

RESEARCH ARTICLE

The Practice of Designing Summative Assessment Tasks and Assuring Their Quality in General Secondary Schools

Erkabayev Khusniddin Sultanovich

Independent Researcher, Tashkent State University of Economics, Uzbekistan

VOLUME: Vol.06 Issue06 2026

PAGE: 92-97

Copyright © 2026 Journal of Social Sciences and Humanities Research Fundamentals, this is an open-access article distributed under the terms of the Creative Commons Attribution-Noncommercial-Share Alike 4.0 International License. Licensed under Creative Commons License a Creative Commons Attribution 4.0 International License.

Abstract

The article analyses practical mechanisms for designing unit and term summative assessment tasks in general secondary schools and for assuring their quality. Drawing on J. Biggs's idea of constructive alignment, G. Wiggins and J. McTighe's idea of backward design, the taxonomy of educational objectives revised by L. Anderson and D. Krathwohl, and the validity theory of S. Messick and M. Kane, the author substantiates a five-stage quality cycle of summative assessment: clarifying learning outcomes, drawing up the assessment specification, developing tasks and assessment criteria, administration and moderation, and analysis of results and feedback. The key actions, instruments and quality indicators of each stage are summarised in a table; the procedure for calculating item difficulty and discrimination indices is presented, and the factors threatening the validity of assessment are identified. The findings contribute to enhancing teachers' assessment literacy and ensuring the reliability of summative assessment results.

KEY WORDS

Summative assessment, unit summative assessment, term summative assessment, assessment specification, constructive alignment, Bloom's taxonomy, assessment criterion, descriptor, rubric, validity, reliability, moderation, item analysis, assessment literacy.

INTRODUCTION

The introduction of a criterion-referenced assessment system in general secondary education has placed new demands on teachers' professional activity. Whereas previously the teacher mainly used ready-made test papers and marked students' work on a five-point scale on the basis of a general impression, he or she is now required to design summative assessment tasks independently, to develop assessment criteria and descriptors for them and to record results on a 100-point scale. Under the regulatory document currently in force, unit summative assessment (BSB) accounts for 50% of a student's term score, term summative assessment (CHSB) for 40% and formative assessment for 10%. Consequently, half of a student's grade depends on the quality of BSB tasks designed

within the school by the teacher. Given that from the 2026/2027 academic year the new system covers more than 5,000 schools, the issue of task quality goes beyond the confines of an individual school and acquires nationwide significance.

The importance of this problem at the level of state policy is also reflected in the National Programme for the Development of School Education in 2022–2026, which sets the task of continuously improving curricula on the basis of the results of assessing students' and teachers' knowledge. However, assessment results can serve as a basis for management decisions only if they are reliable and valid. If a task measures

something other than the outcome specified in the curriculum, or if the same piece of work is marked differently by different teachers, both the grade awarded on that basis and the school ranking lead to erroneous conclusions.

In pedagogical research, issues of summative assessment are often addressed at the level of describing types of test items or forms of assessment, while the integrated quality cycle extending from task design to the analysis of results has not been sufficiently systematised. In particular, no practical algorithm has been developed for teachers on aligning BSB and CHSB tasks with learning outcomes, constructing them on the basis of a specification, clarifying assessment criteria and conducting psychometric analysis of the results. The aim of the article is to provide a theoretical substantiation and systematisation of practical mechanisms for designing summative assessment tasks and assuring their quality in general secondary schools.

LITERATURE REVIEW AND METHODOLOGY

The fundamental principle of designing summative assessment tasks is the idea of constructive alignment developed by J. Biggs. Arguing that learning outcomes, teaching activities and assessment tasks must be aligned with one another, the author emphasised that students learn precisely what assessment requires of them. The practical implication of this principle is that an assessment task must replicate exactly the verb and cognitive level specified in the learning outcome: if an outcome is defined as "is able to analyse", a task that requires only recalling a definition does not measure that outcome. G. Wiggins and J. McTighe developed this idea into the "backward design" model of the learning process: first the expected outcomes are planned, then the evidence of their attainment, that is, the assessment tasks, and only after that the teaching activities. In this approach, summative assessment becomes not the final part of the lesson plan but its starting point.

The taxonomy of educational objectives revised by L. Anderson and D. Krathwohl is widely used to clarify the cognitive level of learning outcomes. It divides cognitive processes into six levels – remembering, understanding, applying, analysing, evaluating and creating – and presents them in the form of a two-dimensional table in which they intersect with four types of knowledge: factual, conceptual, procedural and metacognitive. This two-dimensional table provides a natural basis for drawing up an assessment

specification: it makes it possible to ensure a balanced distribution of tasks both by content and by cognitive level. The concept of a specification has also been formally introduced into national practice: the model programme for teachers' professional development defines the content of assessment as the set of knowledge, skills and competences that a student must demonstrate in a particular test, and assessment criteria as the measures used to determine precisely the level of attainment of learning outcomes.

