the BHM framework, which uses prior information and observed data to update the probability of an outcome (Moeyaert et al., 2017). Its critical feature in this context is a hierarchical structure modeled at several levels (McGlothlin & Viele, 2018). For instance, student performance is at level 1, a specific course is at level 2, which is nested within a department or institution at level 3. This facilitates information sharing across courses and departments.

Notwithstanding the vast amount of literature on learning outcomes assessment and program evaluation, there still appears to be a methodological gap in the modeling of course-level performance as it relates to hierarchical organization and differential enrollment. This study seeks to fill this methodological gap by formalizing a Bayesian hierarchical model that partitions variance, scales course-level effects via partial pooling, and integrates uncertainty into the evaluation inference. The majority of course-level performance assessments at the institutional level are typically descriptive and based on course averages or threshold comparisons that are, in essence, based on assumptions of independence and equal reliability of courses, and do not typically involve uncertainty estimation or variance decomposition. Although presented in the context of an academic program, the significance of this work lies in the development of a transferable modeling paradigm for learning analytics and educational measurement.

Theoretical Framework

Higher education professionals recognize the use of various assessments, from standardized test scores as sources of information to final course grades for students (Chan, 2014). In measuring reliability, consistency of scores is pertinent. At the same time, validity refers to the support for the interpretations that justify the appropriate use of assessment data, ensuring that it measures what it intends to measure (Birenbaum, 2007). The framework for discussing reliability and validity is widely accepted as the standard for guiding practitioners in quality assurance for the development and use of assessments. The increasing importance of higher education and learning in evolving work patterns and society is broadly acknowledged (Ramezani & Mostafavi, 2025). However, sustaining and implementing educational reform presents significant challenges.

The concept of hierarchy in educational organizations is a fundamental factor in the effective functioning of organizational structures. Hierarchy plays a crucial role in defining authority and tasks across management levels, as well as in coordinating among them (Baartman et al., 2007). Accountability and task distribution are based on a structure that enables educational institutions to more easily attain strategic objectives, significantly increasing organizational efficiency (Inglis, 2008). Additionally, a hierarchical structure in educational institutions facilitates leadership and decision-making processes, enabling the implementation of a more systematic and structured assessment method (Whiting et al., 2017). A rational management approach facilitates a more efficient decision-making process.

Intra-organizational communication and innovation are negatively impacted by rigid hierarchical structures, making it challenging for faculty members to effectively communicate their assessment ideas and suggestions to senior management (Bentley et al., 2017). This constraint hinders evaluation flexibility and course outcomes. Thus, a balanced hierarchical structure in an educational organization is crucial for maintaining organized assessment and offering flexibility in evaluating learning outcomes (Birenbaum, 2007). The effectiveness of the assessment process, as well as support for faculty participation and motivation, increases in a balanced hierarchical structure (Kruger & Leuro, 2015, September).

Assessment is deliberately administered and aligned with standards and curriculum; faculty gain a deeper understanding of student progress (Whiting et al., 2017). Similar to curriculum, assessments must align with content and grade-specific requirements to evaluate students' knowledge, skills, and abilities as described in the standards (Chan, 2014). Notably, it is crucial to determine whether the curriculum aligns with the standards and whether the assessments do as well (Feiler et al., 2012). Evaluation does not focus solely on large-scale summative assessments; it also considers formative processes within classrooms and other administered assessments, all of which must be aligned with the curriculum. Alignment with the curriculum involves gathering information pertinent to the specific learning outcomes that students are engaging with.

The four-level model, highly used as an evaluation tool in educational programs, is Kirkpatrick's model of reaction, learning, behavior, and results (Smidt et al., 2009). The first level of the model is applied to formal education settings, encompassing reaction criteria such as the delivery of learning objectives, student learning design, and perceptions (Praslova, 2010). Level two is the learning criteria, which include measuring student performance through skill demonstrations and knowledge tests. Behavioral criteria are the third level of a student's transfer of skills or knowledge beyond the context in which initial learning happened. Result criteria are the final level, which describes long-term outcomes related to service to society, personal stability, and career success (Cheung et al., 2023). Kirkpatrick's first two levels primarily focused on areas in course evaluation, measuring student learning within the specified timeframe (Thörn et al., 2022). In addition to evaluating educational programs, foundational measures are used to assess student reactions to the learning experience. Student attitudes and responses to the delivery and design of learning experiences are used mainly in program evaluation and have been shown to influence learning outcomes (Nawaz et al., 2022).

The methodological basis of this research utilizes Bayesian hierarchical theory as a mathematical model

of information exchange in uncertain contexts. This shrinkage procedure is more than statistical regularization; it implements a normative guideline in evaluation theory, with estimates corresponding to the strength of evidence. Hierarchical Bayes offers a theoretically grounded approach to small-sample problems of instability by shrinking group estimates toward a population mean in proportion to their posterior uncertainty.

This approach redirects educational evaluation from exercises in ranking to uncertainty-calibrated decision-making. This informs methodological debates in learning analytics and educational measurement. It formulates course-level evaluation as probabilistic structural inference. The approach challenges descriptive aggregation and individualistic conceptions of learning achievement. Through the integration of reliability theory, variance decomposition, and Bayesian partial pooling, this research contributes to a methodological integration for learning assessment.

Materials and Methods

Research Design and Framework

Bayesian hierarchical modeling (BHM) is a statistical method used to measure variability at both the individual and group levels, considering prior data to inform analyses and quantify uncertainty in estimates (Baldwin & Fellingham, 2013). BHM is a statistical technique used to analyze data with a multi-level or nested structure, such as students within a course or data gathered from various studies in a meta-analysis (Schmid & Brown, 2000). As a key feature and mechanism, the hierarchical structure recognizes data organized into levels, facilitating parameter estimation at each layer of the hierarchy (Columb & Atkinson, 2016).

Simultaneously, the model assesses students' performance and the average performance of their academic programs. Bayes' theorem forms the basis of BHM, which updates prior beliefs or current knowledge by conditioning on new data and parameters, yielding a full probability distribution (the posterior distribution) for each parameter (Chen et al., 2014). An essential advantage of BHM is the partial pooling of groups with limited data, allowing them to borrow information from the general population distribution (Goodhue et al., 2006, January). This results in more reliable and stable estimates for smaller groups while avoiding overgeneralization from larger groups (Anderson & Maxwell, 2017). Rather than relying on a single point estimate, BHM offers a range of probable values and uncertainty for each parameter, providing a more complete perspective on the results.

