



 Imad S. Achkar¹⁺

 Bassem Kandil²

^{1,2}Faculty of Educational Sciences, Saint Joseph University, Beirut, Lebanon.

¹Email: imad.achkar1@net.usj.edu.lb

²Email: bassem.kandil1@usj.edu.lb



(+ Corresponding author)

Article History

Received: 22 December 2025

Revised: 27 March 2026

Accepted: 4 June 2026

Published: 23 July 2026

Keywords

Authentic assessment
Higher-order thinking
High-stakes testing
Lebanese baccalaureate
Secondary science education
Surface learning
Washback.

ABSTRACT

The predictable nature of the Baccalaureate Biology Exam in Lebanon may lead to "teaching to the test" and an emphasis on developing format-specific proficiency instead of transferable conceptual understanding. Lebanese research has not provided a direct assessment of whether high performance on the traditional Baccalaureate biology exam reflects true understanding or proficiency in responding to the exam's format. This study provides direct performance evidence by examining whether students who perform well on the traditional exam can apply their knowledge to alternative formats with differing cognitive demands. Using a quasi-experimental within-subjects design, 40 Grade 12 students from three Lebanese schools completed Test A (the traditional exam) and Test B (an alternate version with multiple choice, application-based, and analytical tasks). Students performed significantly better on Test A ($M = 85.12\%$, $SD = 8.46\%$) than on Test B ($M = 64.75\%$, $SD = 12.12\%$), with a 20.38-point difference, $t(39) = 9.692$, $p < 0.000001$, and a very large effect size ($d = 1.532$). School-level analysis showed significant variation in performance gaps, $F(2,37) = 11.430$, $p = 0.0001$. The correlation between Test A and Test B was low ($r = 0.203$, $p = 0.210$), suggesting that the two versions assessed different skills despite the same content. Overall, the findings suggest that performance on the traditional Baccalaureate biology exam may reflect familiarity with the exam format rather than strong knowledge transfer, and that the performance gap varies significantly across schools.

Contribution/Originality: This study is the first performance-based Lebanese comparison of Grade 12 biology achievement across traditional Baccalaureate items versus a reformulated assessment that holds content constant while changing format and cognitive demand. It also shows that the size of the cross-format performance gap differs significantly by school, suggesting that instructional approaches may mediate transfer.

1. INTRODUCTION

1.1. Background and Context

The Educational System in Lebanon takes place in a complex socio-cultural environment of language diversity, religious plurality and an examination structure for secondary education that was established during the French Mandate period. The Official Baccalaureate Examination, conducted by the Ministry of Education and Higher Education (MEHE), provides the primary high-stakes assessment at the end of secondary schooling and significantly influences students' ability to enter university programs and academic trajectory thereafter. Biology is one of the heaviest-weighted subject areas in the LS track of the Baccalaureate and is commonly viewed by students, parents, teachers, and school administrators as being a key indicator of students' overall success in the Baccalaureate.

Therefore, performance in biology is directly related to both students' academic success and institutionally related factors, i.e., school reputation and competition among private schools. For over a number of years, the format of the Biology Examination has remained relatively stable and consistent. Each test session generally includes four written assignments on genetics, physiology, molecular biology and ecology/evolution, with common prompts like "pick out," "show that" and "indicate." While the topics covered are important to biological literacy, the repetition of the same types of written assignments over time and similar prompts has resulted in a preparation culture that is based on repeated review of previous examination papers. As a result, in some schools, the instructional focus has shifted from focusing on understanding concepts and processes to repetitive practice of sample questions and expected response patterns. In many cases, the organization of curriculum pacing (Ministry of Education and Higher Education, 1997), lesson plans, and assessment schedules for students are focused on preparing them for anticipated examination formats. While these approaches can lead to high pass rates, there are concerns that student success may be based on mastering the predictable responses rather than developing a flexible and conceptual understanding of biological processes.

1.2. The Washback Effect of High-Stakes Assessments

Research in Assessment Theory indicates that high-stakes assessments can influence instruction not only by measuring what is learned but by shaping how teachers teach and how students learn. The term for this type of effect - washback (or backwash) - refers to the impact of assessments on curriculum design, pedagogical practice and learning strategy development (Alderson & Wall, 1993; Hughes, 1989). When assessment items consistently emphasize certain formats or limit the cognitive processes required to complete them, teachers and students will reasonably adjust their behavior to maximize their performance within those constraints. Stobart (2008) describes this dynamic as a "testing trap," where individuals act rationally based on institutional incentives but collectively undermine broader educational goals such as deep conceptual understanding, application of knowledge, and critical thinking. Additionally, predictable science examinations tend to encourage surface-level learning strategies, such as memorizing model answers, recognizing patterns, and reproducing procedures, rather than integrating concepts and applying scientific reasoning. Although these dynamics have been reported internationally, their empirical impact on student learning in the Lebanese context has not been sufficiently investigated.

1.3. Problem Statement

Given the consistently high success rates on the Lebanese Baccalaureate biology examination, there is increasing concern among educators and researchers that these results do not accurately reflect students' biological competence. Rather, student success on the examination may be the result of prolonged exposure to repetitive task formats and extensive practice on specimen questions. If student success is primarily due to familiarity with the format of the examination, then current Baccalaureate scores may inflate the representation of students' ability to transfer knowledge to new situations and deflate their abilities to engage in higher-level analytical reasoning, knowledge integration, and application of knowledge in unfamiliar contexts. If so, then the Baccalaureate examination functions as a measure of test-specific procedural competency, rather than a valid indicator of biological literacy.

These possibilities raise significant questions about the interpretative validity of Baccalaureate results and their use in making high-stakes decisions, such as university admissions and program placements.

1.4. Research Gap and Study Rationale

While previous studies conducted in Lebanon have analyzed the content of the examination and documented the predominance of lower-order cognitive demands in the examination, no previous research has directly investigated whether students who achieve high performance on the traditional Baccalaureate biology examination are capable of applying the same knowledge when the format and cognitive demands of the assessment are changed. Much of the

existing research either infers learning outcomes from an analysis of the questions used on the examination or from a report of teachers, thereby leaving unanswered whether high performance on the traditional Baccalaureate biology examination reflects a genuine understanding of the concepts being examined or is based upon the students' experience with the specific format of the examination. Without direct performance-based evidence, the extent to which current assessment practices foster the development of transferable scientific knowledge or merely enhance student performance within the narrow framework of a particular testing format is unknown.

The purpose of the current study is to address this gap by empirically investigating student performance on two assessments that include the same biological content but are formatted differently and require different levels of cognitive demands. By keeping the subject matter constant and varying the format of the assessment, the study isolates the role of familiarity with the format of an assessment in affecting student performance and provides direct evidence of the degree to which Baccalaureate success is representative of a robust conceptual understanding of biological processes versus a specialized proficiency in responding to the format of the Baccalaureate examination.

1.5. Research Questions

This study addresses the following research questions.

RQ1. How do Grade 12 students in the Lebanese Life Sciences track perform on traditional Baccalaureate-style biology assessments compared with reformulated assessments testing identical content through different question formats and cognitive demands?

RQ2. What does the magnitude and pattern of any observed performance differences reveal about the nature of learning fostered by current Baccalaureate preparation practices?