The theoretical basis for analysing assessment quality is the concept of validity. S. Messick defined validity as the degree to which the inferences drawn and the decisions made on the basis of test scores are supported by theoretical and empirical evidence. M. Kane operationalised this idea in the form of the "validity argument" model, showing that every claim concerning the interpretation and use of results – that the score has been assigned correctly, that it generalises to other tasks, that it can be extrapolated to the learning outcome, and that a decision can be made on its basis – must be supported by separate evidence. T. Crooks, M. Kane and A. Cohen, in turn, regarding the assessment process as a chain of interconnected links, argued that threats to validity may arise at each of the stages of administration, scoring, aggregation, generalisation, extrapolation, evaluation, decision-making and impact, and that the weakest link in the chain determines the validity of the assessment as a whole. The Standards for Educational and Psychological Testing likewise identify validity, reliability and fairness as the three fundamental requirements of any assessment.

The issue of assessment criteria and their development has been addressed in the studies of D. Sadler and S. Brookhart. Showing that a student cannot improve his or her work without understanding what high-quality work looks like, Sadler argued for the need to communicate criteria and standards to students openly and in advance. Brookhart, for her part, defined a rubric as a tool consisting of criteria and descriptions of quality levels for each criterion, and showed that in a high-quality rubric the criteria should express the features of the learning outcome rather than of the task, and the levels should be differentiated by descriptions of quality rather than by the number of points.

The fulfilment of all the above requirements ultimately depends on teachers' assessment literacy. R. Stiggins introduced this concept into scholarly discourse and defined it as the ability to distinguish between sound and unsound

assessment and to use assessment results appropriately. W.J. Popham, in turn, emphasised that assessment literacy is not a passing fad but a fundamental professional competence for teachers.

The study employed the methods of theoretical analysis, examination of regulatory documents, modelling and generalisation. The difficulty and discrimination indices of classical test theory were selected for the quantitative evaluation of task quality, since they can be calculated by any teacher using an ordinary spreadsheet without specialised software.

DISCUSSION AND RESULTS

On the basis of an analysis of theoretical sources and current regulatory requirements, a five-stage cycle for designing summative assessment tasks and assuring their quality was developed. The essence of the cycle is that each stage builds on the outcome of the preceding stage, while the results of the final stage are fed back into the improvement of the next assessment process.

The first stage is clarifying learning outcomes. The teacher reformulates the outcomes specified in the relevant unit of the curriculum as observable and measurable actions. Instead of vague formulations such as “knows the topic” or “has an understanding of the concept”, precise formulations are used, such as “compares two historical sources and explains the differences between them” or “solves a word problem by setting up an equation”. For each outcome, the cognitive level is determined according to the Anderson–Krathwohl taxonomy. The second stage is drawing up the assessment specification. The specification is a two-dimensional matrix whose rows show the main elements of the unit content, whose columns show the cognitive levels, and whose cells show the number of tasks and their points. The specification ensures full coverage of the unit content by the tasks and a balance across cognitive levels. As a practical recommendation, it is advisable in a BSB paper to allocate approximately 30% of the points to tasks at the levels of remembering and understanding, 50% to applying, and 20% to analysing, evaluating and creating; this ratio is adjusted according to the specific features of the subject and the grade level. The third stage is developing the task and the marking scheme. A task is constructed for each cell of the specification

and accompanied by a marking scheme – an answer key for closed-response tasks and criteria and descriptors for open-response tasks. A descriptor must specify precisely what each point is awarded for: not “the solution is correct”, but “the equation is set up correctly – 1 point; the equation is solved correctly – 1 point; the answer is interpreted in accordance with the conditions of the problem – 1 point”. The assessment criteria are communicated to students in advance.