In cases where different grading scales were used, categorical grades were transformed into percentage equivalents according to institutional guidelines to facilitate comparison across courses. The main outcome variable, cor_average, measures the course-level student performance index based on official institutional grade data for the 2024–2025 academic year. It captures the average student performance score for each course on a continuous scale from 0 to 100. Notably, the variable measures academic performance based on institutional grading systems rather than external standards of competence or ability. Using the standardized model, the variable is converted into z-scores to enable hierarchical analysis and interpretation of course-level effects in terms of deviations from the grand mean. In the unstandardized model, cor_average measures absolute average academic performance, with higher scores indicating better aggregate course performance.

However, several limitations apply. As a course-level aggregate variable, cor_average may obscure differences in instructional design and variability in grading practices at the instructional level. The variable does not account for within-course variability or dynamics over time. Therefore, it is considered a structural performance metric within the institutional setting rather than a direct causal indicator of instructional quality.

Data Sources and Collection

The final dataset comprises anonymized student-level performance records derived from official academic assessment data for the BS Business Administration program. A total of 279 students enrolled in the 2024–2025 academic year, across 22 courses, were included in the observations. Each observation represents a student's performance metric, aggregated at the course level, such as a correlation-based or standardized outcome measure, with course identifiers defining group membership. This study utilized anonymized secondary academic records and did not require direct contact with students or the collection of personally identifiable information. In keeping with institutional research standards for minimal-risk studies using de-identified data, formal ethics board approval was not required, and all data were handled in accordance with data protection and confidentiality standards. Courses with very small enrollments were retained initially to diagnose instability and were subsequently pooled in a refined specification that ensures statistically defensible estimation.

Model Specification

The primary model is a Bayesian random-intercept model (Ha et al., 2014) specified as

$$\gamma_{ij} = \beta_0 + \mu_j + \epsilon_{ij}$$

Where γ_{ij} denotes the outcome for student in course j, β_0 is the grand mean, $\mu \sim N(0, \tau^2)$ captures the course-level random effect, and $\epsilon_{ij} \sim N(0, \sigma^2)$ represents the course residual variation. Weakly informative priors were used for fixed effects and variance components to stabilize estimation without imposing strong assumptions (Moeyaert et al., 2017). A second, refined specification pooled minimal enrollment courses into a single group, enabling more substantial shrinkage and reducing spurious extremity in course-level estimates.

Estimation and Computation

Model estimation was performed using MCMC sampling via the stan_lmer implementation, with 4,000 to 8,000 posterior samples drawn across multiple chains (Monnahan et al., 2017). This method provides complete posterior distributions for all parameters, thereby enabling probabilistic interpretation of course effects and their

uncertainty (Bocquel et al., 2013). Computation focused on achieving stable posterior exploration (Nguyen et al., 2018), with sufficient warm-up iterations and thinning avoided to preserve an adequate sample size (Kim et al., 2017).

Model Validation and Checking

A combination of convergence diagnostics and posterior predictive checks was used to assess model adequacy. Convergence was evaluated using both the potential scale reduction factor (Stern & Sinharay, 2005) and Monte Carlo standard errors, as well as effective sample size (Koch, 2018), all of which indicated excellent convergence (R-hat < 1.01 for all parameters) (Chen et al., 2025). For model adequacy, posterior predictive distributions were examined to ensure that the key features of the observed data were accurately replicated, with a focus on the overall mean and dispersion (Gajewski et al., 2008). Comparisons of unpooled and pooled specifications were used to assess the effects of partial pooling on stability and interpretability (Feng et al., 2024).

From a Bayesian viewpoint, statistical strength is assessed through the concentration of posterior distributions and the stability of credible intervals. Courses with larger enrollment sizes have narrower intervals and higher posterior precision, while small-enrollment courses have wider intervals and stronger shrinkage toward the grand mean. This adaptive borrowing of strength partially offsets the problem of small sample sizes but does not eliminate imprecision in sparse groups.

Limitations of Methodology

A drawback of the study design is that the number of courses and students limits the precision of estimates for individual courses, particularly for small classes. While hierarchical shrinkage reduces imprecision, some posterior intervals remain wide, suggesting that there is only limited evidential support for precise comparisons. Future studies with multiple years of data across several institutions would greatly improve inferential precision and allow more robust structural inferences.

The study uses data from complete institutional records for a single year; no classical a priori power analysis was necessary, as the size of the enrollment was determined by administrative necessity rather than experimental design. In Bayesian hierarchical analysis, goodness of fit is assessed through the precision of posterior distributions, effective sample size, and interval width rather than null hypothesis–based power calculations. The total sample (N = 279 across 22 courses) is sufficient to provide information for the estimation of overall variance components, although precision for individual courses varies proportionally with enrollment size.

Results

The descriptive statistics in Table 1 indicate that student performance is highly dispersed across the BS Business Administration program. While the overall mean is 72.79 and the median is noticeably higher at 79.44, the large standard deviation of 22.32 and the full range of 0 to 100 depict strong heterogeneity in outcomes. At the course level, this noticeable average enrollment is reasonable (about 13 students per course), yet the courses cover a wide range from 1.69 to a perfect 100. This array suggests that raw differences in performance cannot be attributed solely to differences in student performance, but rather reflect structural differences across courses in alternative assessments and evaluation practices. These descriptive results from the observed motivation for the subsequent multilevel evaluation, due to basic aggregation, clearly mask the principal effects of the course and exacerbate instability arising from small enrollments.

Table 1. Descriptive Statistics of Student Performance

Statistic	Value
Number of Observations	279
Number of courses	22
Mean	72.79
Median	79.44
Standard Deviation	22.32
Minimum	0
Maximum	100
Average student per course	12.7
The range of the course means	1.69-100

The variance decomposition in Table 2 clearly indicates a structural imbalance in the BS Business Administration program, with nearly 69% of the total performance variability explained by between-course rather than within-course variability, resulting in a high ICC of 0.688. This implies that the program the student attends is far more important than the student. This level would not be expected in an optimally integrated academic program, indicating significant variability in the standards and performance measures across courses. The minor variability within courses suggests that, once students are grouped in the same course, their performance tends to be relatively similar, confirming the observation that course-level factors largely determine performance.