RQ3. To what extent do performance gaps vary across schools, and what might such variation suggest about the role of instructional approaches in mediating assessment effects?

1.6. Research Hypotheses

H₁: Students will achieve significantly higher scores on the traditional Baccalaureate-format biology assessment than on the reformulated assessment, despite identical biological content.

H₂: The performance difference between formats will be large in magnitude, indicating that traditional examination success reflects, in part, format-specific test-taking proficiency rather than exclusively transferable conceptual understanding.

H₃: The size of the performance gap will differ significantly across schools, with institutions that emphasize conceptual instruction and varied assessment practices exhibiting smaller declines under the reformulated format.

1.7. Structure of The Paper

This section of the paper will continue as follows. Section 2 will review theoretically and empirically existing literature regarding the washback effect on science education, as well as assessment practices. Section 3 will detail the quasi-experimental methodological strategy used in this study, including participant characteristics and assessment tools. Section 4 will outline the quantitative statistical comparisons between student achievement under each type of assessment tool. Section 5 will describe the results of the study and relate them to the theoretical frameworks and research internationally. Section 6 will discuss the implications of the study for assessment policy and practice, and outline the potential avenues of research for further investigation.

2. LITERATURE REVIEW

2.1. Theoretical Background of Washback

Washback refers to the way in which institutional assessment impacts teachers' practices. As such, it can now encompass the whole field of language testing research, including all assessment contexts where high stakes exist. Alderson and Wall (1993) wrote one of the first research papers on washback. They hypothesized that assessments

can have an effect on curricula, pedagogies, and learning; however, the nature of this effect could be highly complex, and should therefore be researched empirically, rather than being taken for granted. Furthermore, they produced a model outlining different dimensions of assessment effects. These dimensions include: the entities affected (e.g., students, teachers, administrators); the process versus the product; the content emphasis; and the attitudinal impact. Messick (1989) and Messick (1996) outlined a broader theoretical framework for understanding consequential validity. He stated that comprehensive validity includes not only construct validity (the degree to which tests measure what they intend to measure) but also value implications and social consequences of assessment use. High-stakes exams, which emphasize the use of predictable formats, may lead to negative consequences, including curriculum narrowing, the adoption of pedagogical methods focused on exam preparation, and the use of shallow learning strategies, even if the exams are valid measures of recall ability.

2.2. International Research on the Backwash Effects of Examinations

In addition to the negative relationship demonstrated between assessment and instructional practices, there is considerable evidence of this relationship being observed globally. Similarities in the results of the studies conducted in Morocco (Belkbir, 2023), Jordan (Allouh, 2016), and Pakistan (Mehmood, Iqbal, Abdullah, & Farooq, 2016) included assessments that emphasized lower order cognition (60-82 % of the items were at the remember/understand level); teachers organized their instruction around previous exams; and there was a trend toward teachers ignoring curriculum objectives that required students to use their analytical/critical thinking skills. The research consistently demonstrates that when surveyed, teachers report very strong positive correlations between the cognitive levels of assessment used and the cognitive levels emphasized in their instructional practices. Virtually all surveys of teachers have shown that the pressure to prepare students for high-stakes exams has led to a decrease in laboratory work and inquiry-based activities. The research also indicates that backwash effects are being documented across a variety of educational systems; therefore, providing empirical support for the notion that backwash effects continue to occur throughout the world despite widespread recognition of the limitations of using high-stakes exams as a pedagogy. Khan, Hassan, and Cheng (2025) examined the impact of school-based assessment reform in Malaysia. Although they noted that the Malaysian government had implemented policies to reduce the pressure on students created by high-stakes exams and promote meaningful learning, they found that policy and school-level factors contributed to the continuation of test-oriented instructional practices among teachers and school administrators. In particular, Khan et al. (2025) study of lower secondary English language assessment found that while the government shifted from high-stakes centralized exams to school-based assessment to eliminate the washback of high-stakes exams, teachers and school administrators still felt pressured to create test preparation environments. Therefore, the findings of Khan et al. (2025) suggest that the format of assessment is only one factor that influences the degree to which backwash occurs; the degree to which systemic and cultural factors influence how assessment reforms are enacted in classrooms is far more important.

Longitudinal studies that have analyzed the cognitive demands of science assessments across long periods of time have consistently shown that lower-order thinking skills are emphasized in science assessments. Azzopardi and Camilleri (2025) analyzed 20 years of Biology and Physics practical exams in Malta (72 examination papers containing 394 Biology and 467 Physics questions) and concluded that there were significant differences in cognitive demands between subjects. Using Bloom's Revised Taxonomy, Azzopardi and Camilleri (2025) found that Biology exams contained a significantly higher percentage of HOCS (Higher-order cognitive skills) questions (70.7%) than Physics exams (10.5%). Furthermore, Physics exams did not contain any "create" level questions (The highest level in Bloom's taxonomy). As such, the findings of Azzopardi and Camilleri (2025) demonstrate that the design of exams directly influences the cognitive processes employed by students, and therefore the cognitive processes emphasized by teachers in their instructional practices. Furthermore, the large difference in cognitive demands between Biology and Physics exams demonstrates that the formats of exams and the types of questions asked can become institutionalized

over time and lead to discipline-specific cognitive emphases that may not align with the curriculum's intended outcomes.

More recently, studies have documented how high-stakes exams have continued to hinder pedagogical reform efforts in other contexts more closely related to the educational system of Lebanon. Syed and Butt (2025) conducted a meta-synthesis of multiple studies to examine the nature of classroom assessment practices in secondary education in Pakistan during the 2021-2023 academic year. Their findings indicated that teachers continued to adopt teaching to the test approaches, despite the fact that national curriculum frameworks emphasize critical thinking, problem solving, and inquiry-based learning. Teachers further indicated that the examination pressure created conflicting goals between the curriculum objectives and the reality of the classroom, leading them to make rational choices to prioritize test preparation over conceptual development. This pattern of conflict between policy and assessment-driven instructional practices is consistent with the experiences of Lebanese educational researchers (Achkar, 2024; Shehayeb, 2023).

Rahman, Seraj, Hasan, Namaziandost, and Tilwani (2021) have provided empirical documentation of the washback mechanisms that exist within the Secondary School Certificate (SSC) examination system in Bangladesh. They found that the structure of the exam resulted in an explicitly exam-oriented approach to teaching that eliminated students' opportunities to engage in holistic language learning. Teachers reported that they spent most of their instructional time preparing students to answer specimen questions, organizing their entire curriculum around previously administered exam papers, and purposefully omitting curriculum content that would not be expected to appear on the exam. Students responded to these conditions by engaging in surface learning, focusing on memorizing model answers as opposed to increasing their depth of understanding. Surface learning strategies due to the focus on specimen questions have also been documented in the context of Lebanese Baccalaureate preparation; however, they have not been empirically tested through comparative assessment performance (Achkar, 2024).