The fourth stage is administration and moderation. BSB and CHSB are administered under uniform conditions, at the scheduled time and in compliance with the procedural rules set out in the regulatory document. The main means of ensuring the reliability of open-response tasks is moderation: within the school’s subject methodological association, at least a proportion of randomly selected scripts is independently marked by a second teacher, discrepancies between scores are discussed and the interpretation of the criteria is standardised. The fifth stage is analysis of results and feedback. At this stage, two tasks are addressed: providing feedback to students and checking the quality of the tasks. Within the first task, results are analysed by learning outcome, and additional work is planned on outcomes that have not been mastered. Within the second task, two indicators are calculated for each item. The difficulty index is determined by the formula $p = R / N$, where R is the number of students who answered the item correctly and N is the number of students who took the test; in practice, items with a p value in the range of 0.30–0.70 are considered acceptable. The discrimination index is determined by the formula $D = (U - L) / n$, where U and L are the numbers of students answering the item correctly in the highest- and lowest-scoring groups by total score (usually 27% of the students each), and n is the number of students in each group. According to the recommendation of R. Ebel and D. Frisbie, an item with $D \geq 0.40$ is considered very good, 0.30–0.39 – good, 0.20–0.29 – in need of improvement, while an item with a value of 0.19 or below should be revised or discarded. An item with a negative discrimination index, that is, an item on which stronger students performed worse than weaker ones, usually indicates an error in the key or an ambiguous item stem.

The key actions, instruments and quality indicators of the identified stages are summarised in the table below (see Table 1).

Table 1. The cycle of designing summative assessment tasks and assuring their quality

Stage	Key action	Instrument	Quality indicator
1. Clarifying learning outcomes	Expressing curriculum outcomes as observable actions; determining the cognitive level	Anderson–Krathwohl taxonomy, curriculum	Each outcome is expressed with a measurable verb
2. Drawing up the specification	Distributing tasks and points across content elements and cognitive levels	Two-dimensional specification matrix	Full coverage of unit content; balance of cognitive levels
3. Developing the task and marking scheme	Preparing the task text, answer key, criteria and descriptors	Item bank, rubric, descriptors	Alignment between task and outcome; unambiguity of criteria
4. Administration and moderation	Administration under uniform conditions; independent re-marking of a proportion of scripts	Procedural rules, moderation record	Inter-rater agreement
5. Analysis and feedback	Analysing results by learning outcome; calculating item statistics	Difficulty (p) and discrimination (D) indices, results map	$p = 0.30\text{--}0.70$; $D \geq 0.30$; written feedback provided to the student

The data in Table 1 show that the quality of summative assessment is determined not by the task itself but by the entire chain extending from its design to the use of the results. From the perspective of the Crooks, Kane and Cohen model, the links that pose the greatest threats to validity in school practice are as follows. The first is the scoring link: when descriptors are imprecise, the same answer is marked differently by different teachers, and the result reflects the marker's personal expectations rather than the student's knowledge. The second is the generalisation link: since a BSB paper consists of only a few tasks, if the tasks cover only part of the unit content, interpreting the result as the level of mastery of the unit as a whole is unjustified. The third is the extrapolation link: if the tasks are constructed mainly at the level of remembering, the scores become unsuitable for drawing conclusions about the competences envisaged in the curriculum. The fourth is the impact link: if results are merely recorded in the gradebook and no feedback is given to the student, summative assessment loses its potential to improve teaching. The cycle presented in the table is aimed precisely at strengthening these four links: the specification secures generalisation and extrapolation, descriptors and moderation secure scoring, and analysis and feedback secure impact.

The role of the school's subject methodological association is decisive in putting the cycle into practice. BSB tasks designed

by an individual teacher inevitably reflect his or her personal experience and expectations. It is therefore advisable to subject tasks and specifications to collective review within the methodological association, to discuss item statistics at the end of each term and to accumulate high-quality tasks in the school item bank. Such an approach, on the one hand, increases the comparability of BSB results across classes and teachers and, on the other, develops teachers' assessment literacy in the course of their practical work. As Stiggins and Popham emphasised, assessment literacy is formed not as theoretical knowledge but through the practice of constructing, analysing and refining assessment instruments.

In the course of the study, two serious methodological risks in the organisation of summative assessment were identified. The first is the disconnection of the assessment instrument from the learning outcome, that is, a situation in which a task measures not the outcome specified in the curriculum but whatever is easy for the teacher to construct. This risk most often manifests itself in papers consisting solely of closed-response tasks at the level of remembering. The second is the summative result remaining merely a final record, that is, BSB and CHSB results being used only for grading and not for adjusting teaching. The condition for preventing both risks is the consistent implementation of the stages of the assessment cycle and the transformation of the analysis of results into a

mandatory part of the teacher's work.

It should be specifically emphasised that the cycle proposed in the article is the result of theoretical analysis. To evaluate its effectiveness, it is necessary to organise experimental work, to measure the difficulty and discrimination indicators of BSB tasks in the experimental and control groups, the level of inter-rater agreement and the dynamics of teachers' assessment literacy, and to process the results using mathematical and statistical methods. This task has been identified as the next stage of the research.