Table 2. Variance Components and Descriptive Summary
Variance Components Decomposition

Component	Variance	Percentage	ICC
Between Course	776.06	68.78	0.688
Within-Course	352.33	31.22	-
Total	497.97	100	-

Descriptive statistics in Table 1 confirm this diagnosis: although the grand mean is 72.79, the class means exhibit an extreme range, from 1.69 to 1.00, despite the average class size being only about 13 students. This combination of small enrollments with extreme means is a red flag. It maximizes course effects and inflates between-course variance, to make some courses appear "excellent" artificially and others catastrophically weak. The results taken together suggest a quality assurance problem at the program level rather than one related to isolated student underperformance and, again, call for tighter alignment of assessment criteria, stronger

moderation of grading practices, and closer monitoring of low- and high-outlier courses as a way of restoring equity and defensibility in academic outcomes.

The Bayesian random-intercept models in Table 3 show that course membership is a primary structural driver of variation in the outcome-cor_average. The overall intercept represents a baseline latent mean, while the significant, estimated course-level variance (≈1,416) relative to the residual variance (σ ≈ 18.9) suggests substantial heterogeneity across courses. Several courses, particularly 3_ENTE 222, 3_FM 211, and selected GE offerings, exhibit significant positive deviations from the grand mean, whereas 3_FME 311, 3_FME 322, and 3_FM 221 show significant adverse effects with credible intervals that exclude zero. The result suggests that these courses have systematically lower outcomes, rather than experiencing random fluctuations. This set of results aligns with the earlier variance decomposition results, as most of the variation is between courses rather than within them.

Table 3. Bayesian Random-Intercept Model (Unpooled Courses) Fixed Effect (Overall Intercept)

Parameter	Mean	SD	2.50%	97.50%
Intercept (Grand Mean)	42.14	9.49	1.82	58.14

Table 4. Course-Level Random Effect (Selected)

Course	Mean	SD	2.50%	97.50%
3_ENTE 222	44.73	20.47	7.5	86.32
3_FM 211	46.12	12.37	23.55	71.17
3_GE 5	40.83	12.76	17.07	66.07
3_GE 7	39.39	10.82	20.08	62.36
3_GE 9	38.89	19.87	0.85	79.88
3_FM 221	-24.02	14.41	-51.71	5.11
3_FME 311	-34.19	14.45	-61.55	-4.88
3_FME 322	-34.49	14.06	-60.83	-6.01

Positive values indicate above-grand mean performance, negative values indicate below-grand mean performance

Table 5. Variance Component

Component	Mean	SD	2.50%	97.50%
Residual SD (σ)	18.87	0.83	17.35	20.56
Course Level Variance	1415.68	751.9	462.54	3378.2

Table 6. Posterior Predictive Check

Statistic	Mean	SD	2.50%	97.50%
Mean_PPD	72.21	1.62	69.04	75.29

Table 7. MCMC Diagnostics (Summary)

Metric	Result
Rhat (all parameters)	≈ 1.00-1.01
Effective Sample Size	500-3900
MCSE	Small relative to posterior SD
Convergence	Satisfactory

From the point of view of model quality, diagnostics are clean: The Rhat values are all essentially 1; effective sample sizes are good; and the posterior predictive check shows that the model reproduces the observed mean very well (mean PPD ≈ 72.2, close to the empirical average). The hierarchal estimates provide information on relative location within the sample and the degree of uncertainty associated with these locations, rather than suggesting direct causal links between course organization, teaching, and policy decisions and the observed differences. The estimates presented are measures of statistical association within the observed grading pattern and are not considered for causal interpretation. The variance and random effects for the between course levels capture patterns of dispersion in the measured performance under the grading structure of the institution.

Table 8. Bayesian Random-Intercept Model with Pooled Small Courses. Fixed Effect (Overall intercept)

Parameter	Mean	SD	2.50%	50.00%	97.50%
Intercept	0.018	0.157	-0.293	0.018	0.322

Table 9. Course-Level Random Effects

Course	Mean	SD	2.50%	97.50%	Interpretation
3_FM_211	0.48	0.331	-0.135	1.171	Moderately positive, uncertain
3_GE 1	0.206	0.19	-0.167	0.58	Small positive
3_GE 11	0.136	0.289	-0.43	0.705	Weak, inconclusive
3_GE 2	-0.305	0.376	-1.077	0.404	Weak negative
3_GE 20	0.271	0.258	-0.23	0.796	Small positive
3_GE 3	-0.136	0.205	-0.545	0.268	Near zero
3_GE 4	-0.604	0.211	-1.028	0.2	Credible negative effect
3_GE 5	0.32	0.335	-0.312	0.999	Positive, uncertain
3_GE 6	-0.322	0.235	-0.782	0.131	Weak negative
3_GE 7	0.319	0.264	-0.197	0.652	Small positive
3_GE 8	0.234	0.2	-0.151	0.641	Small positive
3_UGE 1	0.155	0.236	-0.319	0.626	Near zero
3_UGE 2	0.211	0.268	-0.307	0.741	Small positive
POOLED_SMALL_COURSES	-0.962	0.281	-1.537	-0.432	Strong negative effect

Table 10. Variance Components

Component	Mean	SD	2.50%	97.50%
Residual SD (σ)	0.917	0.041	0.641	1.001
Couse-Level Variance	0.27	0.154	0.088	0.663

Table 11. Posterior Predictive Check

Statistic	Mean	SD	2.50%	50%	97.50%
Mean_PPD	-0.0004	0.078	-0.153	-0.001	0.153

Table 12. MCMC Diagnostic (Summary)

Metric	Result
Max R-hat	1.0029
Convergence	Excellent (all <1.01)
Effective Sample Size	1,700-8,200
MCSE	Negative relative to SD

The overall intercept depicted in Table 4 is essentially zero, indicating that cor_average is now correctly centered. Consequently, course effects are interpreted as deviations from a neutral baseline, rather than being inflated by scale artifacts. Course-level variance is substantially reduced compared with the earlier model, confirming that much of the previously observed between-course extremity was driven by sparse data rather than genuine structural differences. Most individual courses have credible intervals that overlap zero, indicating their effects are statistically indistinguishable from the program average after controlling for slight sample instability.