Abate and Mishore (2024) analysis of science examinations in Ethiopia provides additional data on the relationship between the cognitive demand of assessments and the instructional practices of teachers. Abate and Mishore (2024) examination of 205 teacher-created test items across three disciplines (Biology, Chemistry, and Physics) found that there was a consistent emphasis on lower-order cognitive skills and significant misalignment between national curriculum learning objectives and the assessment practices actually engaged in by teachers. Using Bloom's Revised Taxonomy, Abate and Mishore (2024) determined that the overwhelming majority of the questions in the assessments targeted the remembering and understanding levels of cognitive demand, with minimal representation of the analyzing, evaluating, and creating levels of cognitive demand. Importantly, they found that the misalignment between the intended curriculum objectives and the actual assessment practices impacted students' performance on national exams. Therefore, their findings indicate that when classroom assessments do not represent the cognitive demands that are articulated in curriculum documents, students develop competencies that will not transfer to performance situations that require authentic demonstration of knowledge - a condition that is directly relevant to the current study's investigation of whether the success of Lebanese Baccalaureate students reflects the format-specific skills of their assessments, or the transferable biological understanding that is the intent of the assessments.

In total, the recent studies demonstrate that negative washback effects are not episodic occurrences, but rather are a systematic pattern of behavior when high-stakes exams emphasize predictable formats and lower-order cognitive skills. The consistency of the findings from studies conducted in Malaysia, Malta, Pakistan, Bangladesh, and Ethiopia - regardless of language, culture, or educational system - suggests that the mechanisms responsible for generating negative washback effects function through common dynamics: examination pressure, format predictability, and the rational responses of teachers and students to assessment-related incentives. These global patterns provide a useful backdrop for understanding the Lebanese Baccalaureate system and indicate that the

concerns expressed by Lebanese educators regarding the potential washback of the Baccalaureate system are likely to be supported by empirical evidence that is consistent with that found in other high-stakes exam contexts.

2.3. Surface-Level Learning and Assessment Format Effects

Marton and Säljö (1976) provided an essential framework for understanding how assessment influences learning processes in their research, distinguishing between surface and deep learning approaches. Students who take a surface approach to learning are primarily concerned with memorizing information so that they can reproduce it, view learning tasks as external requirements to accomplish with minimal effort. Students who take a deep learning approach, on the other hand, attempt to comprehend the underlying meanings of information and establish meaningful associations between current and prior learning. Importantly, Marton and Säljö (1976) established that students' learning approaches are shaped by the perceived demands of assessment tasks. When assessments are designed to elicit only recalling information and not to demonstrate understanding, students rationally adopt surface approaches a finding that is directly pertinent to understanding how predictable examination formats may limit the growth of transferable knowledge.

2.4. The Quality of Assessments in Lebanon: Findings from Research

Lebanese research on assessment quality has shown a number of problems in assessing quality. Vlaardingerbroek, Jaber, Rizk, and Bayoud (2009) identified three major issues in the Brevet examination process. Firstly, they indicated that there was a lack of coverage of all curriculum areas. Secondly, they stated that many questions were ambiguous and did not clearly state what was being asked. Thirdly, they demonstrated that the topics covered were very predictable and allowed students to strategically narrow down their curriculum preparation. Hokayem, Jin, and Zhu (2014) examined biology examinations pre and post-curriculum reform and showed that despite significant reform efforts to improve students' ability to reason scientifically based on authentic scientific evidence, reform efforts failed to provide desired results and assessments remained focused on reproducing content rather than promoting authentic scientific reasoning. More recently, Shehayeb (2023) conducted an analysis of mathematics examinations and demonstrated a "prescriptive pedagogy" emphasizing rote learning and memorization due to the format of assessment design. The primary researcher's dissertation (Achkar, 2024) analyzed science examinations in biology, chemistry, and physics administered to grade 12 students in Lebanon from 2010-2022. He demonstrated a consistent emphasis on lower-order thinking skills, a high degree of question repetition across years and inherent contradictions between reform goals of assessment designers and practical implementation challenges such as limitations in examiner training, reliability of scoring, and potential legal challenges.

2.5. Gaps in Current Research

While a great deal of research has been done examining the content of the examinations given to Lebanese students, little to no research has investigated if student examination performance accurately reflects their true comprehension of course material or simply their ability to complete the type of question(s) that will be on the examination. This gap in research is important since the theoretical literature demonstrates that the format of assessment can greatly affect cognitive processes; therefore, simply analyzing the questions on the examination will not allow us to determine the learning outcomes of students. Although recent international research has demonstrated the existence of washback effects in various countries, such as Malaysia (Khan et al., 2025), Malta (Azzopardi & Camilleri, 2025), Pakistan (Syed & Butt, 2025), Bangladesh (Rahman et al., 2021), and Ethiopia (Abate & Mishore, 2024); to date, no Lebanese research has used empirical studies to compare student performance on traditional versus reformed assessments that tested equivalent content. International research has demonstrated that both the format of the assessment and the level of cognitive demand significantly affect instructional practice and student learning approaches; however, no Lebanese research has utilized experimental design to assess whether the

success of students on the Baccalaureate Examination reflects transferable understanding versus format-specific skills.

The present study addresses this methodology gap by conducting the first direct empirical comparison of Lebanese students' performance across assessment formats that test the same biological content and thus examines actual learning outcomes attributable to students' familiarity with the format of the examination versus their conceptual understanding of the biological content.

3. METHODOLOGY

The present study utilizes a quasi-experimental within-subject design to investigate whether the success of Lebanese grade 12 students on traditional Baccalaureate biology examinations reflects their true understanding of biology or is due in part to their familiarity with the format of the examinations. Prior research in Lebanon has primarily focused on either the content of the examinations or educators' perceptions of the examinations; therefore, the present design will directly measure how well students do when the same biological content is assessed in different formats and with varying levels of cognitive demands.

Each participant will take two assessments: Test A, which represents the familiar traditional Baccalaureate format; and Test B, which represents a reformulated version of the assessment that was developed using the same biological content and images but with different item types (e.g., multiple choice, data-based analysis, and extended response items). Since the design is a within-subjects design, each student will serve as their own control, and this will help minimize the impact of confounding variables related to students' prior knowledge, ability, and instructional exposure. Therefore, if the results show that students performed at a high level on the traditional format but significantly declined on the reformatted format despite both versions containing identical content, these results will provide direct evidence that the outcomes of students on the examinations are influenced by students' familiarity with the format of the examinations and the cognitive demands associated with the structural elements of the item. The magnitude of the effect of the format is measured using effect sizes and school-level analyses examine the extent to which the context of the institution affects the transferability of knowledge across formats.

Therefore, the current design aligns with the consequential validity perspective of Messick (1996) which emphasizes the need to evaluate the consequences of test development, including how test design affects learning outcomes and instructional priorities.

3.1. Research Design

The quasi-experimental research design is based on the within-subjects comparative design to compare the results of Lebanese grade twelve students who were assessed by traditional assessment formats and reformulated assessment formats of biology. In this case, the use of a within-subjects design allows researchers to utilize the following methodological advantages: increased statistical power and the elimination of the need to assume between-group equivalency.

Independent Variable: Assessment Format (Traditional Baccalaureate format vs. Reformatted format).

Dependent Variable: Student Performance - Percentage Score.

Control Measures: Identical Biological Content, Equivalent Time Allocation and Standardized Administration Conditions.

Participants:

N = 40 Grade Twelve Life Science Students were recruited from three different Lebanese Schools using Purposive Sampling. The number of participants will provide enough statistical power to detect large effect sizes ($d \geq 0.8$) at an alpha level of .05, when utilizing a paired t-test.

