JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(2), 196–210. https://doi.org/10.18608/jla.2025.8753

How Much is Enough? Formative Assessment Dynamics: Quantifying the Relationship Between Intermediate Quiz Performance and Final Exam Scores

Michael J. Parker[1], Matt Bunch[2] and Andrew Pike[3]

Abstract

While the educational value of formative assessment is widely acknowledged, the precise amount needed to effectively predict student performance on summative assessments remains unclear. This study investigates the relationship between intermediate formative assessment performance and final exam scores, addressing the critical question of how much assessment is needed for accurate prediction. Using a large dataset encompassing over 20,000 student enrollments across 127 course runs of 15 online biomedical sciences courses, we examined the correlation between intermediate assessment scores and final exam performance. Our results show that after completing about 40% of the formative assessments in a course, student scores demonstrate a strong correlation (Pearson r > 0.7) with their final exam scores. The correlation after taking additional formative assessments reaches a maximum of approximately 0.75. This finding was consistent across different course types and lengths, suggesting that the relative amount of assessment taken, rather than the absolute number, is key. Surprisingly, we found that random sampling of assessments was even more predictive than chronological sampling, suggesting that the proportion of questions used, relative to the total number of assessment questions, is more important than their specific sequence. These findings contribute to a deeper understanding of the predictive capabilities of formative assessment, and enable educators to identify at-risk students earlier, optimize assessment design, and develop more efficient and targeted interventions.

Notes for Practice

- Formative assessment through quizzes is widely used in education, but the optimal amount needed for predicting student performance has been unclear.
- After students complete approximately 40% of course quiz questions, their performance strongly correlates with final exam scores, enabling early identification of at-risk students.
- The predictive threshold is based on a percentage of total quiz questions rather than an absolute number, making findings applicable across different course lengths.
- Random sampling of quiz questions provides equal or better predictive power compared to sequential sampling, suggesting representative coverage matters more than specific sequencing.
- These insights can inform automated early warning systems and optimize quiz loads while maintaining predictive power for student outcomes.

Keywords: Assessment, formative assessment, learning analytics, learning outcomes, outcomes prediction

Submitted: 30/10/2024 — **Accepted:** 23/04/2025 — **Published:** 04/07/2025

Corresponding author [1]Email: mparker@alum.mit.edu Address: HMX, Office for External Education, Gordon Hall, Suite 013, 25 Shattuck Street, Boston, Massachusetts 02115, USA. ORCID iD: https://orcid.org/0000-0003-4739-5217

[2]Email: matt_bunch@hms.harvard.edu Address: Office for External Education, Gordon Hall, Suite 013, 25 Shattuck Street, Boston, Massachusetts 02115, USA.

[3]Email: apike1@oberlin.edu Address: HMX, Office for External Education, Gordon Hall, Suite 013, 25 Shattuck Street, Boston, Massachusetts 02115, USA. ORCID iD: https://orcid.org/0000-0003-3155-9969

1. Introduction

Formative assessment through quizzes is a widely used educational tool, serving as a cornerstone in both traditional and online learning environments (Black & Wiliam, 1998; Gikandi et al., 2011). The practice of using ongoing assessments to guide instruction and learning has a long history in education, with its roots tracing back to the 1960s when researchers began to distinguish between summative and formative evaluation (Scriven, 1966). Formative assessment, defined as the process of gathering evidence about student learning to inform instruction and provide feedback, has since evolved into a key pedagogical strategy (Hattie & Timperley, 2007; Sadler, 1989; Nicol & Macfarlane-Dick, 2006; Wiliam, 2011). Its effectiveness has been well-documented across various disciplines and educational levels, from primary education to higher learning (Black & Wiliam, 1998; Sortwell et al., 2024).

Over the past several decades, research on formative assessment has undergone significant development. Early work focused on the theoretical foundations and potential benefits of formative practices (Bloom, 1968). As the field progressed, researchers began to investigate specific strategies for implementing formative assessment, such as questioning techniques, peer assessment, and self-evaluation (Black et al., 2003). The adoption of digital technologies in education has further transformed formative assessment practices, enabling more frequent, automated, and data-driven approaches (Spector et al., 2016).

The impact of formative assessment on student learning outcomes has been a subject of extensive study. Meta-analyses have consistently shown positive effects of formative assessment on academic achievement, with particularly strong benefits for low-achieving students (Sortwell et al., 2024; Yeh, 2010; Kluger & DeNisi, 1996). Moreover, formative assessment has been linked to improved metacognitive skills, increased student engagement, and enhanced self-regulated learning (Nicol & Macfarlane-Dick, 2006; McLaughlin & Yan, 2017; Tempelaar et al., 2013, 2015). These findings have led to widespread adoption of formative assessment strategies in educational policies and practices worldwide (OECD, 2005).

In recent years, the transition to online and blended learning models has accelerated the adoption of automated formative assessments, particularly in the form of multiple-choice questions (MCQs; Bennett, 2011; Hrastinski, 2019; Gikandi et al., 2011; Nicol, 2007). This shift has coincided with an increased focus on predicting student outcomes in education, driven by the broader context of educational effectiveness and accountability (Siemens, 2013). Stakeholders across the educational spectrum, including educators, institutions, and students, have vested interests in understanding and improving academic performance (Ferguson, 2012). This focus on prediction has become even more pronounced with the proliferation of online learning platforms and learning management systems (LMSs), which generate large amounts of data amenable to analysis (Long & Siemens, 2011).

The insights gained from predicting student performance using formative assessments are particularly relevant to the design of adaptive learning systems. These systems, as highlighted in research on data analytics and adaptive learning (Moskal et al., 2023), often utilize formative MCQs to guide students through personalized learning experiences as a form of just-in-time intervention. Consequently, investigations into the predictive power of formative assessment, especially concerning the optimal quantity needed for accurate prediction, directly inform the development of more effective adaptive learning architectures.

Learning analytics, as a field, has emerged as a key tool for outcome prediction in digital learning environments (Greller & Drachsler, 2012; Siemens, 2012; Dawson et al., 2014; Clow, 2013). By leveraging data from various sources, including formative assessments, learning analytics aims to provide insights that can lead to targeted interventions and personalized learning experiences (Knight et al., 2014; Greller & Drachsler, 2012; Hattie, 2009). The potential for using formative assessment as part of a broader collection of data to predict course outcomes, particularly final exam scores, has been a subject of considerable research interest (Johnson et al., 2015; Gašević et al., 2015). However, the optimal use of this data for early identification of at-risk students and the design of effective support strategies remains an area of active investigation (Wolff et al., 2013; Foster & Siddle, 2020; Glick et al., 2019; West et al., 2016). Figure 1 shows the close relationship between some of the conceptual entities that relate to formative assessment and learning analytics.

Despite the growing body of literature on formative assessment and learning analytics, several critical gaps remain in our understanding of the relationship between formative assessment performance and final outcomes. Current research has primarily focused on aggregate formative assessment data (Bulut et al., 2023; in other words, using a single aggregate score from all of a student's formative assessment in a course), often overlooking the predictive power of intermediate assessments. Moreover, there is a lack of investigation into the optimal quantity of assessment needed for accurate prediction, a question with significant implications for assessment design, student workload, and the development of more efficient and effective assessment practices.

The potential of intermediate formative assessment data for early prediction of course performance is particularly relevant in the context of timely interventions (Foster & Siddle, 2020). However, determining the optimal assessment quantity for prediction presents challenges, balancing the need for accurate prediction with the skill and time required to create well-constructed assessment questions (van der Vleuten & Schuwirth, 2005). This balance is crucial for the design of early warning

ISSN 1929-7750 (online).

systems in educational settings, which could lead to more timely and targeted interventions for struggling students (Macfadyen & Dawson, 2010).