Two results clearly stand out and are defensible. First, 3_GE 4 demonstrates a credibly negative effect, as its entire 95% credible interval falls below zero, indicating systematically weaker outcomes. Even after pooling and shrinkage, this conveys a substantive signal rather than noise. Second, the POOLED_SMALL_COURSES group shows a large and robust negative effect, confirming that very small-enrollment courses, when combined, consistently underperform relative to the program norm. Model diagnostics are exemplary, with R-hat values near 1, high effective sample sizes, and good posterior predictive checks, indicating excellent reproduction of the observed mean. This model is statistically clean and stable; substantially, it narrows the problem from "many extreme courses" to a focused set of underperforming course structures, thus offering a far stronger basis for academic review, policy action, and defensible reporting.

Figure 1 below shows the posterior distributions of rank correlations (p) by course. The message is blunt: while only a handful of courses display extreme, degenerate posteriors, most cluster tightly around zero. Several major-field courses, notably FM and MM, and selected ENTE/FME subjects, produce razor-thin spikes at p = +1 or p = -1, suggesting that practically deterministic monotonic relationships are statistically suspicious and due to small samples, limited score dispersion, or structurally constrained grading rather than accurate, perfect alignment. By contrast, the General Education (GE) and University GE (UGE) courses present bell-shaped posteriors centered close to zero with relatively narrow spreads, which imply weak to modest associations and greater heterogeneity in student performance, precisely what one would want from service courses taken by diverse cohorts. A few of the GE subjects do lean positive or negative, but one is as extreme as those from major courses. Overall, the figure reinforces a clear structural pattern within the BS Business Administration Program: high correlation stability exists only where assessment and enrollment are tightly controlled. At the same time, most courses contribute independently to student outcomes, and any interpretation of "perfect" correlation should be viewed as a data artifact rather than evidence of curricular inevitability.

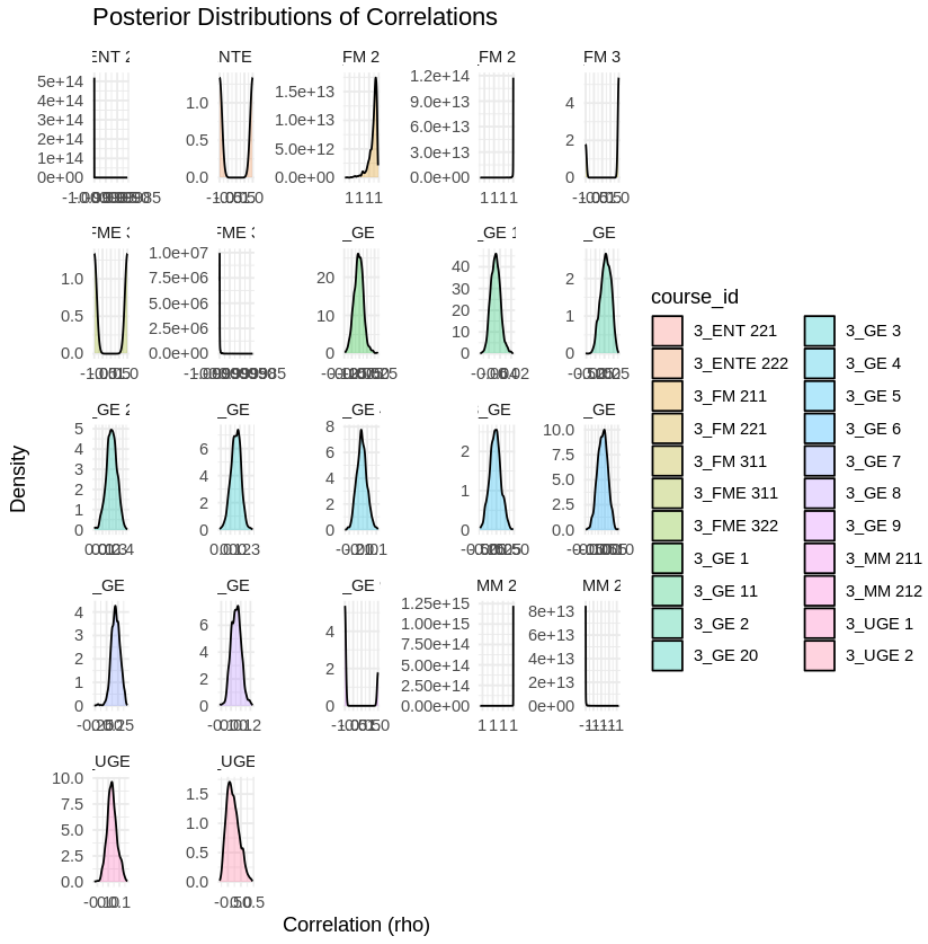

Fig. 1 Posterior Distribution of Course-Level Correlations

Figure 2 below shows a caterpillar plot of posterior course-level random effects from a Bayesian hierarchical model. Points are posterior means, and horizontal bars are 95% credible intervals, all centered around the red dashed zero line, which represents the grand mean. Most course clusters tightly gather around zero, with their intervals overlapping the reference line, indicating that, controlling for overall variability. These are not statistically distinguished from the program average. In contrast, a small subset of courses exhibits decidedly negative effects. For instance, courses 3_FME 311 and 3_FME 322 are identified as having credible intervals that lie completely below zero, suggesting systematic underperformance rather than random fluctuations. Several courses have positive effects, such as 3_FM 211 and 3_ENTE 222, as well as some GE courses. However, the wide intervals associated with those estimates reflect significant uncertainty, primarily driven by small enrollments. The wide interval for the course-level variance parameter indicates substantial heterogeneity across courses overall. Figure 2 supports the notion that, despite variation between courses, few courses exhibit credibly different performance, which in turn supports the need for pooled estimation and targeted academic review rather than broad program-level conclusions.