School Characteristics:

School A (n=13, 32.5%): Denominational school within the private school system. Serves a middle-class population. Biology is taught in French. Baccalaureate performance is very high, with a pass rate typically exceeding 90%.

School B (n=14, 35.0%): Secular school within the private school system. Diverse socio-economic population. Biology is taught in English. Baccalaureate performance is moderate, with a pass rate typically between 75% and 85%.

School C (n=13, 32.5%): Private school serving middle to upper-middle-class students. Biology is taught in French. Baccalaureate performance is strong (typically pass rate >85%).

Inclusion Criteria: Enrolled in Grade Twelve LS Track; Studying according to the Lebanese Biology Curriculum; Planning to take the Official Baccalaureate Examination; Parental Consent and Student Assent Provided.

3.2. Assessment Instruments

Both assessments tested the same biological material (Genetics, Immunity, Hormonal Regulation, and Molecular Biology), utilizing the same experimental documentation and illustrations from the 2022 Lebanese Baccalaureate Session. Test B was developed in early 2025 using the publicly available 2022 examination as its template. When the official June 2025 examination (Test A) was administered, it was found to be structured and thematic in relation to the 2022 session, thus confirming content equivalence among the two tests. Although this convergence was fortunate, it was also not surprising considering the Ministry's history of standardizing examinations across sessions (Achkar, 2024). Thus, the critical experimental manipulation existed solely in the form of question type and cognitive demands and not in the content itself. Test A was the official June 2025 Lebanese Baccalaureate biology examination, which followed the same format used in all national assessments for over two decades: four document-based exercises with typical prompts such as "Select from the text", "Demonstrate that" and "Indicate", focusing on procedural reasoning and short answer responses.

Test B reconfigured the same biological material into multiple-choice questions, data analysis tasks, and extended-response items, designed to stimulate higher-order thinking skills requiring students to integrate, evaluate and explain information, instead of simply reproducing procedural responses.

Each content area was aligned with the Revised Bloom's Taxonomy Levels (Anderson & Krathwohl, 2001; Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Therefore, detailed test blueprints ensured the equivalence of content coverage while providing a basis for systematic comparisons of cognitive demands and response formats.

Table 1 summarizes the cognitive profiles of both instruments.

Table 1. Content and cognitive distribution across both assessments.

Domain / Content Area	Representative Subtopics	Weight (%)	Test A Format & Cognitive Level	Test B Format & Cognitive Level
Genetics	Pedigrees (Hemophilia B, Albinism), Gene mutation, Chromosomal locus analysis	25	Structured document-based questions; Understand → Apply	MCQ + data analysis + short constructed response; Apply → Analyze
Immunity	COVID-19 vaccine mechanisms, Immune complex formation	25	Short text-based prompts; Remember → Understand	Graph interpretation and mechanism schematization; Analyze → Evaluate
Endocrine & Reproductive Physiology	Hypothalamus–Pituitary–Ovary axis, feedback mechanisms	25	Experimental interpretation from provided docs; Apply	Scenario-based MCQ + explanatory item; Apply → Evaluate
Molecular Biology / Cell Processes	DNA replication, Protein synthesis, Mutation impact	15	Definition + diagram labeling; Remember → Understand	MCQ + graphical data interpretation; Analyze → Create
Integration / Extended Response	Immune response comparison & Menstrual cycle essay	10	Two 10-mark essay-style questions; Apply → Analyze	Two extended responses requiring synthesis across domains; Analyze → Evaluate → Create

Table 2 provides a comprehensive comparison of both instruments' developmental, structural, and psychometric characteristics.

Table 2. Comparative features of test A and Test B.

Feature	Test A (Traditional Baccalaureate – 2025 Official)	Test B (Reformulated Assessment – Developed from 2022 Session)
Development Timeline	Administered June 2025 as official MEHE examination; format unknown during Test B design.	Developed early 2025 using 2022 session as a template before 2025 exam release.
Purpose	Benchmark traditional assessment representing familiar, high-stakes context.	Experimental instrument assessing knowledge transfer across novel formats and cognitive levels.
Content Domains	Genetics, Immunity, Hormonal Regulation, Molecular Biology.	Identical domains using same 2022 documents and figures.
Source Materials	2025 official documents and graphs (structurally parallel to 2022 session).	Reformatted 2022 session materials into new item types.
Question Formats	Document-based short-answer (≈ 30 items, 4 exercises).	Mixed formats (≈ 35 items): 25 MCQs + 10 constructed responses, including 2 extended tasks.
Duration	150 minutes.	150 minutes.
Cognitive Distribution	Levels 1–3 (Remember→Apply) $\approx 85\%$.	Levels 2–6 (Understand→Create) $\approx 80\%$.
Scoring Rubric	Official MEHE marking scheme.	Analytic rubric + MCQ answer key.
Reliability	Cronbach's $\alpha=0.82$; Inter-rater $r=0.94$	KR-20=0.86; Inter-rater $r=0.91$
Cognitive Intent	Surface-learning orientation: procedural and format-specific knowledge.	Authentic-reasoning orientation: conceptual understanding and knowledge transfer.

3.3. Data Collection

Data collection took place in June 2025. Test A was taken at designated testing centers as per Ministry of Education protocols. Test B was taken by students during their regular school day at their own desks with the same protocol and timing as the 2022 exam given in Lebanon (the original 2022 exam was used as a model for the development of Test B), thus providing a control for the instructional delivery, timing, and conditions for each test administration.

Three experienced biology teachers who had no affiliation with the participating schools acted as raters after attending training sessions. Both tests were scored blindly with any identifying information eliminated. Two raters independently scored the exams. Inter-rater reliabilities were $r = .94$ for Test A and $r = .91$ for Test B. Any discrepancies greater than 5 points were resolved by a third rater who consulted with the other raters until a consensus could be reached.

Descriptive statistics provided an overview of the performance distribution of the students on each assessment. A paired samples t-test was utilized to compare the mean performance of students on the two assessments, along with the effect size calculated by Cohen's d . School differences were investigated using one-way ANOVA. The relationship between students' performance on Test A and Test B was analyzed via Pearson's product-moment correlation coefficient. To confirm the results using a non-parametric approach, the Wilcoxon signed-rank test was also used. All statistical procedures were conducted using SPSS version 28, and a significance level of $\alpha = .05$ was employed.

Prior to conducting the research, the researcher obtained Institutional Review Board (IRB) approval from the Ethical Committee of Saint Joseph University (fsedu) under protocol number CER-2025-275. Additionally, the researchers obtained written informed parental consent and student assent prior to collecting data from the students. Student names were anonymous throughout the entire process and each student received a code identifier that was accessible only to members of the research team responsible for data collection and analysis. Officially, Test A results were recorded as part of the students' official Lebanese Baccalaureate grades. In addition, the results of Test B contributed to the students' final grades at their respective schools.

There are several limitations associated with this study that should be acknowledged. Most notably, the sample size ($N = 40$) does not provide sufficient statistical power to detect potential small effects or to conduct more complex statistical analyses, although it is likely to be sufficient to detect the large effects hypothesized to occur based upon the previous research. Furthermore, since the sample consisted of convenience sampling of schools located in Lebanon, there is a concern regarding whether the results will generalize to all Lebanese schools.