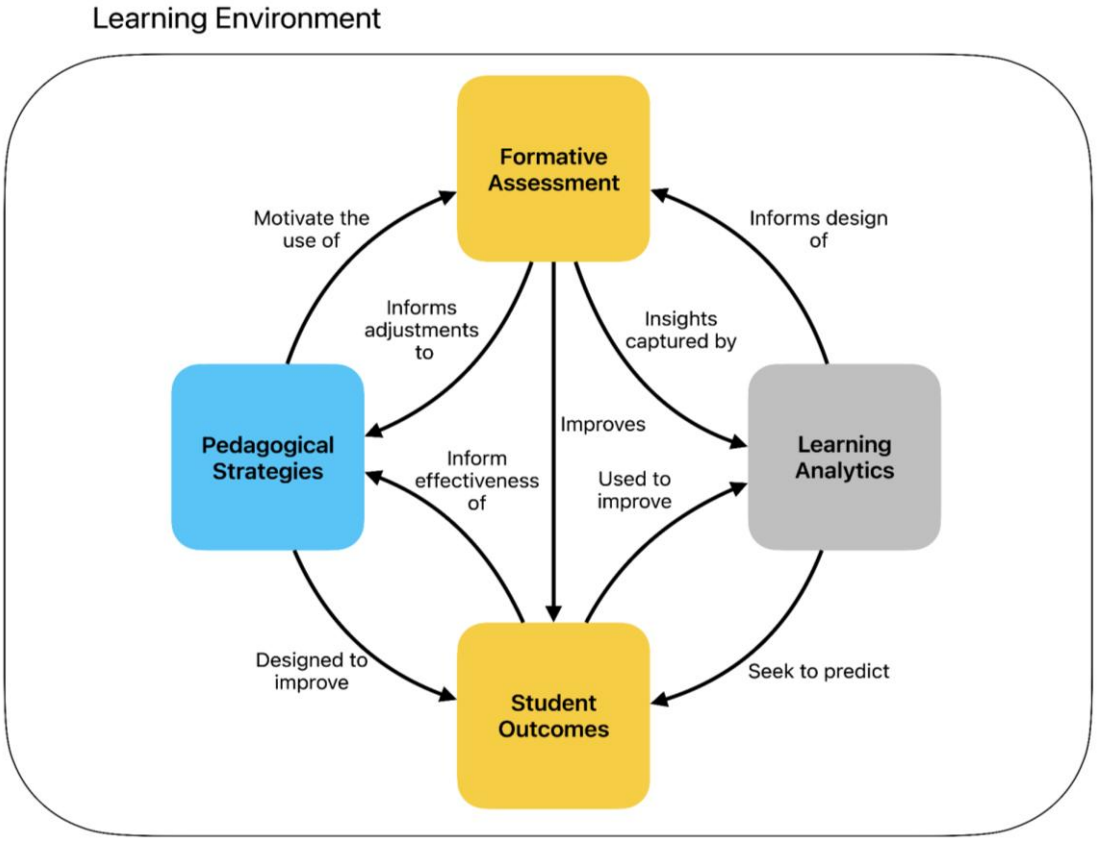

Figure 1. The roles and relationships of pedagogical strategies, formative assessment, learning analytics, and student outcomes in the learning environment.

Our study aims to address these gaps by investigating the dynamics of formative assessment in relation to final exam scores. To explore this relationship, we conducted a large-scale analysis across multiple online biomedical science courses, encompassing 21,306 enrollments across 127 course runs. Our focus on automatically graded questions (primarily MCQs) reflects their widespread use in both online and traditional educational settings. We chose to focus on formative assessment questions because their use is pedagogically grounded, and the data is easily accessible by teachers.

The main research question guiding this study is this: What is the relationship between intermediate formative assessment results and final exam scores? More specifically, we seek to answer the following questions:

1. How much formative assessment is needed to provide robust predictive power for final exam scores?
2. What is the relationship between intermediate formative assessment results and final exam scores?
3. Is the predictive power of intermediate formative assessment based on absolute question count or relative to overall course content?

By examining these questions, we aim to provide insights that can inform assessment design and predictive modelling in education. This research contributes to the "learning analytics loop" by examining how intermediate assessment data can be effectively analyzed and potentially translated into actionable feedback (Ferguson et al., 2023; Motz et al., 2023). Through investigating the relationship between formative assessments and final outcomes, this work aims to inform when and how assessment data might be optimally used to enable timely interventions, helping to "close the loop" between data collection and improving student outcomes.

2. Methods

2.1. Courses and Learners

Data were collected from a series of courses focused on foundational and cutting-edge biomedical science topics (Table 1). The courses are offered for credit at our university and externally as paid public course runs (PCRs) open to individuals from

the general public who apply to and are accepted into the courses, with the opportunity to earn a certificate. This study included all individuals enrolled in the PCRs between April 2020 and November 2022. Demographic information was not collected for learners. The study was conducted using anonymized data after the completion of the courses and awarding of any relevant certificates and was determined not to be human subjects research by the HMS Office of Human Research Administration.

This study includes 127 instances of 15 course types run quarterly over 10.5 weeks, with course content, including videos and assessment questions, released weekly, followed by a summative final exam (Table 1). New course types were introduced at different points during the study period and, therefore, the course types have differing numbers of course instances.

Table 1. Courses Included in This Study, with Enrollment Numbers

Course Type	Course Category	Number of Enrollments During Study Period	Number of Instances (Course Runs) During Study Period
Immunology	Fundamentals	3,924	11
Pharmacology	Fundamentals	2,099	11
Genetics	Fundamentals	1,908	11
Physiology	Fundamentals	1,710	11
Biochemistry	Fundamentals	1,497	11
Immuno-oncology	Pro	3,122	11
Cancer Genomics and Precision Oncology	Pro	2,514	11
Novel Therapies for Chronic Inflammation, Autoimmunity, and Allergy	Pro	1,071	8
Genetics Testing and Sequencing Technologies	Pro	982	8
Pharmacology Essentials	Pro	688	11
Genetics Essentials	Pro	629	11
Vaccines and Viral Immunology	Pro	508	4
Drug Delivery	Pro	357	6
Drug Discovery and Development	Pro	154	1
Gene Therapy	Pro	143	1
Grand total		21,306	127

Of the 21,306 enrollments during the study period, 16,707 (78%) completed the courses as defined by having attempted at least 85% of graded assessments in the body of the course (excluding the final exam) and having a non-zero final exam score. This study focuses on completers, given that the focus is on prediction of final exam scores. There were 879,971 assessment attempts overall, of which 831,065 came from those who completed the courses (further breakdown shown in Table 2).

Table 2. Number of Attempts on Assessment Questions Taken (Completers)

Number of Assessment Attempts	Count (percent of assessment questions)
1	551,948 (77.15)
2	162,369 (22.7)
3	1,095 (0.15)
Total assessment questions taken	715,412
Total attempts*	879,971

** Note: Total attempts is the sum of the number of attempts times count.*

2.2. Course and Assessment Question Design

Each course comprises a series of lessons that include videos, assessment questions, text, interactive visualizations, and discussion forums. Videos are typically short (5–10 minutes) and utilize different modalities, including writing with narration (chalk-talk style), clinical pieces involving doctors and patients, interviews, and animations. The number of assessment questions and videos are shown for the courses in Table 3.