Figure 3 reveals a striking and concerning pattern in the data: the most extreme course means, both very high and very low, are concentrated in courses with very small enrollments. At the same time, larger classes cluster tightly around the grand mean, represented by the dashed line at approximately 73. Courses with one to three students exhibit wildly inflated means near 100, or else collapsed means close to zero that are statistically unstable and almost certainly. Courses with one to three students exhibit wildly inflated means near 100, or else collapsed means close to zero, which are statistically unstable and almost certainly reflect grading idiosyncrasies or data artifacts rather than actual instructional effectiveness or failure. As the sample size increases, the course means regress toward the grand mean, and mid- to large-enrollment courses show far less dispersion and more credible performance estimates. This pattern directly explains the high between-course variance and significant random effects observed earlier: small-n courses are driving artificial polarization in outcomes. In practical terms, the figure makes it evident that unmoderated small classes are distorting program-level performance metrics, and any evaluative or accountability use of course means without adjusting for enrollment size would be methodologically indefensible.

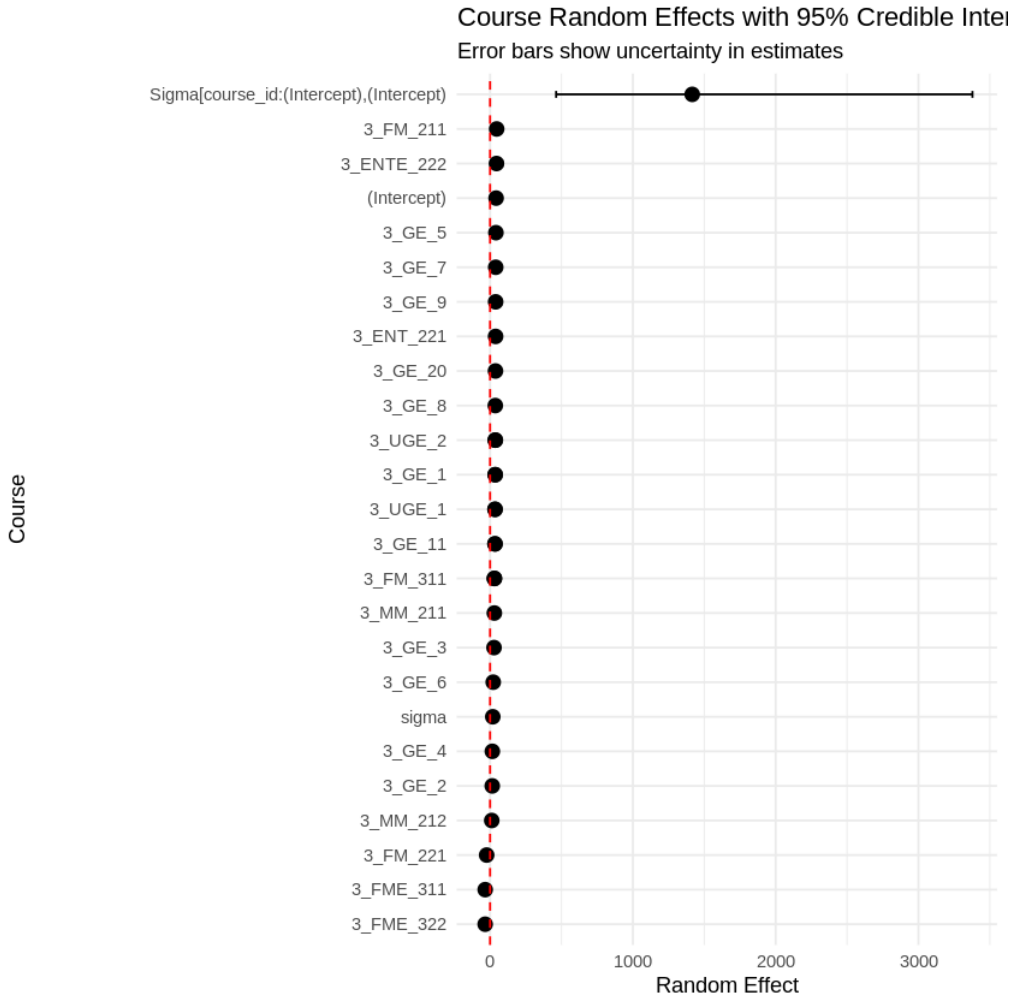

Fig. 2. Course Random Effects with 95% Credible Intervals

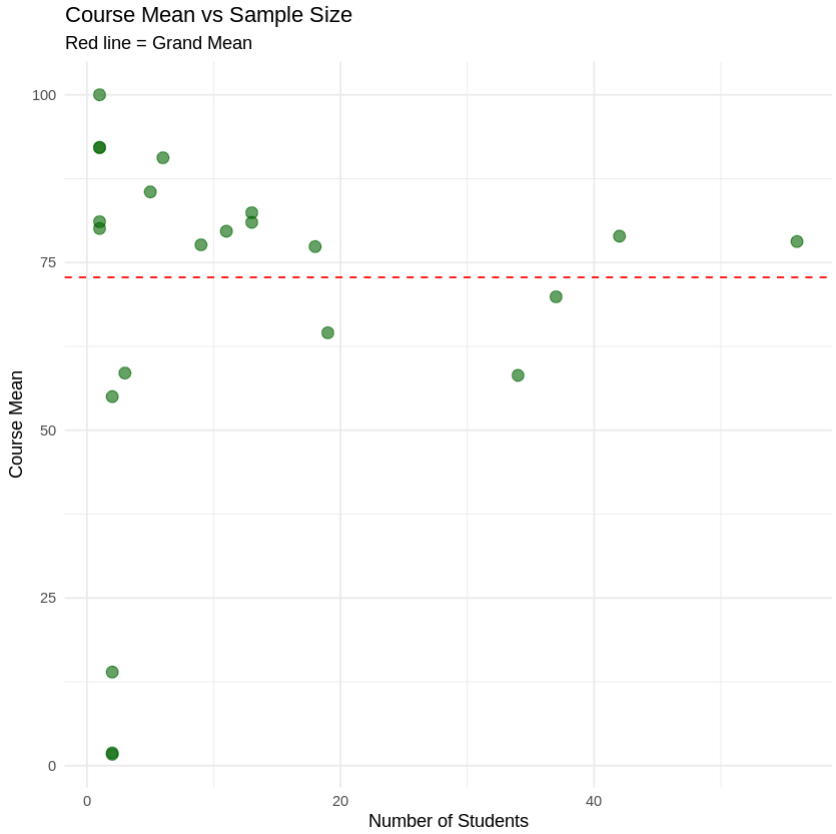

Fig. 3. Course Mean vs Sample Size

Discussions

The study describes structural differences in aggregated performance metrics rather than providing normative judgments about teaching effectiveness, course quality, or administrative efficacy. Within the Bayesian hierarchical framework, a credible interval that does not include zero indicates statistical distinction in model-based estimation, rather than evidence of instructional superiority or deficiency. Accordingly, the identified course-level effects represent probabilistic deviations from the program mean, explicitly adjusted for sampling uncertainty and uneven enrollment, and should be interpreted as properties of the measurement structure rather than causal attributes of courses or instructors.