It is possible that students' lower performance levels on Test B resulted in part from their lack of familiarity with the MCQ format. Therefore, some degree of the decrease in performance can be attributed to the fact that students were less familiar with the format of Test B, but the lack of familiarity with the MCQ format is exactly what the study is investigating the "backwash" effect where students become proficient in specific formats, but do not necessarily have the ability to apply these proficiencies to different contexts.

Although the content of the 2022 and 2025 official examinations was similar, students performed significantly worse on Test B than on Test A. This suggests that format-specific proficiency, rather than unfamiliar content, was the main factor influencing the performance difference between the two tests.

4. RESULTS

4.1. Descriptive Statistics

The statistics in Table 3 provide evidence of significant differences in student test performance for each test type. Student test performance was significantly greater on the traditional baccalaureate test (Test A: $M = 85.12$; $SD = 8.46$) than it was for the atypical test (Test B: $M = 64.75$; $SD = 12.12$). There is a 20.38-point difference between Test A and Test B scores, which represents a decrease of 23.9 % in performance levels as measured by Test A.

Table 3. Overall performance statistics.

Assessment	N	Mean (%)	SD	Median	Range	IQR
Test A (Traditional)	40	85.12	8.46	88.00	60.00-95.00	10.50
Test B (Reformulated)	40	64.75	12.12	65.50	36.00-87.20	12.22
Difference (A-B)	40	20.38	13.30	20.50	-	-

Note: $p < 0.01$.

Distribution analysis revealed Test A was negatively skewed (Shapiro-Wilk $W = 0.8927$, $p = 0.0012$) with scores concentrated in the 75-95% range. Test B showed approximately normal distribution ($W = 0.9606$, $p = 0.1751$) with broader dispersion. Figure 1 presents the Histogram of student score distributions for Test A and Test B.

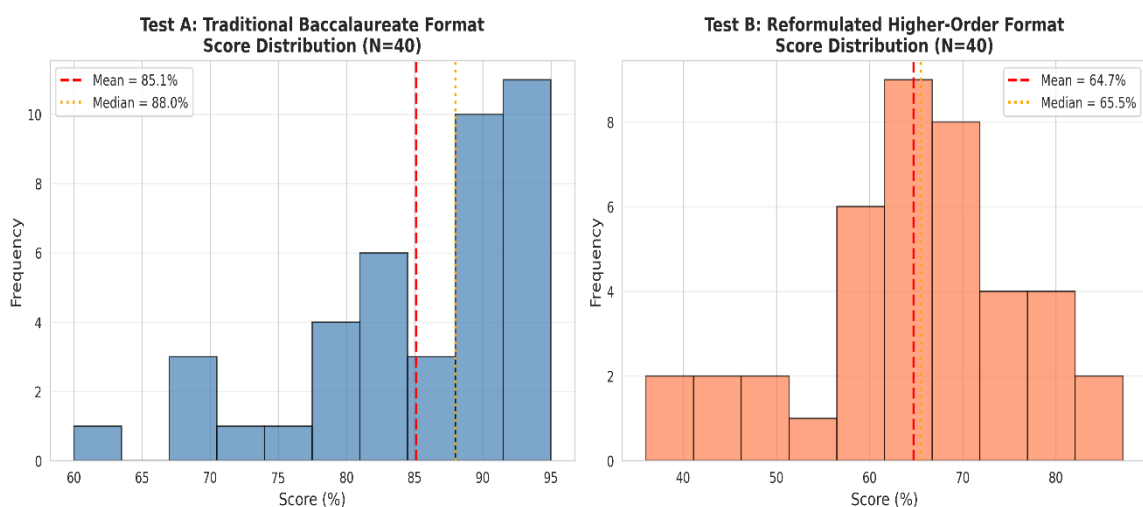


Figure 1. Score distribution.

Table 4 reveals significant institutional differences in performance patterns with schools showing significantly different magnitudes of decline. These differences provide crucial insight into varying instructional approaches and their relationship to format-dependent versus transferable learning.

Table 4. Performance by school.

School	n	Test A Mean (SD)	Test B Mean (SD)	Difference	% Decline
School A	13	86.15 (9.60)	54.62 (12.69)	31.54	36.6%
School B	14	81.79 (9.40)	69.63 (8.36)	12.16	14.9%
School C	13	87.69 (4.84)	69.62 (8.71)	18.08	20.6%

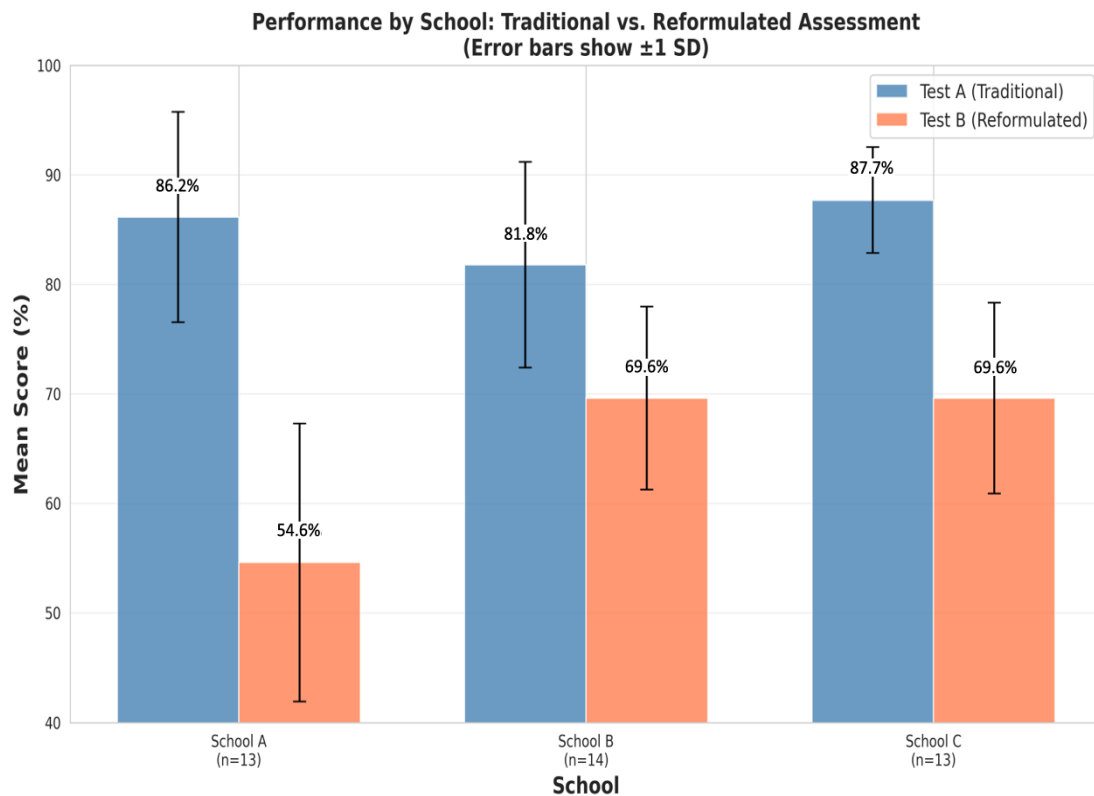


Figure 2. Mean Test A and Test B scores by school.