Table 3. Number of Videos and Number of Graded Assessment Questions in the Lessons (Body of the Course, or Formative Assessment Component) and on the Final Exam (Summative Assessment Component) by Course Type

Course Type	Course Category	Videos	Graded Assessments – Lessons	Assessments – Final Exam
Biochemistry	Fundamentals	98	204	26
Genetics	Fundamentals	115	293	27
Immunology	Fundamentals	91	275	29
Pharmacology	Fundamentals	88	255	27
Physiology	Fundamentals	114	172	25
Cancer Genomics and Precision Oncology	Pro	49	101	20
Genetics – Essentials	Pro	42	103	20
Genetic Testing and Sequencing Technologies	Pro	54	125	20
Gene Therapy	Pro	55	129	20
Novel Therapies for Chronic Inflammation, Autoimmunity, and Allergy	Pro	60	109	20
Immuno-oncology	Pro	56	103	20
Vaccines and Viral Immunology	Pro	55	107	20
Drug Delivery	Pro	55	102	20
Drug Discovery and Development	Pro	77	160	20
Pharmacology – Essentials	Pro	46	123	20

Courses vary in length, with longer courses (Fundamentals courses) consisting of approximately 10 lessons, and shorter, more advanced courses (Pro courses) consisting of approximately five lessons. Lessons, including the assessment questions, are typically released on a weekly basis during a course, with one lesson released per week.

All courses were created as collaborations between faculty and a dedicated curriculum team, with videos and visualizations built in conjunction with a creative media team. The curriculum team faculty received formal pedagogical training, including design of assessment questions. Courses underwent an extensive review process for pedagogical practices and accuracy.

All courses follow a similar format, with videos and interactive visualizations interleaved with assessment questions, followed by a final exam. Assessment questions typically follow videos, allowing the learner to perform knowledge checks and incorporating spaced repetition and interleaving into the courses (Cepeda et al., 2006; Rohrer, 2012). All assessment questions in the lessons are considered *formative assessment,* as opposed to the final exam, which is considered *summative assessment*. Formative assessment is considered a critical component of all courses, both as a tool for learning and for evaluation.

Most assessment questions are MCQs, including multi-select checkbox questions, with a smaller number of open response questions including short answer questions and thought questions (Figure A1). Thought questions were graded on participation only and were excluded from the analysis of formative assessments. Questions were designed based on best practices (for example, no “all of the above” type MCQs, plausible distractors, etc.; Butler, 2018). All questions in the lessons, except for a small number of “drag-and-drop” style questions, allowed two attempts for mastery learning. Grades were recorded after each attempt; however, only the score after all attempts was counted toward students’ final overall scores. Final exam questions were single attempt only.

Raw assessment data consisted of score and correctness data for each student’s graded assessment attempts across the body of a course and the final exam. The raw assessment attempt data was further processed to derive the fraction of multiple-

attempt assessments correct on the first attempt (referred to as *fraction after first attempts*, a value between 0.0 and 1.0), and the fraction of assessments correct overall (i.e., correct after all attempts, referred to as *fraction after all attempts*, also a value between 0.0 and 1.0).

2.3. Data Collection

The courses were delivered via a version of the Open edX LMS platform, with assessment question scoring data available via the platform database and log files.

2.4. Statistics

Scikit-learn 1.2.0 was used for statistical tests, along with numpy 1.23.5 and Pandas 2.0.1 for data analysis.

To determine whether fraction on first attempt correlation was significantly different than fraction after all attempts at each decile, we used Fisher's z-transformation on the correlation coefficients, then calculated z-statistic based on the standard error and determined significance based on a two-tailed test for a standard normal distribution.

To systematically identify where the predictive power stabilizes across graphs, we established a threshold criterion based on the Pearson correlation coefficient (r) for first-attempt performance. This threshold was set at 95% of the maximum observed r value for first attempts. We selected this criterion because it corresponds to the point where each curve demonstrates asymptotic behaviour, capturing approximately 90% of the maximum possible explained variance in final exam scores (since variance explained is proportional to r^2, with $0.95^2 \approx 90\%$). This approach provides a standardized method for comparing different correlation curves and identifying the point beyond which additional formative assessments yield diminishing returns in predictive power.

3. Results

To better understand the relationship between intermediate formative assessment results and final exam scores, we examined data from the 831,065 assessment attempts from completers in the set of 127 biomedical science courses. In section 3.1, we address the research question of how much formative assessment is needed to predict final exam scores and quantify this relationship using data from all 127 courses. In section 3.2, we explore the question of whether reaching the demonstrated predictive threshold of correlation (see section 2.4) requires completing a quantity of quiz questions that is an absolute number or whether it is more closely based on a *percentage* of the total number of quiz questions in a course. In section 3.3, we focus on whether prediction requires using quiz question data in sequential order (the order in which the quiz questions appear in the course) or whether a randomized subset of quiz questions is equally informative in terms of predictive power.

3.1. Final Exam Correlation to Formative Assessment Results

We aggregated data from the 127 courses to determine how well student scores, after a cumulative subset of quiz questions in a course, taken as varying percentages of total formative assessment questions in that course, correlated with their final exam scores. The aggregate distribution for fraction on first attempts, fraction after all attempts, and final exam scores is shown in Figure 2. This suggests that students, on average, did slightly worse on first attempts, at least judged by mean scores, than they did on the final exam (which allowed a single attempt per problem), although the overall distribution had a lower interquartile range for fraction after first attempts. Fraction correct after all attempts is substantially higher than fraction after first attempts, indicating that students are usually able to determine the correct answer after two attempts.

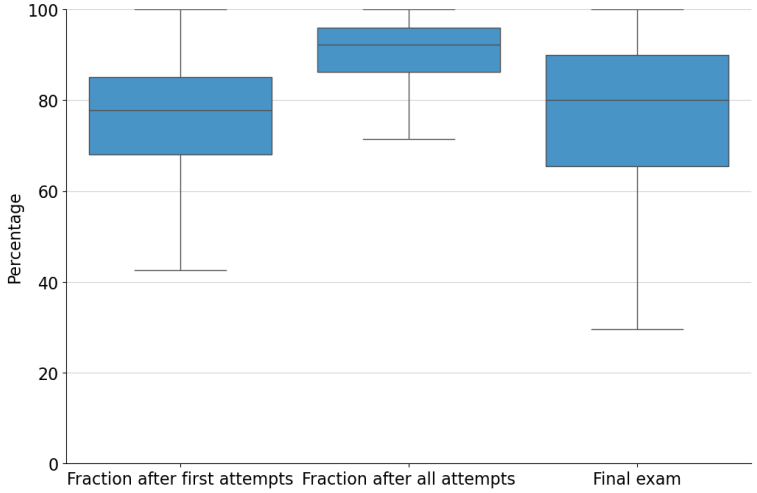

Figure 2. Average fraction correct (score) after first attempts, after all attempts, and final exam scores.

To determine the correlation of intermediate assessment results with final exam scores, fraction correct on first attempts and fraction correct after all attempts was determined for every student at intervals corresponding to cumulative deciles. Cumulative deciles were defined as having taken the first 10% of quiz questions in the course, the first 20% of quiz questions in the course, etc. up to 100% of the quiz questions in the body of the course (excluding the final exam, which is the outcome variable). Figure 3 shows the correlation of intermediate assessment scores with final exam scores versus the percentage of formative assessments completed (shown at each cumulative decile), using fraction correct on first attempt and fraction correct after all attempts.