This interpretation is reinforced by caterpillar plots of random effects, in which courses are positioned relative to the grand mean. Variance component estimates and consistently high intraclass correlation coefficients indicate that a substantial proportion of observed variability is expressed at the course level rather than arising from differences among students within the same course (Yang Hansen et al., 2024). Importantly, this does not imply that course-level factors are causal determinants of student performance. Instead, it reflects the structuring of assessment and grading practices within the observed institutional context, a pattern that has also been noted in prior analyses of performance differences in educational systems (Gyamfi et al., 2022).

At the same time, the joint inspection of course-level effects and enrollment size highlights a second critical issue: statistical instability associated with small samples. Small classes are not only noisier; under naïve aggregation, they produce unstable estimates that can distort program-level inference (Allenby & Rossi, 2006). Courses with very small enrollments exhibit extreme means and wide credible intervals, creating the appearance of exceptional performance or underperformance that is largely driven by sparse data. These patterns account for the inflated between-course variance observed in the unpooled model and demonstrate the limitations of direct comparisons based on raw course averages. The wide posterior uncertainty associated with these estimates further confirms that they should not be interpreted as reliable indicators of substantive instructional differences but rather as manifestations of estimator instability (Mascha & Vetter, 2018).

Methodologically, the application of Bayesian hierarchical pooling reframes course-level evaluation from descriptive comparison to statistically grounded inference. Partial pooling reduces spurious extremity by shrinking imprecise estimates toward the program mean while preserving differences supported by sufficient evidence. In the pooled specification, most course effects contract toward zero and lose apparent distinctiveness, whereas a limited subset retains consistent deviations, indicating structurally meaningful differences. In this sense, hierarchical modeling functions as a filtering mechanism that distinguishes structural signal from measurement noise and aligns course-level evaluation with principles of uncertainty-aware inference (Greenland et al., 2016).

Conclusions

This study concludes that students' performance in the BS Business Administration program is primarily determined by structures at the course level rather than by individual student differences. Consequently, a significant proportion of the variability between courses is attributed to how courses are structured, assessed, and taught. It also shows that raw comparisons between courses are highly unreliable when enrollments are small because sparse data inflate apparent extremes and mask actual instructional effects. These distortions are corrected through Bayesian hierarchical pooling, which shrinks and stabilizes unstable estimates without eliminating credible signals of systematic underperformance in the small set of identified courses. The evidence supports targeted, course-specific academic reviews rather than broad or punitive program-wide interventions, emphasizing that sound methodological choices are essential to fair and defensible evaluations. Overall, the study indicates that meaningful program assessment requires both recognition of structural course effects and statistical discipline in handling small-sample variability.

Although hierarchical modeling offers a more statistically informed framework for understanding structural differences in course-level performance, any administrative or policy intervention would necessarily require additional qualitative and contextual information beyond the scope of the current analysis.

he study suggests that courses with a persistent negative effect after Bayesian pooling, such as 3_FME 311 and 3_FME 322, should be considered within a broader curricular reform and intervention strategy. These courses exhibit a strong negative random effect, with credible intervals below zero, suggesting they are high-priority candidates for immediate academic review in instructional delivery, content alignment, grading standards, and assessment design. Moreover, 3_GE 4, despite shrinkage, remains negative and requires similar attention, indicating the presence of structural issues rather than sampling noise. Courses in the POOLED_SMALL_COURSES category that are not individually evaluated should be addressed through policy measures aimed at consolidating courses and standardizing assessment frameworks. On the other hand, courses such as 3_FM 211, 3_ENTE 222, and 3_ENTE 221 exhibit a significant positive effect, albeit with considerable uncertainty, and require further validation before being considered benchmarks of excellence. This is due to the potential for student performance estimates to be biased by small sample sizes. The recommendations include initiating diagnostic reviews and implementing uniform assessment procedures for underperforming courses, establishing regular program evaluation to institutionalize hierarchical modeling, and preventing decision-making based on unadjusted course means.

Ethical Considerations

This study is based exclusively on anonymized secondary academic records obtained from official institutional assessment data. No personally identifiable information was accessed, collected, or processed at any stage of the research. The dataset was fully de-identified prior to analysis, and individual students cannot be re-identified from the reported results.

In accordance with institutional research guidelines for minimal-risk studies using de-identified administrative data, formal ethics committee approval was not required. All procedures complied with applicable standards for data protection, confidentiality, and responsible research conduct.

Data Availability

The data supporting the findings of this study consist of anonymized institutional academic records and are subject to confidentiality and data protection restrictions imposed by the hosting institution. Therefore, the raw data are not publicly available.

Aggregated data summaries, model specifications, and analytical procedures necessary to reproduce the reported results are described in sufficient detail within the manuscript. Reasonable requests for additional methodological clarification may be considered by the corresponding author, subject to institutional approval.

AI Transparency Statement

The authors declare that no generative artificial intelligence (AI) tools were used for data generation, statistical analysis, or result interpretation. Computational modeling and statistical inference were conducted using established statistical software and fully specified Bayesian procedures.

AI-based tools may have been used solely for language editing or stylistic refinement of the manuscript text. Such use did not influence the scientific content, data analysis, interpretation of results, or conclusions of the study. The authors retain full responsibility for the integrity and originality of the work.

Acknowledgment

The researchers would like to acknowledge the Administration and the Research and Publication Center (RPC) for their assistance, as well as their colleagues for their moral support.