Figure 2 illustrates the mean performance of students across schools.

4.2. Inferential Statistics

The paired t-test revealed a highly statistically significant difference, $t(39) = 9.692$, $p < 0.000001$, strongly supporting Hypothesis 1. Table 5 presents the paired-samples t-test results for the comparison between Test A and Test B.

Table 5. Paired samples t-test.

Comparison	Mean Diff	SD	t	df	p	95% CI
Test A vs B	20.38	13.30	9.692	39	<0.000001	[16.13, 24.63]

Cohen's d effect size analysis quantified the practical magnitude of performance differences, addressing whether gaps are educationally meaningful beyond statistical significance. Table 6 presents the effect size of the difference between Test A and Test B.

Table 6. Effect size.

Measure	Value	Interpretation
Cohen's d	1.532	Very Large
η^2	0.707	70.7% variance explained

Note: Cohen's d substantially exceeds the conventional threshold for a large effect (≥ 0.8), based on benchmarks proposed by Cohen (1988).

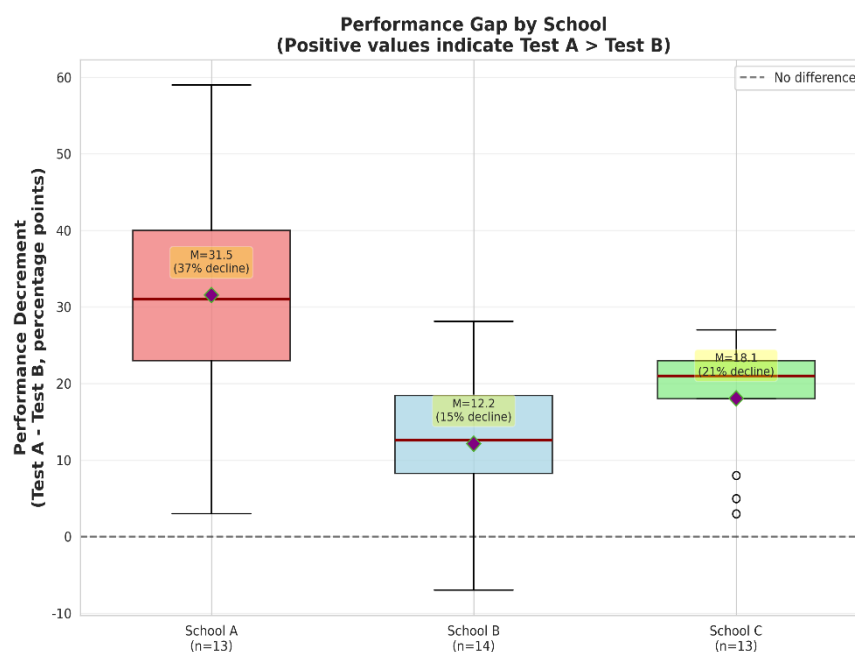
Cohen's d of 1.532 indicates a very large effect according to conventional guidelines (Small=0.2, medium=0.5, large=0.8), substantially exceeding the threshold for large effects. Assessment format explained 70.7% of the variance in student scores ($\eta^2=0.707$), indicating that format familiarity was the dominant factor determining performance differences.

One-way ANOVA examined whether schools differed in performance patterns, addressing Research Question 3 about institutional variations. Table 7 presents the ANOVA results assessing the school-level differences between Test A and Test B

Table 7. ANOVA results.

Variable	F	df	p	η^2	Interpretation
Test A by School	1.865	2,37	0.169	0.092	Not significant
Test B by School	9.748	2,37	<0.001	0.345	Significant
Difference by School	11.430	2,37	<0.001	0.382	Significant

ANOVA results revealed that schools performed similarly on the traditional format ($F(2,37) = 1.865$, $p = 0.169$) but differed significantly on the atypical format ($F(2,37) = 9.748$, $p = 0.0004$). Most importantly, the magnitude of performance gaps varied significantly by school ($F(2,37) = 11.430$, $p = 0.0001$), indicating that institutional approaches substantially influenced students' ability to transfer knowledge across assessment formats. Figure 3 illustrates the performance gap between Test A and Test B scores across schools.

**Figure 3.** Performance gap by school.

Pearson correlation between Test A and Test B scores examined whether students' relative performance remained consistent across formats or whether different competencies were being assessed. Table 8 presents the correlation between Test A and Test B scores

Table 8. Correlation.

Measure	r	p	r ²	Interpretation
Test A & B	0.203	0.210	0.041	Weak, not significant

The weak, non-significant correlation ($r=.203$, $p=.210$) indicates that only 4.1% of variance was shared between assessment formats. If both tests measured identical biological understanding, correlation should exceed $r=.70$. The observed weak correlation demonstrates that Test A and Test B assessed largely different competencies despite identical biological content.

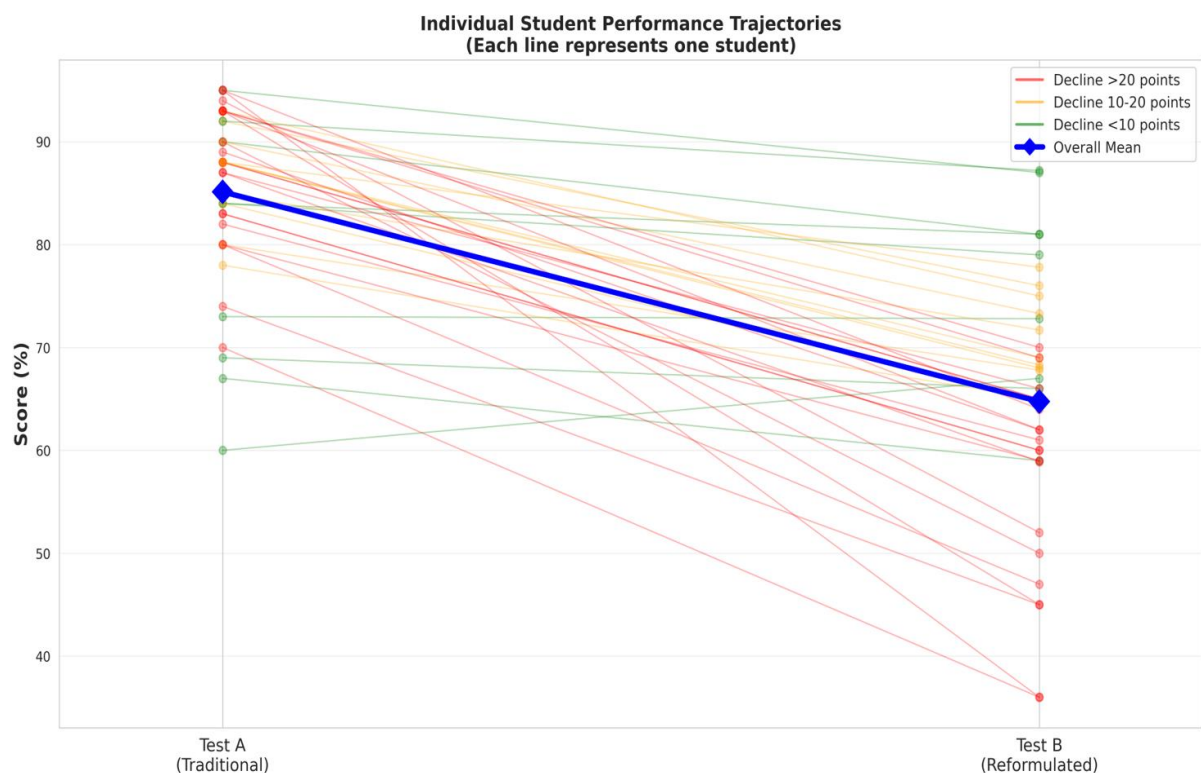
4.3. Additional Performance Metrics

Supplementary analyses revealed the practical significance of format effects for student academic outcomes. Table 9 presents the categorical distribution of student performance in Test A and Test B.