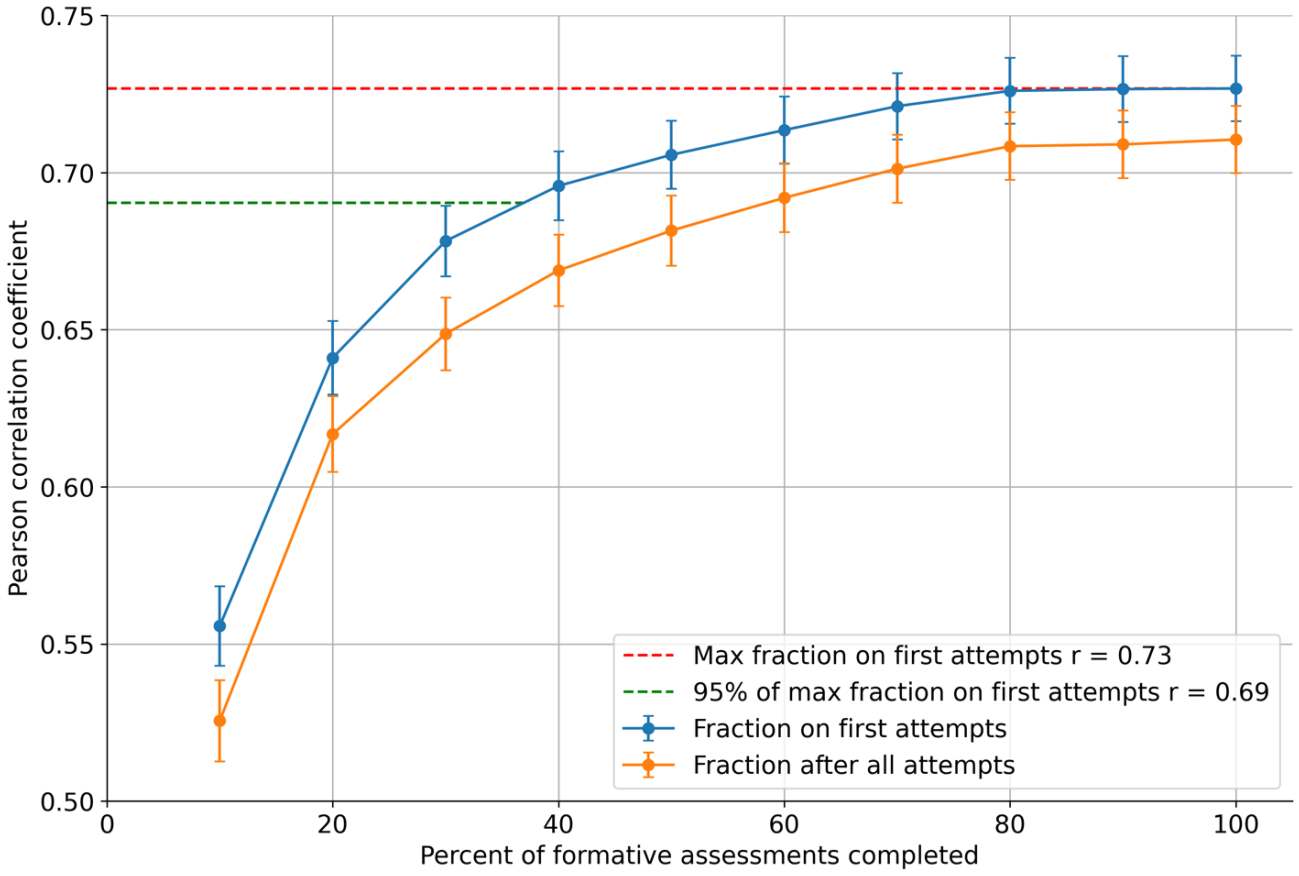

Figure 3. Correlations between fraction correct on first assessment attempt, fraction correct after all attempts, and final exam scores across all 127 courses in the study. Error bars indicate 95% confidence intervals. Note that the y-axis runs from 0.5 to 0.75 in order to better show differences between the curves. The maximum r value is 0.73 and 95% of that is indicated by the lower dashed line.

In Figure 3, the correlation between fraction on first attempts and final exam score crossed the 95% threshold (95% of the maximum r value) after 40% of formative assessments were completed. Because data were collected at decile intervals, the point corresponding to the larger decile is used when the threshold line intersects between two points.

Fraction on first attempts correlation was significantly different than fraction after all attempts correlation at every decile based on the z-statistic ($p = 0.0002$ at 10%, $p = 0.0005$ at 20%, $p < 10^{-5}$ at 30%, $p < 10^{-5}$ at 40%, $p = 4.5 * 10^{-5}$ at 50%, $p = 0.0002$ at 60%, $p = 0.0005$ at 70%, $p = 0.002$ at 80%, $p = 0.002$ at 90%, $p = 0.004$ at 100%). Given that fraction on first attempts was a monotonically better predictor (higher r value) than fraction after all attempts, we focus on that for the remainder of this paper. However, we also include fraction after all attempts, given that this may be the value most easily accessible to educators in their courses.

These results suggest that fraction on first attempts is strongly correlated with final exam scores, even after only approximately 40% of formative assessment in a course has been completed. The results also indicate that fraction on first attempts explains a higher percentage of the variance in final exam scores than using fraction after all attempts, and that both metrics yield diminishing incremental gains in explanation of final exam score variance as more quiz questions are used. Based on the shape of the correlation versus percent of formative assessments completed curve, the results suggest that we get most of the predictive value by the time we have included approximately 40% of formative assessment.

ISSN 1929-7750 (online).

3.2. Threshold for Prediction: Absolute Number of Quiz Questions Versus Percentage

Given that courses may vary significantly in length and number of quiz questions, we explored the question of whether the predictive threshold — 95% of the maximum correlation between intermediate formative assessment scores and final exam scores (shown in Figure 3 above) — is reached after a student has taken an *absolute number* of quiz questions or after a *percentage* of the total number of quiz questions in a course.

Courses differed in number of assessment questions, with Fundamentals courses having approximately twice the number of quiz questions as Pro courses (Table 3). To better compare the correlation relationship by percentage versus absolute number of quiz questions, we chose courses that had at least 500 completers over the study period. Ten of the 15 courses were included on that basis (as shown in Figure 4 below and Table A1).

Figure 4A shows the correlation coefficient versus the absolute number of formative assessments completed and Figure 4B shows the correlation coefficient versus the percentage of formative assessments completed. In Table A1, the 95% thresholds (criteria defined in the Methods section) are shown for each course by absolute number of formative assessments completed and by percent of formative assessments completed.