Conflict of interest

None

References

Anwar, M.A., Ahmed, N., & Al Ameen, A.M. (2012). An Outcome-Based Assessment and Improvement System for Measuring Student Performance and Course Effectiveness. *Contemporary Issues in Education Research, 5*(4), 279-294. https://doi.org/10.19030/cier.v5i4.7272

Cabrera, A.F., Colbeck, C.L., & Terenzini, P.T. (2001). Developing performance indicators for assessing classroom teaching practices and student learning. *Research in higher education, 42*(3), 327-352. https://doi.org/10.1023/A:1018874023323

Hristov, S., Nakov, D., & Miočinović, J. (2023). Constructive alignment between objectives, teaching and learning activities, student competencies and assessment methods in higher education. *Journal of Agriculture and Plant Sciences, 21*(2), 21-36. https://doi.org/10.46763/JAPS23212021h

Lewis, E. (2021). Best practices for improving the quality of the online course design and learners experience. *The Journal of Continuing Higher Education, 69*(1), 61-70. https://doi.org/10.1080/07377363.2020.1776558

Kennedy, D. (2008). Linking Learning Outcomes and Assessment of Learning of Student Science Teachers. *Science Education International, 19*(4), 387-397. https://eric.ed.gov/?id=EJ890648&utm_source=chatgpt.com

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., & Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience, 14*(5), 365-376. https://doi.org/10.1038/nrn3475

Anderson, S.F., & Maxwell, S.E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate behavioral research, 52*(3), 305-324. https://doi.org/10.1080/00273171.2017.1289361

Mascha, E.J., & Vetter, T.R. (2018). Significance, errors, power, and sample size: the blocking and tackling of statistics. *Anesthesia & Analgesia, 126*(2), 691-698. https://doi.org/10.1213/ANE.0000000000002741.

Berry, S.M., Broglio, K.R., Groshen, S., & Berry, D.A. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clinical Trials, 10*(5), 720-734. https://doi.org/10.1177/1740774513497539

Vandendijck, Y., Faes, C., Kirby, R.S., Lawson, A., & Hens, N. (2016). Model-based inference for small area estimation with sampling weights. *Spatial Statistics,* 18, 455-473. https://doi.org/10.1016/j.spasta.2016.09.004

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods, 22*(4), 760. https://doi.org/10.1037/met0000136

McGlothlin, A.E., & Viele, K. (2018). Bayesian hierarchical models. *Jama, 320*(22), 2365-2366. https://doi.org/10.1001/jama.2018.17977

Chan, E.K. (2014). *Standards and guidelines for validation practices: Development and evaluation of measurement instruments.* In Validity and validation in social, behavioral, and health sciences (pp. 9-24). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_2

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*(1), 29-49. https://doi.org/10.1016/j.stueduc.2007.01.004

Ramezani, S.G., & Mostafavi, Z.S. (2025). Developing and validating a comprehensive scale for accreditation standards and quality assurance in e-learning institutions. *Education and Information Technologies,* 1-49. https://doi.org/10.1007/s10639-025-13587-5

Baartman, L.K., Bastiaens, T.J., Kirschner, P.A., & Van der Vleuten, C.P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational research review, 2*(2), 114-129. https://doi.org/10.1016/j.edurev.2007.06.001

Inglis, A. (2008). Approaches to the validation of quality frameworks for e-learning. *Quality Assurance in Education, 16*(4), 347-362. https://doi.org/10.1108/09684880810906490

Whiting, P., Wolff, R., Mallett, S., Simera, I., & Savović, J. (2017). A proposed framework for developing quality assessment tools. *Systematic reviews, 6*(1), 204. https://doi.org/10.1186/s13643-017-0604-6

Bentley, T.G., Cohen, J.T., Elkin, E.B., Huynh, J., Mukherjea, A., Neville, T.H., ... & Broder, M.S. (2017). Validity and reliability of value assessment frameworks for new cancer drugs. *Value in Health, 20*(2), 200-205. https://doi.org/10.1016/j.jval.2016.12.011

Kruger, T., & Leuro, J. (2015, September). *Using Quality Assurance Principles to Help Ensure the Validity and Reliability of Competency Assessments.* In SPE Offshore Europe Conference and Exhibition (pp. SPE-175491). SPE. https://doi.org/10.2118/175491-MS

Feiler, P.H., Goodenough, J.B., Gurfinkel, A., Weinstock, C.B., & Wrage, L. (2012). *Reliability validation and improvement framework* (No. CMUSEI2012SR013). https://www.sei.cmu.edu/documents/1918/2012_003_001_34081.pdf

Smidt, A., Balandin, S., Sigafoos, J., & Reed, V.A. (2009). The Kirkpatrick model: A useful tool for evaluating training outcomes. *Journal of Intellectual and Developmental Disability, 34*(3), 266-274. https://doi.org/10.1080/13668250903093125

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in higher education. *Educational assessment, evaluation and accountability, 22*(3), 215-225. https://doi.org/10.1007/s11092-010-9098-7

Cheung, V.K. L., Chia, N.H., So, S.S., Ng, G.W. Y., & So, E.H. K. (2023). Expanding scope of Kirkpatrick model from training effectiveness review to evidence-informed prioritization management for cricothyroidotomy simulation. *Heliyon, 9*(8). https://doi.org/10.1016/j.heliyon.2023.e18268

Thörn, J., Strandberg, P.E., Sundmark, D., & Afzal, W. (2022). Quality assuring the quality assurance tool: applying safety-critical concepts to test framework development. *PeerJ Computer Science,* 8, e1131. https://doi.org/10.7717/peerj-cs.1131

Nawaz, F., Ahmad, W., & Khushnood, M. (2022). Kirkpatrick model and training effectiveness: a meta-analysis 1982 to 2021. *Business & Economic Review, 14*(2), 35-56. https://doi.org/10.22547/BER/14.2.2

Baldwin, S.A., & Fellingham, G.W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological methods, 18*(2), 151. https://doi.org/10.1037/a0030642

Schmid, C.H., & Brown, E.N. (2000). Bayesian hierarchical models. Methods in enzymology, 321, 305-330. https://doi.org/10.1016/S0076-6879(00)21200-7

Columb, M.O., & Atkinson, M.S. (2016). Statistical analysis: sample size and power estimations. *Bja Education, 16*(5), 159-161. https://doi.org/10.1093/bjaed/mkv034

Chen, C., Wakefield, J., & Lumely, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spatial and spatio-temporal epidemiology,* 11, 33-43. https://doi.org/10.1016/j.sste.2014.07.002

Goodhue, D., Lewis, W., & Thompson, R. (2006, January). PLS, small sample size, and statistical power in MIS research. *In Proceedings of the 39th Annual Hawaii International Conference on System Sciences* (HICSS'06) (Vol. 8, pp.