CONCLUSION

The theoretical study conducted leads to the following conclusions. First, the fact that in the new assessment system half of a student's term score depends on BSB tasks designed within the school makes the quality of summative assessment tasks a decisive factor in ensuring the quality of education. Second, the principles of constructive alignment and backward design require summative assessment to be regarded not as the final part of the learning process but as the starting point of its planning; a task must replicate exactly the cognitive level specified in the learning outcome. Third, the design and quality assurance of summative assessment tasks should be organised as an integrated cycle comprising the stages of clarifying learning outcomes, drawing up the specification, developing the task and marking scheme, administration and moderation, and analysis and feedback. Fourth, indicators of classical test theory such as the difficulty and discrimination indices are simple tools that any teacher can calculate in a school setting and that enable an objective check of task quality. Fifth, the greatest threats to the validity of assessment arise in the scoring, generalisation, extrapolation and impact links; the condition for eliminating them is the systematic application of specifications, descriptors, moderation and feedback.

On the basis of the findings, the following recommendations are put forward:

- 1) to develop model specification matrices and sample descriptors for BSB by subject and grade and to make them available to teachers in electronic form;
- 2) to introduce into the work of school methodological associations a procedure for the collective review of BSB tasks and for the moderation of open-response papers;
- 3) to add to the electronic gradebook platform an analytical

module that automatically calculates item difficulty and discrimination indices and displays results by learning outcome;

- 4) to introduce into teacher professional development programmes a module based on practical sessions on designing and analysing summative assessment tasks.

The implementation of these recommendations will enhance the validity and reliability of summative assessment results and support their effective use for improving teaching.

REFERENCES

1. Decree of the President of the Republic of Uzbekistan No. PF-134 of 11 May 2022 "On Approval of the National Programme for the Development of School Education in 2022–2026" // National Database of Legislation. – URL: <https://lex.uz/docs/-6008663>
2. Order of the Minister of Preschool and School Education of the Republic of Uzbekistan No. 1-mh of 4 February 2025 (registration No. 2117-2, 05.02.2025) // National Database of Legislation. – URL: <https://lex.uz/uz/docs/-7362905>
3. Model Curriculum of the Differentiated Continuous Professional Development Course for General Secondary School Teachers. – Tashkent, 2023. – URL: <https://www.avloniy.uz/uploads/educational/7aaa6e967f2a04e39e154533141f5186.pdf>
4. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. – Washington, DC: AERA, 2014. – 230 p.
5. Anderson L.W., Krathwohl D.R. (Eds.) A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. – New York: Longman, 2001. – 352 p.
6. Biggs J. Enhancing Teaching through Constructive Alignment // Higher Education. – 1996. – Vol. 32, No. 3. – P. 347–364. – DOI: 10.1007/BF00138871.
7. Brookhart S.M. How to Create and Use Rubrics for Formative Assessment and Grading. – Alexandria, VA: ASCD, 2013.
8. Crooks T.J., Kane M.T., Cohen A.S. Threats to the Valid Use of Assessments // Assessment in Education:

Principles, Policy & Practice. – 1996. – Vol. 3, No. 3. – P. 265–286. – DOI: 10.1080/0969594960030302.

- 9.** Ebel R.L., Frisbie D.A. Essentials of Educational Measurement. 5th ed. – Englewood Cliffs, NJ: Prentice Hall, 1991.
- 10.** Kane M.T. Validating the Interpretations and Uses of Test Scores // Journal of Educational Measurement. – 2013. – Vol. 50, No. 1. – P. 1–73. – DOI: 10.1111/jedm.12000.
- 11.** Messick S. Validity // Linn R.L. (Ed.) Educational Measurement. 3rd ed. – New York: American Council on Education; Macmillan, 1989. – P. 13–103.
- 12.** Popham W.J. Assessment Literacy for Teachers: Faddish or Fundamental? // Theory Into Practice. – 2009. – Vol. 48, No. 1. – P. 4–11. – DOI: 10.1080/00405840802577536.
- 13.** Sadler D.R. Formative Assessment and the Design of Instructional Systems // Instructional Science. – 1989. – Vol. 18, No. 2. – P. 119–144. – DOI: 10.1007/BF00117714.
- 14.** Stiggins R.J. Assessment Literacy // Phi Delta Kappan. – 1991. – Vol. 72, No. 7. – P. 534–539.
- 15.** Wiggins G., McTighe J. Understanding by Design. Expanded 2nd ed. – Alexandria, VA: ASCD, 2005. – 370 p.