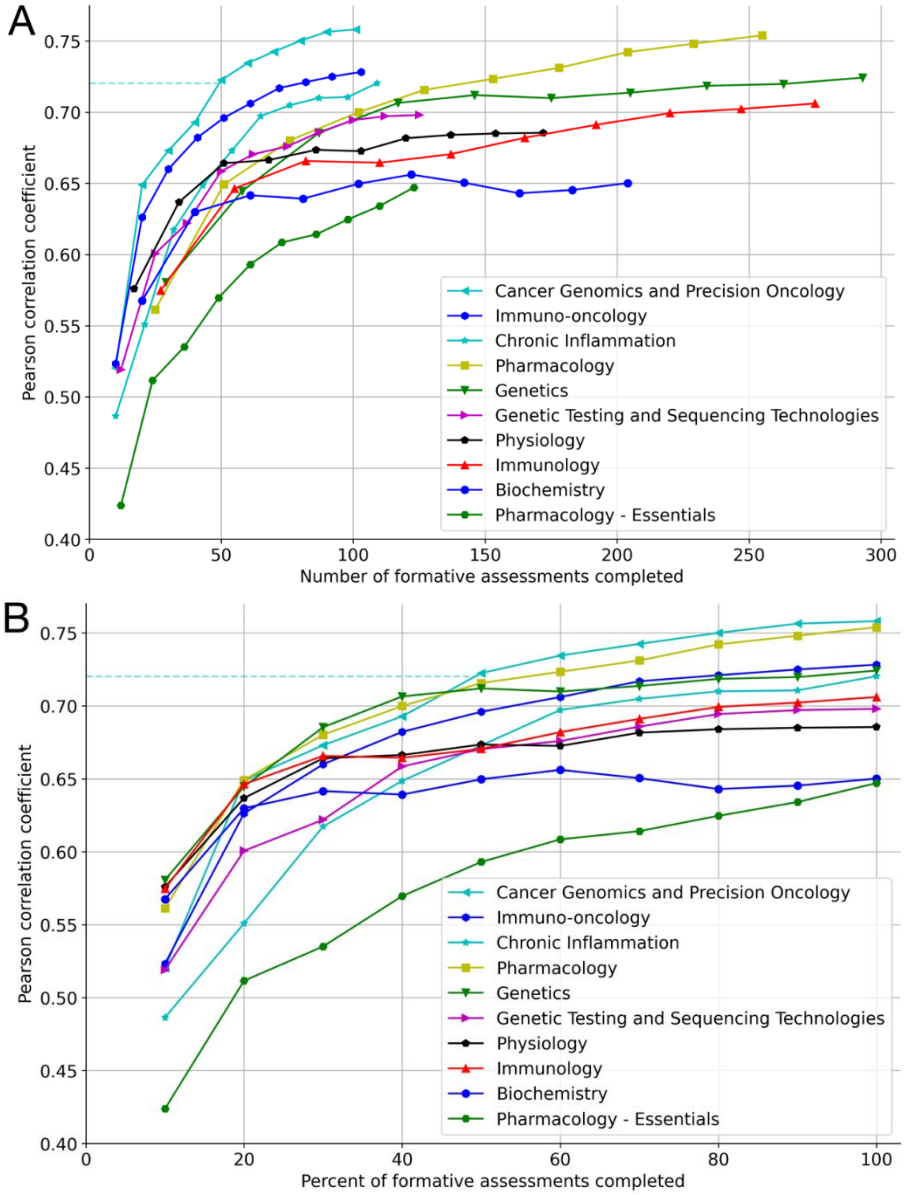

Figure 4. (A) Correlation of fraction on first attempts to final exam score versus absolute number of formative assessment questions taken for 10 courses with at least 500 completers; (B) Correlation of fraction on first attempts to final exam score versus percentage of formative assessment questions taken for 10 courses with at least 500 completers. For each panel, an example of the 95% of maximum r value threshold for one course (Cancer Genomics and Precision Oncology) is shown as a horizontal dashed line.

ISSN 1929-7750 (online).

Figure 5 compares the distribution of the absolute number of assessment questions needed to reach the 95% threshold versus the distribution for percent of assessment questions needed to reach the 95% threshold. The coefficient of variation for these thresholds based on the absolute number of formative assessments is 0.54 and the coefficient of variation based on the percent of formative assessment questions is 0.34. This suggests that there is greater relative variability in the absolute values compared to the percentages.

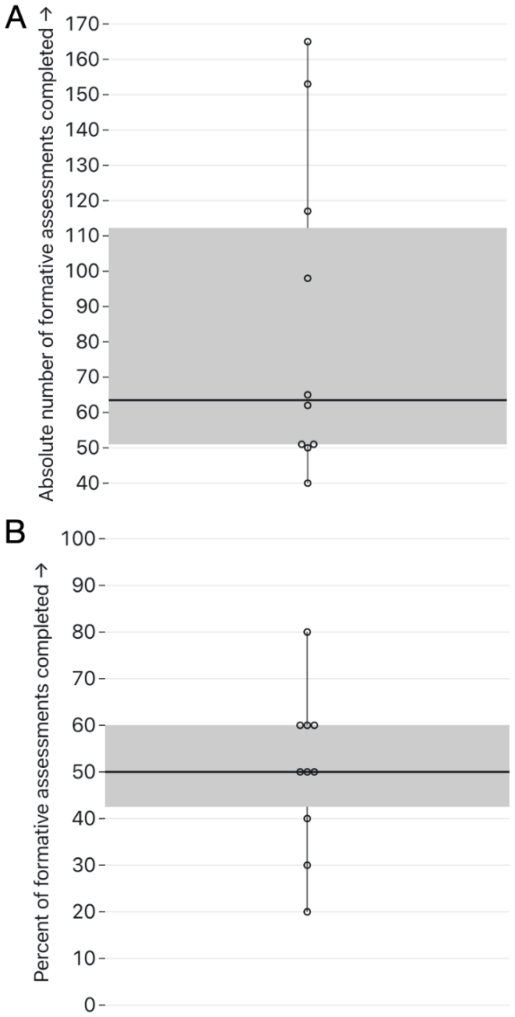

Figure 5. Based on the fraction after first attempts, the absolute number of formative assessments needed to reach the 95% threshold (A) and percent of formative assessments needed to reach the 95% threshold (B). The boxplots show the median and interquartile range. All data points are shown.

These results indicate that the threshold for correlation is more closely based on a percent of total amount of assessment than on an absolute number of assessment questions. This observation holds for both relatively short (Pro courses) and longer courses (Fundamentals).

3.3. Dependence on Using a Sequential Subset of Quiz Questions versus a Random Subset

We explored the question of whether the correlation results visualized in Figure 3 are related to the sequencing of questions in courses — in other words, is it necessary to determine correlation with final exam scores based on the *first* x% of questions, or can it be *any* x% of questions in a course. To do this, we analyzed the largest grouping of courses that had shared assessment questions in the same order and with the same set of questions on the final exam. Eight Immunology course instances shared assessment questions in this way and therefore allowed for direct comparison. There were 2,071 completers across these Immunology course instances.

The quiz questions were randomized into 10 different orders, using random seeds, and the correlations with the final exam scores were calculated for the cumulative deciles for each random order. These correlations for each decile across the 10 random orderings were then averaged. We followed this procedure to minimize the possibility that a particular random ordering might display favourable (or unfavourable) correlations. Results are shown in Figure 6.

ISSN 1929-7750 (online).