Table 9. Categorical performance.

Metric	Test A	Test B	Difference
Students $\geq 70\%$	37 (92.5%)	12 (30.0%)	-62.5 points
Students $\geq 80\%$	30 (75.0%)	5 (12.5%)	-62.5 points
Students >20-point decline	-	-	20 (50.0%)
Students improving on Test B	-	-	1 (2.5%)

While 92.5% of students achieved passing scores on the traditional format, only 30.0% passed the atypical format, a 62.5% reduction in pass rate despite identical biological content. Half of all students experienced performance decrements exceeding 20 points, and only one student (2.5%) scored higher on the atypical format. These findings demonstrate that format effects have profound practical consequences for academic achievement measures. Figure 4 illustrates the individual student performance trajectories between Test A and Test B.

**Figure 4.** Individual student performance.

5. DISCUSSION

5.1. Evidence of Format-Dependent Achievement

The primary conclusion of this study is the significant drop in student performance when the exact same biological content was examined using unfamiliar formats that required greater cognitive processing (i.e., higher order). The 20.38-point decrease in student performance when examining identical content in traditional versus reformulated formats, along with the extremely large effect size ($d = 1.532$ Cohen's), clearly demonstrates that how an assessment is structured has a large impact on what students will achieve.

This pattern aligns with washback theory, which suggests that prolonged exposure to the same assessment format encourages students to optimize their learning strategies for that specific format rather than develop broader, flexible thinking skills. In Lebanon, where biology instruction has historically mirrored the repetitive use of national examination formats, students' strong performance on the Baccalaureate exam likely reflects both their biological knowledge and familiarity with common task types. This consistency in assessment design reinforces procedural skills and knowledge, but may limit the development of adaptable thinking. Such an approach emphasizes rote learning and procedural familiarity, potentially impacting students' ability to apply knowledge flexibly in varied contexts. Overall, the exam format influences learning behaviors and skill development, aligning with washback effects.

5.2. Format Familiarity and Limited Knowledge Transfer

Students were consistently able to navigate the typical document-based exercises on Test A, i.e., extracting relevant information, identifying expected procedural steps, and utilizing previously practiced analytical routines. However, when the same concepts were presented in multiple choice, scenario-based, and integrative analytical formats in Test B, students experienced a significant decline in performance across all of the content areas.

This pattern of performance decline across all content areas when the format changed is indicative of students' knowledge being stored in a manner that is strongly linked to the specific response procedures in which they had been trained. These students were not supported in developing flexible ways of reasoning about the content across various formats of representation; instead, the training they received focused on developing retrieval strategies that were specific to particular formats. This type of retrieval strategy is effective in an environment with high stakes, but it limits students' ability to restructure knowledge, justify mechanisms, and synthesize evidence when there are no familiar cues available.

5.3. Implications of the Weak Cross-Format Correlation

The weak and statistically insignificant ($p < 0.01$) cross-format correlation ($r = 0.203$) between student Test A and Test B scores provides additional insight into the nature of the competences being assessed. If both Tests were measuring general biological knowledge, then a much higher degree of consistency in the ordering of individuals would be expected. However, since the amount of shared variance between Test A and Test B was minimal (4.1%), it appears that the knowledge being utilized by students to succeed on the traditional Baccalaureate format and the knowledge being utilized to succeed on the reformulated format are comprised of largely different skill sets. Student success on the traditional format assessment appeared to depend almost entirely on recognition of familiar patterns, recalling procedural information, and developing strategies that were format dependent. On the other hand, student success on the reformulated format assessment required the integration of concepts, interpretation of unfamiliar representations, and the utilization of explanations to justify reasoning. From a validity standpoint, this divergence raises significant concerns as to the extent to which current Baccalaureate biology scores accurately measure the broad biological competence of students versus the competency to perform well on an assessment that utilizes a highly predictable format.

5.4. Institutional Differences and Instructional Mediation

All of the schools studied here showed large drops in performance on the same content area as assessed in Test B, which used unfamiliar formats. The school showing the least loss (12.16 points) was School B, whereas School A showed a decline of over 30 points. Thus, although the washback effects of the structure of an assessment can depend upon the instruction employed by teachers, the instructional practices that include students experiencing a variety of formats of assessments, encouraging students to explain their thought processes, and providing routine instruction in analytical reasoning allow students to demonstrate the ability to transfer knowledge between formats. The data collected from School B indicates that it is possible to support high performance on traditional assessments while simultaneously supporting students' adaptation to a variety of different formats. Consistent with constructive alignment theory, the findings that School B's students performed at high levels on traditional assessments and that they were able to adapt to reason about different formats are consistent with the principles of constructive alignment theory. Constructive alignment theory (Biggs, 1996; Biggs, 1999) supports the idea that there should be coherence between the intended learning objectives, the instructional activities provided to students, and the assessment procedures provided to measure student performance.

5.5. Positioning within International Evidence

This study expands the global literature on the washback effects of performance-based assessment items and prior research by establishing a connection between instructional practices or curriculum content changes and specific cognitive performance measures. Prior studies in Morocco, Jordan, Pakistan, Bangladesh, Malaysia, and Ethiopia indicate a narrowing of the curriculum and a shift in pedagogy due to high-stakes testing. This research builds on this body of literature by demonstrating how the cognitive implications of curriculum and pedagogy shifts influence student performance. Previous studies may have implied the potential learning consequences of high-stakes testing, whereas this study indicates actual learning consequences: the format of an assessment is altered, yet the content remains constant, leading to significant differences in student performance.

5.6. Limitations

The present study has two major limitations that need to be considered. The first limitation is that the sample size is too small, and only three private schools were included in the study, therefore limiting the generality of the conclusions made from this study. The second limitation is that the students had little to no experience with the multiple-choice format of Test B, and therefore may have been less successful on Test B; however, the limited experience with the multiple-choice format is indicative of the students having been exposed to a very narrow assessment environment for an extended period of time. Although alternative explanations for the results of the present study, such as test anxiety, time management problems, or slight differences in task difficulty, are plausible, it is unlikely that such alternative explanations can account for the magnitude and consistency of the format-dependent performance across content areas and schools. Therefore, additional longitudinal research would be required to determine whether the format-dependent success of students in the 12th grade is predictive of later difficulties in university science courses or professional settings that require analytical reasoning.