202b-202b). IEEE. https://doi.org/10.1109/HICSS.2006.381

Monnahan, C.C., Thorson, J.T., & Branch, T.A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution, 8*(3), 339-348. https://doi.org/10.1111/2041-210X.12681

Bocquel, M., Papi, F., Podt, M., & Driessen, H. (2013). Multitarget tracking with multiscan knowledge exploitation using sequential MCMC sampling. IEEE *Journal of Selected Topics in Signal Processing, 7*(3), 532-542. https://doi.org/10.1109/JSTSP.2013.2251317

Nguyen, T.D., Gupta, S., Rana, S., & Venkatesh, S. (2018). Stable bayesian optimization. *International Journal of Data Science and Analytics, 6*(4), 327-339. https://doi.org/10.1007/s41060-018-0119-9

Kim, M., Ding, Y., Malcolm, P., Speeckaert, J., Siviy, C.J., Walsh, C.J., & Kuindersma, S. (2017). Human-in-the-loop Bayesian optimization of wearable device parameters. *PloS one, 12*(9), e0184054. https://doi.org/10.1371/journal.pone.0184054

Stern, H.S., & Sinharay, S. (2005). Bayesian model checking and model diagnostics. *Handbook of Statistics,* 25, 171-192. https://doi.org/10.1016/S0169-7161(05)25007-1

Koch, K.R. (2018). Bayesian statistics and Monte Carlo methods. *Journal of Geodetic Science, 8*(1), 18-29. https://doi.org/10.1515/jogs-2018-0003

Chen, J.J., Lai, P.C., & Huang, Y.T. (2025). Bayesian reanalysis reinforces the potential mortality benefit of TNF-α inhibitors in COVID-19: a methodological perspective. *Critical Care, 29*(1), 250. https://doi.org/10.1186/s13054-025-05506-4

Gajewski, B.J., Simon, S.D., & Carlson, S.E. (2008). Predicting accrual in clinical trials with Bayesian posterior predictive distributions. *Statistics in medicine, 27*(13), 2328-2340. https://doi.org/10.1002/sim.3128

Feng, Y., Gao, K., & Lacasse, S. (2024). Bayesian partial pooling to reduce uncertainty in overcoring rock stress estimation. *Journal of Rock Mechanics and Geotechnical Engineering, 16*(4), 1192-1201. https://doi.org/10.1016/j.jrmge.2023.05.003

Поза межами курсових середніх: узагальнена байєсівська ієрархічна методологічна рамка оцінювання результатів навчання на рівні курсів

Вінсенте Е. Монтано[1ABCD], Арчі Г. Рейєс[1ABCD]

[1]Університет Мінданао

Авторський вклад: A – дизайн дослідження; B – збір даних; C – статаналіз; D – підготовка рукопису; E – збір коштів

Реферат. Стаття: 12 с., 12 табл., 3 рис., 40 джерел.

Обґрунтування. Оцінювання результатів навчання на рівні курсів у вищій освіті часто базується на порівнянні середніх показників, що передбачає незалежність курсів та однакову надійність оцінок. За умов малих і нерівномірних контингентів така практика призводить до статистичної нестабільності та перебільшення крайніх значень, ускладнюючи інтерпретацію курсових відмінностей.

Мета. Метою дослідження є обґрунтування узагальненої методологічної рамки застосування байєсівського ієрархічного моделювання (Bayesian hierarchical modeling, BHM) для оцінювання результатів навчання на рівні курсів з урахуванням невизначеності та нерівномірності вибірок.

Матеріали і методи. У дослідженні використано ієрархічну байєсівську модель, у якій результати навчання студентів моделюються на індивідуальному рівні з урахуванням їх належності до конкретних курсів, що відображає багаторівневу організацію освітніх даних. Модель передбачає декомпозицію загальної дисперсії на внутрішньокурсову та міжкурсову складові з оцінюванням курсових ефектів на основі апостеріорних розподілів. Для зменшення спотворень, зумовлених малими обсягами вибірок, застосовано механізм часткового пулінгу. Як емпіричну ілюстрацію використано знеособлені дані про результати навчання 279 студентів у 22 курсах.

Результати. Показано, що наївні порівняння курсів за середніми значеннями систематично перебільшують крайні оцінки за малих обсягів вибірок, формуючи нестабільні та потенційно хибні висновки. Застосування ієрархічного байєсівського підходу з частковим пулінгом суттєво знижує штучну екстремальність оцінок і водночас зберігає структурно обґрунтовані міжкурсові відмінності.

Висновки. Запропонована методологічна рамка забезпечує статистично обґрунтовану альтернативу описовому агрегуванню та ранжуванню курсів, орієнтуючи оцінювання результатів навчання на ймовірнісну структурну інтерпретацію з урахуванням невизначеності.

Ключові слова: байєсівське ієрархічне моделювання, багаторівневий аналіз, курсовий рівень оцінювання, нестабільність малих вибірок, освітні вимірювання.

Information about the Authors:

Montano, V.E.: vicente_montano@umindanao.edu.ph; https://orcid.org/0000-0001-9117-568X; Business Economics Department, College of Business Administration Education, University of Mindanao, Bolton St., 8000, Davao City, Philippines.

Reyes, A.G.: archiereyes@umindanao.edu.ph ; https://orcid.org/0009-0005-7443-3022; Human Resource Management Department, College of Business Administration Education, University of Mindanao, Bolton St., 8000, Davao City, Philippines.

Cite this article as: Montano, V., & Reyes, A. (2026). Beyond Course Averages: A Generalized Bayesian Hierarchical Framework for Course-Level Learning Evaluation. *Journal of Learning Theory and Methodology, 7*(1), 37-48. https://doi.org/10.17309/jltm.2026.7.1.04

Received: 17.01.2026. Accepted: 17.02.2026. Published: 30.04.2026