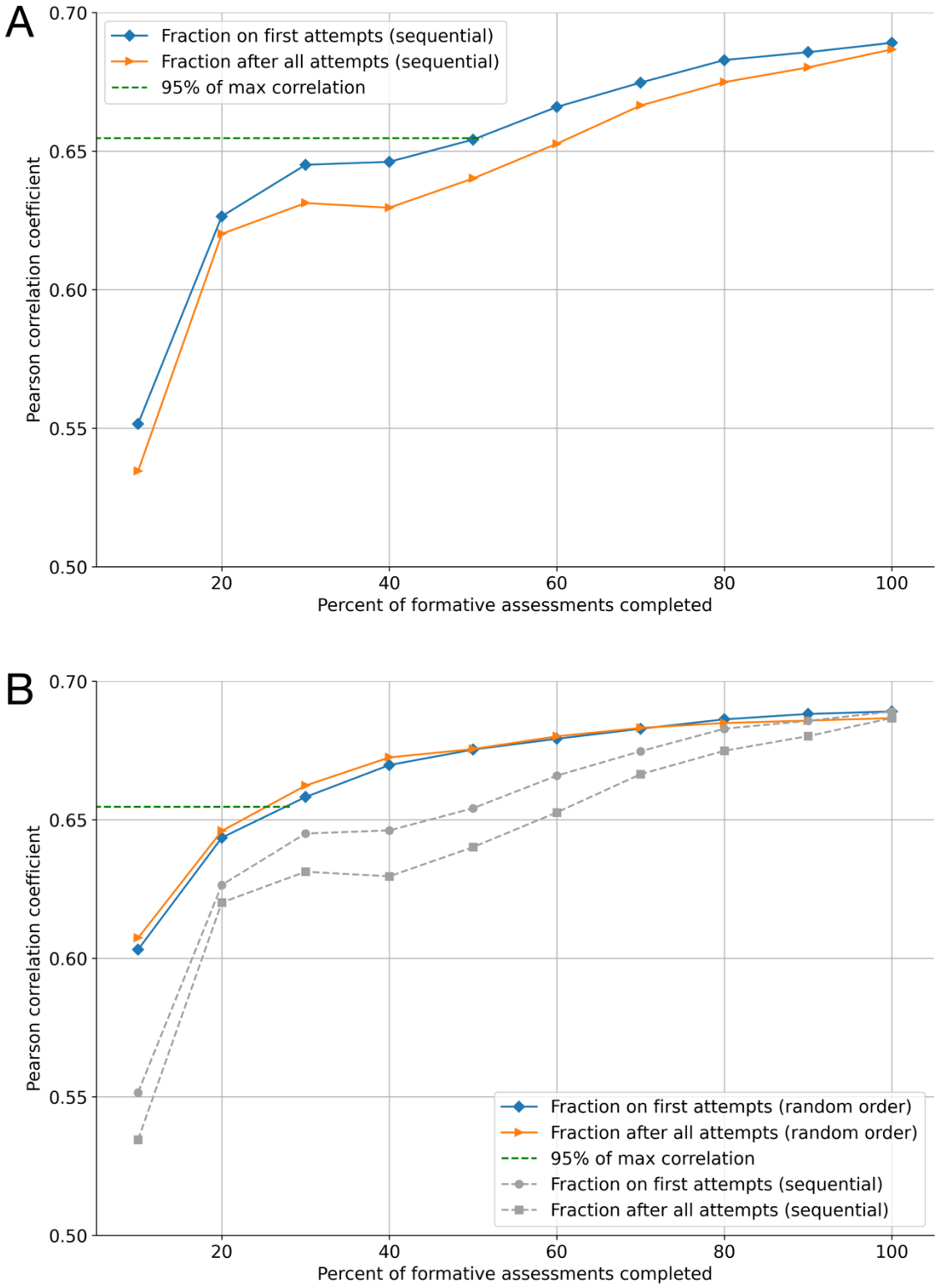

Figure 6. (A) Correlation for fraction on first attempts and fraction after all attempts, non-randomized (with questions in sequential order), for eight IMMU instances that have questions in common, (B) random ordering (10 random orderings with correlations at each cumulative decile average) for fraction on first attempts and fraction after all attempts, randomized and non-randomized comparison, for eight IMMU instances that have questions in common. Non-randomized (same plot as A) is displayed grayed out for reference in B. Horizontal dashed lines indicate the 95% threshold for fraction on first attempts.

ISSN 1929-7750 (online).

The 95% of maximum r value threshold for fraction on first attempts correlation is reached at 50% of formative assessments completed for the questions taken sequentially (Figure 6A) and at approximately 30% of formative assessments completed for the questions taken in random order (Figure 6B).

Fraction on first attempts correlation was significantly different for the randomized order questions versus the sequential order questions at three deciles ($p < 10^{-5}$ at 10%, $p = 0.0648$ at 20%, $p = 0.1748$ at 30%, $p = 0.0037$ at 40%, $p = 0.0091$ at 50%, $p = 0.1386$ at 60%, $p = 0.5209$ at 70%, $p = 1.2684$ at 80%, $p = 1.4474$ at 90%, $p = 2.0000$ at 100%), with randomized monotonically better and reaching $p < 0.05$ significance levels at deciles 10%, 40%, and 50%.

Fraction after all attempts correlation was significantly different for the randomized order questions versus the sequential order questions at seven deciles ($p < 10^{-5}$ at 10%, $p = 0.0025$ at 20%, $p = 0.0001$ at 30%, $p < 10^{-5}$ at 40%, $p < 10^{-5}$ at 50%, $p = 0.0004$ at 60%, $p = 0.0435$ at 70%, $p = 0.3259$ at 80%, $p = 0.8693$ at 90%, $p = 2.0000$ at 100%), with randomized monotonically better and reaching $p < 0.05$ significance levels at deciles 10%, 20%, 30%, 40%, 50%, 60%, and 70%.

The results suggest that taking a random subset of questions achieves a higher correlation with final exam scores. Although this difference does not reach significance for all deciles, using the randomized version is monotonically better for all deciles than using the sequentially taken questions.

4. Discussion

Formative assessment, particularly through automatically graded questions like MCQs, has become integral to modern courses (Einig, 2013). Given the prevalence of quizzes, understanding their predictive power for student outcomes helps educators anticipate and address learning challenges. While past studies have demonstrated a strong link between overall formative assessment performance and final exam scores, our work extends this analysis by focusing on *intermediate* formative assessment results (Bulut et al., 2023). This provides a more dynamic perspective, exploring how early in a course such predictions become reliable.

Our research explored this relationship by focusing on three key questions: 1) How much formative assessment is needed to provide robust predictive power for final exam scores? 2) What is the relationship between intermediate formative assessment results and final exam scores? and 3) Is the predictive power of formative assessment based on an absolute question count or relative to overall course content? By analyzing a large dataset encompassing over 20,000 enrollments and courses with 100–300 formative assessment questions, we aimed to quantify the relationship between intermediate quiz performance and final exam scores, examining this across diverse course types and lengths.

Our key findings reveal that, on average, the correlation coefficient between formative assessment scores and final exam scores reaches a threshold (95% of its maximum value) after students complete only about 40% of the quiz questions (Figure 3). While individual courses showed some variation (Figure 4), this 40–60% threshold held consistently when viewed as a percentage of total quiz questions, regardless of course length or type. This percentage-based relationship supports the hypothesis that a representative sample of formative assessment, covering roughly half of the course content, captures a substantial portion of the variance in final exam performance.

Furthermore, we investigated whether the *sequence* of quiz questions influences predictive power. Surprisingly, using a random sample of quiz questions yielded even *stronger* correlations with final exam scores than using sequentially ordered quiz results (Figure 6). This suggests that a random sample might better reflect the comprehensive nature of a final exam, which typically covers material from across the entire course. Sequential quizzing, on the other hand, can over- or under-emphasize the difficulty of specific lessons based on their position in the course. Randomization mitigates this bias by sampling questions more evenly across the course content. However, in practice, teachers typically observe student performance sequentially (with the potential exception of certain online courses that release all material at once, including formative assessments). Therefore, Figure 3 provides a realistic representation of the predictive power available to educators as a course progresses.

The concept of a learning analytics loop proposes multiple stages — from collecting student data, through analysis and insight generation, to implementing pedagogical changes (Ferguson et al., 2023; Motz et al., 2023). Our study focuses on the foundational stages of this cycle: data collection from formative assessments, measurement of student performance patterns, and analysis of their predictive relationships with final outcomes. Our finding that a strong predictive correlation emerges after approximately 40% of formative assessment completion provides a quantitative foundation for subsequent stages of the learning analytics cycle, particularly the design of early intervention systems and optimization of assessment strategies. While our analysis focuses primarily on establishing the predictive relationship between formative assessment and final outcomes, it serves as a methodological bridge toward closing the learning analytics loop by enabling evidence-based implementation of systems that balance prediction accuracy with intervention timeliness.