6. CONCLUSION

Lebanon's current approach to teaching and assessing biology results in high levels of achievement for students on the traditional Baccalaureate examination, but at a cost. Students are able to successfully complete the Baccalaureate examination, but the degree to which students are able to apply their knowledge in a variety of formats is limited. It is neither necessary nor uniform for students to be forced to choose between achieving high levels of success on a test and the ability to apply that knowledge in various ways. By altering the instructional practices of schools, students can be allowed to demonstrate their knowledge in different formats. Reforming both the structure

of assessments and the instructional delivery practices of schools will be required if Lebanon is to find a balance between holding students accountable for meeting certain standards and allowing students the opportunity to learn meaningfully. If no reform occurs, high levels of success on the Baccalaureate examination may continue to mask fragile understanding adequate to pass exams, yet insufficient to enable students to engage in scientific reasoning beyond them.

6.1. Policy Recommendations

In order to assure the feasibility of a more diversified assessment format (Wiggins, 1998), MEHE should begin the implementation of a diversified assessment format via a series of pilot programs, cognitive demand audits using Bloom's revised taxonomy, and selective benchmarking against international assessments (Such as TIMSS and PISA).

At the Ministry of Education and Higher Education (MEHE), the MEHE should begin to gradually incorporate a variety of item formats into the assessment process of students, while keeping the same content as the traditional Baccalaureate. Some examples of formats which could be incorporated are: multiple choice, data interpretation, and extended analytical response formats. By incorporating these formats, MEHE will be able to create an assessment environment that is less predictable for students and encourage instructional practices that result in the development of transferable competencies.

At school level: Schools should change the focus of their instructional time to have a minimum of 60–70% devoted to helping students develop their conceptual understanding of biology and analytical reasoning skills. Schools should allocate a maximum of 30–40% of their time to helping students prepare for the Baccalaureate examination format. As indicated in school B, this ratio can be achieved even in an environment with a high level of competition for student enrollment.

At Teacher Level: Teachers should plan their instruction so that, prior to developing students' procedural fluency in performing specific tasks, students develop their explanatory understanding of biological mechanisms. Teachers should view past Baccalaureate exams as a means to assist students in reinforcing their procedural fluency and should not view the past Baccalaureate exams as a means to guide their instruction. Professional development programs should be designed to help teachers develop the ability to create tasks that require students to explain and justify their reasoning related to biological mechanisms, rather than just reproducing fixed response patterns.

6.2. Directions for Future Research

Future studies should attempt to replicate the design of this study using larger and more diverse samples of students and extend the comparison to include chemistry and physics. Additionally, future research could involve qualitative case studies of schools demonstrating strong knowledge transfer abilities among students. It could also examine the predictive validity of Baccalaureate scores on students' performance in university science courses and professional settings requiring analytical reasoning.

Funding: This study received no specific financial support.

Institutional Review Board Statement: This study was approved by the Institutional Review Board of the Saint Joseph University, Lebanon, under protocol number (CER-2025-275), dated [25/04/2025]. Informed verbal consent was obtained from all participants, and all data were anonymized to protect participant confidentiality.

Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.

Disclosure of AI Use: During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to aid in editing the clarity of writing, grammar and overall format. The scientific aspects of the research, as well as the study's design, data analysis and academic judgments were conceived and verified solely by the authors; therefore, the authors assume sole responsibility for the quality and accuracy of the information presented in this document.

REFERENCES

- Abate, T., & Mishore, E. (2024). Alignment analysis between teacher-made tests with the learning objectives in a selected school of central regional state of Ethiopia. *Heliyon*, 10(11), e31869. <https://doi.org/10.1016/j.heliyon.2024.e31869>
- Achkar, I. S. (2024). A comprehensive study on the complexity and patterns of questions in the Lebanese Baccalaureate examinations for LS biology, chemistry, and physics (2010–2022). Unpublished Master's Thesis. Saint Joseph University, Beirut, Lebanon.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-129. <https://doi.org/10.1093/applin/14.2.115>
- Allouh, Z. M. (2016). Evaluation of chemistry questions in the general secondary examination in Jordan for years 2010–2015. *International Journal of Education*, 8(4), 10-20. <https://doi.org/10.5296/ije.v8i4.9927>
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. United States: Longman.
- Azzopardi, M., & Camilleri, L. (2025). Examining the cognitive demands in biology and physics practical advanced level examinations in Malta. *Symposia Melitensia*, 20, 15-32.
- Belkbir, R. (2023). The backwash effect of the national exam on upper secondary Moroccan teaching. *Jurnal Arbitrer*, 10(2), 185-195. <https://doi.org/10.25077/ar.10.2.185-195.2023>
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347-364. <https://doi.org/10.1007/BF00138871>
- Biggs, J. (1999). *Teaching for quality learning at university: What the student does*. Buckingham, England: Society for Research into Higher Education & Open University Press.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. United States: David McKay.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hokayem, H., Jin, H., & Zhu, M. (2014). Scientific literacy in Lebanese biology national exams: A case study comparing pre/post exams. *International Education*, 44(1), 73–87.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge, England: Cambridge University Press.
- Khan, A., Hassan, N., & Cheng, L. (2025). Investigating the contextual factors mediating washback effects of a learning-oriented English language assessment in Malaysia. *Language Testing in Asia*, 15(1), 20. <https://doi.org/10.1186/s40468-025-00359-8>
- Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology*, 46(1), 4–11. <https://doi.org/10.1111/j.2044-8279.1976.tb02980.x>
- Mehmood, T., Iqbal, M., Abdullah, & Farooq, M. (2016). Analysis of question papers of physics at secondary level in Pakistan in the light of revised Bloom's taxonomy. *Journal of Applied Environmental and Biological Sciences*, 6(4S), 124–131.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. In (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256. <https://doi.org/10.1177/026553229601300302>
- Ministry of Education and Higher Education. (1997). *National curriculum for the Lebanese educational system (Decree No. 10227/97)*. Beirut: Ministry of Education and Higher Education.
- Rahman, K. A., Seraj, P. M. I., Hasan, M. K., Namaziandost, E., & Tilwani, S. A. (2021). Washback of assessment on English teaching-learning practice at secondary schools. *Language Testing in Asia*, 11(1), 12. <https://doi.org/10.1186/s40468-021-00129-2>
- Shehayeb, S. N. (2023). Exploring learning loss in mathematics for Lebanese secondary school students during economic and health crisis. *International Journal on Studies in Education*, 5(3), 269-287. <https://doi.org/10.46328/ijonse.132>
- Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Abingdon, England: Routledge.

- Syed, K. A. S., & Butt, M. (2025). Reimagining classroom assessment in Pakistan: A multi-study synthesis toward future-oriented assessment-informed pedagogy. *Futurity Education*, 5(4), 238-251. <https://doi.org/10.57125/FED.2025.12.14>
- Vlaardingerbroek, B., Jaber, L. Z., Rizk, N. G., & Bayoud, J. M. (2009). The Lebanese Brevet chemistry and physics examinations: An exploratory study. *International Journal of Science and Mathematics Education*, 7(6), 1227-1247. <https://doi.org/10.1007/s10763-009-9159-x>
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Views and opinions expressed in this article are the views and opinions of the author(s), International Journal of Education and Practice shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.