These findings have significant implications for educational practice, particularly in the context of adaptive learning systems, which offer an individualized and finely tuned way of intervening to alter outcomes. The ability to make reliable

predictions using early formative data enables more sophisticated approaches to personalized learning pathways and targeted interventions. By identifying at-risk students earlier in their learning journey, educators can implement interventions before students fall critically behind.

Interventions based on predictive insights can take various forms. For example, a tiered approach could involve flagging students whose quiz performance falls below a predetermined threshold and then providing them with personalized feedback, remedial materials, or one-on-one tutoring sessions to address specific gaps (Arnold & Pistilli, 2012). In parallel, technology-enhanced adaptive learning systems can automatically adjust content difficulty and sequence, ensuring that each student practises foundational concepts until mastery is achieved (Chi et al., 2011). Additionally, integrating peer-led study groups or mentoring programs offers a collaborative framework where at-risk students can benefit from the guidance and support of peers who have demonstrated stronger performance, potentially increasing engagement and improving outcomes (Springer et al., 1999; Topping, 2005).

Furthermore, understanding the quantity of assessment needed for reliable prediction also empowers educators to optimize assessment design, balancing predictive validity with student workload. This knowledge enables formative assessment to be used as a dynamic instrument for predicting and improving student outcomes. Specifically, our results suggest the following:

Predictive Power of Intermediate Assessments: A significant correlation between formative and summative assessment emerges early in the course, after only about 40–60% of the formative assessments are completed.

Percentage-Based Threshold: This predictive threshold is relative to the total number of formative assessment questions, not the absolute count, suggesting generalizability across courses of varying lengths.

Benefit of Randomization: Randomly sampling formative assessment questions may offer even stronger predictive power than using sequentially administered quizzes, potentially by creating a more representative sample of overall course content.

Implications for Intervention: The early emergence of predictive power allows for timely identification of struggling students and implementation of targeted support strategies.

Assessment Design Optimization: These findings can inform more efficient assessment design, balancing the need for robust prediction with minimizing student workload and maximizing learning value.

While the correlation between intermediate formative assessment and final exam scores is strong, it is not perfect. Several factors might contribute to this. Students may approach low-stakes, multiple-attempt formative assessments differently than high-stakes, single-attempt final exams, leading to differences in effort and performance (Figure 2). Variations in question difficulty between formative assessments and the final exam, despite being authored by the same instructor, could also play a role. Finally, the close proximity of formative quiz questions to the relevant course material might make them easier than final exam questions, which assess more distant learning.

Our findings also inspire important questions about the role and necessity of summative assessments. While our results show that formative assessment performance can robustly predict final exam scores even with a relatively small number of assessments completed, summative assessments can serve a complementary role. When used well, they offer an evaluation of cumulative learning, ensure that the integration and retention of knowledge over time is assessed, and provide standardized benchmarks. Our findings about the predictive power of formative assessment, combined with ongoing innovations in formative assessment design, may contribute to discussions about optimizing the balance between formative and summative assessment. Rather than undermining one another, finding the right blend has the potential to create a more nuanced and effective evaluation ecosystem.

5. Limitations

Several factors may limit the generalizability of these findings. Our study focused on high-quality, online biomedical science courses with consistent design principles, including frequent formative assessments and a summative final exam. The results may not generalize to courses with different structures, such as those with infrequent assessments or a greater emphasis on summative evaluation. Courses significantly shorter than those in our study (10 weeks for Fundamentals, five weeks for Pro courses) might also exhibit different dynamics, since the role of cumulative assessment may change when there are few assessments or the course duration is very short (e.g., a single session). In addition, we focused on students who completed the courses, given that those were the ones with final exam scores. Most non-completers in our study disengaged after having completed either no formative assessments or a small number. Therefore, generalization to this population may require other techniques. Future research across diverse educational contexts and course designs is needed to investigate these limitations and confirm the generalizability of our findings.

6. Concluding Remarks

The widespread adoption of formative assessment in education necessitates a deeper understanding of its predictive capabilities. This large-scale study extends prior work by examining the dynamic relationship between *intermediate* formative assessment results and final exam scores. Our findings demonstrate that a substantial portion of predictive power is achieved relatively early in a course, using only about half of the available formative assessment questions. This insight provides valuable information for educators, enabling earlier and more targeted interventions for at-risk students. Moreover, the finding that randomized quiz samples enhance predictive power raises important questions about assessment design and the optimal distribution of quiz questions throughout a course. Future work should explore these questions further and investigate how these findings translate to other learning environments and assessment modalities. While this study focuses on multiple-choice quizzes as a proxy for formative assessment, it is important to recognize that these represent only one approach among many. Other formats — such as open-ended responses, peer assessments, and interactive tasks — can provide nuanced insights into student learning processes. With the advent of large language models, future research may harness automated analysis of unstructured data to further enhance formative assessment practices. By continuing to refine our understanding of formative assessment dynamics, we can empower educators to make data-informed decisions that maximize student learning and success.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

Acknowledgements

The authors would like to thank the Harvard Medical School HMX curriculum team and HMX faculty for their creation of the assessment questions studied.

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In S. Dawson, C. Haythornthwaite, S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), *LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267–270). ACM Press. https://doi.org/10.1145/2330601.2330666

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594x.2010.513678

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Open University Press. https://oro.open.ac.uk/24157/

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bloom, B. S. (1968). *Learning for mastery*. Regional Education Laboratory for the Carolinas and Virginia. https://files.eric.ed.gov/fulltext/ED053419.pdf

Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., Lai, K. W., & Shin, J. (2023). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology, 54*(1), 19–39. https://doi.org/10.1111/bjet.13276

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3): 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21*(1–2), 137–180. https://doi.org/10.1007/s11257-010-9093-1

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683–695. https://doi.org/10.1080/13562517.2013.827653

Dawson, S., Gašević, D., Siemens, G., & Joksimović, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. In M. Pistilli, J. Willis, D. Koch, K. Arnold, S. Teasley, & A. Pardo (Eds.), *LAK '14: Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 231–240). ACM Press. https://doi.org/10.1145/2567574.2567585

Einig, S. (2013). Supporting students' learning: The use of formative online assessments. *Accounting Education, 22*(5), 425–444. https://doi.org/10.1080/09639284.2013.803868

Ferguson, R., Khosravi, H., Kovanović, V., Viberg, O., Aggarwal, A., Brinkhuis, M., Buckingham Shum, S., Chen, L. K., Drachsler, H., Guerrero, V. A., Hanses, M., Hayward, C., Hicks, B., Jivet, I., Kitto, K., Kizilcec, R., Lodge, J. M., Manly, C. A., Matz, R. L., … Yan, V. X. (2023). Aligning the goals of learning analytics with its research scholarship: An open peer commentary approach. *Journal of Learning Analytics, 10*(2), 14–50. https://doi.org/10.18608/jla.2023.8197

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5–6), 304–317. https://doi.org/10.1504/ijtel.2012.051816

Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment and Evaluation in Higher Education, 45*(6), 842–854. https://doi.org/10.1080/02602938.2019.1682118

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Glick, D., Cohen, A., Festinger, E., Xu, D., Li, Q., & Warschauer, M. (2019). Predicting success, preventing failure. In D. Ifenthaler, D.-K. Mah, & J. Y.-K. Yau (Eds.), *Utilizing learning analytics to support study success* (pp. 249–273). Springer Cham. https://doi.org/10.1007/978-3-319-64792-0_14

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, *15*(3), 42–57. https://www.jstor.org/stable/jeductechsoci.15.3.42

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://www.routledge.com/Visible-Learning-A-Synthesis-of-Over-800-Meta-Analyses-Relating-to-Achievement/Hattie/p/book/9780415476188

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hrastinski, S. (2019). What do we mean by blended learning? *TechTrends*, *63*(5), 564–569. https://doi.org/10.1007/s11528-019-00375-5

Johnson, L., Adams Becker, S., Estrada, V., & Freeman, A. (2015). *NMC horizon report: 2015 higher education edition*. The New Media Consortium. https://library.educause.edu/resources/2015/2/2015-horizon-report

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Knight, S., Buckingham Shum, S., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics, 1*(2), 23–47. https://doi.org/10.18608/jla.2014.12.3

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 31–40. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *Journal of Computer Assisted Learning*, *33*(6), 562–574. https://doi.org/10.1111/jcal.12200

Moskal, P. D., Dziuban, C. D., & Picciano, A. G. (Eds.). (2023). *Data analytics and adaptive learning*. Routledge. https://doi.org/10.4324/9781003244271

Motz, B. A., Bergner, Y., Brooks, C. A., Gladden, A., Gray, G., Lang, C., Li, W., Marmolejo-Ramos, F., & Quick, J. D. (2023). LAK of direction: Misalignment between the goals of learning analytics and its research scholarship. *Journal of Learning Analytics*, *10*(2), 1–13. https://doi.org/10.18608/jla.2023.7913

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64. https://doi.org/10.1080/03098770601167922

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

OECD. (2005). *Formative assessment: Improving learning in secondary classrooms*. OECD Publishing, Paris. https://doi.org/10.1787/9789264007413-en

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24*(3), 355–367. https://doi.org/10.1007/S10648-012-9201-3

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/bf00117714

Scriven, M. (1966). *The methodology of evaluation*. Social Science Education Consortium. https://eric.ed.gov/?id=ED014001

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. In S. Dawson, C. Haythornthwaite, S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), *LAK '12: Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 4–8). ACM Press. https://doi.org/10.1145/2330601.2330605

Sortwell, A., Trimble, K., Ferraz, R., Geelan, D. R., Hine, G., Ramirez-Campillo, R., Carter-Thuiller, B., Gkintoni, E., & Xuan, Q. (2024). A systematic review of meta-analyses on the impact of formative assessment on K–12 students' learning: Toward sustainable quality education. *Sustainability, 16*(17), Article 7826. https://doi.org/10.3390/su16177826

Spector, J. M., Ifenthaler, D., Sampson, D., Yang, L., Mukama, E., Warusavitarana, A., Dona, K. L., Eichhorn, K., Fluck, A., Huang, R., Bridges, S., Lu, J., Ren, Y., Gui, X., Deneen, C. C., San Diego, J., & Gibson, D. C. (2016). Technology enhanced formative assessment for 21st century learning. *Journal of Educational Technology & Society, 19*(3), 58–71. https://www.jstor.org/stable/jeductechsoci.19.3.58

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research, 69*(1), 21–51. https://doi.org/10.3102/00346543069001021

Tempelaar, D. T., Heck, A., Cuypers, H., van der Kooij, H., & van de Vrie, E. (2013). Formative assessment and learning analytics. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *LAK '13: Proceedings of the third international conference on learning analytics and knowledge* (pp. 205–209). ACM Press. https://doi.org/10.1145/2460296.2460337

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157–167. https://doi.org/10.1016/j.chb.2014.05.038

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology, 25*(6), 631–645. https://doi.org/10.1080/01443410500345172

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, *39*(3), 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

West, D., Heath, D., & Huijser, H. (2016). Let's talk learning analytics: A framework for implementation in relation to student retention. *Online Learning*, *20*(2). https://doi.org/10.24059/olj.v20i2.792

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *LAK '13: Proceedings of the third international conference on learning analytics and knowledge* (pp. 145–149). ACM Press. https://doi.org/10.1145/2460296.2460324

Yeh, S. S. (2010). Understanding and addressing the achievement gap through individualized instruction and formative assessment. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 169–182. https://doi.org/10.1080/09695941003694466

Appendix

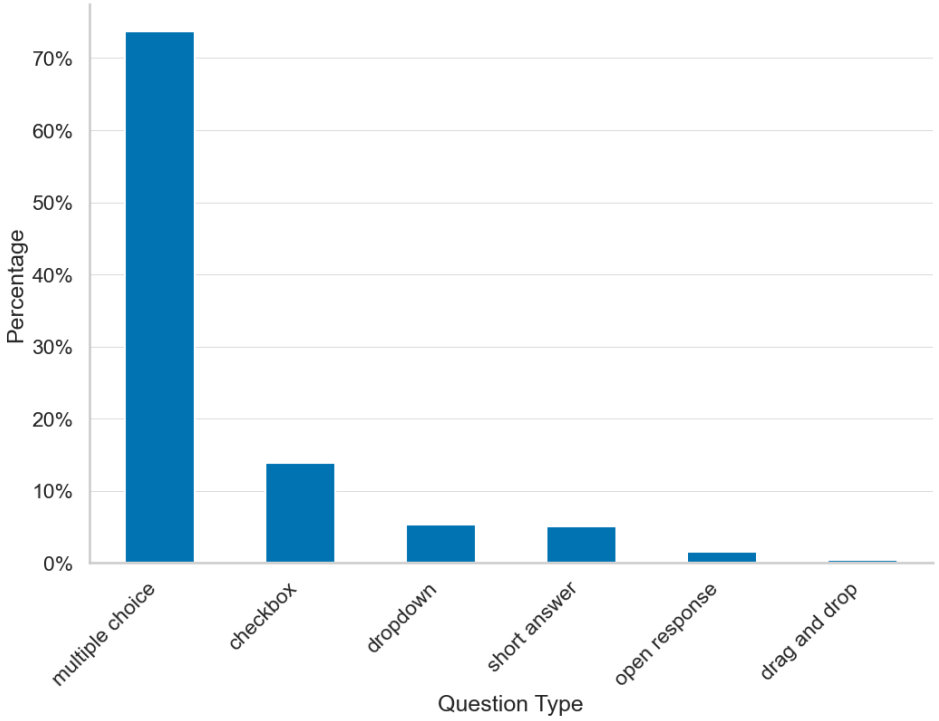

Figure A1. Breakdown of course questions by type. Checkbox is a form of MCQ that allows for multi-select. Short answer and open response questions allow freeform input.

Table A1. The Number and Percent of Questions in Each Course Type Where the Correlation Reaches a Threshold (Greater Than or Equal to 95% of the Maximum Correlation Value)

Course Type	Threshold for Fraction After First Attempts (absolute number of formative assessments completed)	Total Number of Formative Assessments in the Course Lessons	Threshold for Fraction After First Attempts (percent of formative assessments completed)
Cancer Genomics and Precision Oncology	50	101	50
Immuno-oncology	51	103	50
Novel Therapies for Chronic Inflammation, Autoimmunity, and Allergy (Chronic Inflammation)	65	109	60
Pharmacology	153	255	60
Genetics	117	293	40
Genetic Testing and Sequencing Technologies	62	125	50
Physiology	51	172	30
Immunology	165	275	60
Biochemistry	40	204	20
Pharmacology - Essentials	98	123	80