



Applications of Generative AI in Summative Assessment

4.1 INTRODUCTION

For a long time, summative assessments have been a cornerstone in the ever-evolving education landscape. By summative assessment, we refer to the evaluation of learners' academic performance at the conclusion of an instructional period, such as the end of a project, unit, course, semester, or academic year. It is often contrasted with formative assessment, which is ongoing and allows for continuous feedback and adjustment of teaching and learning strategies in line with learners' needs (Man Sze Lau, 2016). Summative assessments are typically structured and standardised, designed to measure each learner's achievement in relation to learning objectives and outcomes. Examples include final examinations, end-of-unit tests, dissertations, presentations, and portfolios.

While their exact forms have changed over time, summative assessments remain the standard way to evaluate learners' knowledge and determine whether learning outcomes have been achieved at the end of a unit or course. However, while summative assessments are an essential part of the educational process, they are most effective when used in conjunction with formative assessments, which provide ongoing feedback and support for learning (Man Sze Lau, 2016; Taras, 2009). In Kolade et al. (2024) we provide a discussion of how Generative AI (GenAI) tools influence assessments especially in higher education where

written essays are a widely used type of summative assessment. We highlight, for instance, the critique that summative assessments sometimes promote memorisation rather than understanding and application that are associated with deep learning (Rawlusyk, 2018; Dixon and Worrell, 2016).

However, the rapid diffusion of GenAI tools has significantly expanded the scope and potential of summative assessments to enhance learning. This is true despite the downsides associated with the use GenAI tools. Two downsides are worth noting here. Firstly, tools like ChatGPT and Google Gemini now make it easier for learners to cheat by passing off AI-generated work as their own. Secondly, dependence on GenAI tools could limit genuine learning, and this raises the question of where to draw the line. These concerns notwithstanding, the application of GenAI in summative assessment will enable innovation and effectiveness in education. By automating routine tasks such as question generation and grading, GenAI frees educators to focus on the specialised tasks that demand their unique skills and insights. This collaborative approach optimises educational outcomes and ensures that assessment remains both effective and deeply human-centred.

Contrary to the conventional critiques, recent research suggests that learners show positive attitudes and perceptions towards GenAI-mediated instruction and evaluation (Fathi et al, 2024). The capacity to create customised GPTs makes this even more interesting. As we show in the previous chapter (Chapter 3), instructors can create and train their own GPT based on specific content that is relevant to their learners. Such customised GPTs are a nimble way to teach and provide real-time feedback. And they are much faster at these than human instructors. Thus, the future of education appears not as a battleground between technology and tradition, but as a fertile ground for synergy. The goal of preparing learners for the challenges and opportunities of the future can be achieved more creatively and efficiently with, rather than despite, GenAI.

This chapter explores the application of GenAI in summative assessment, focusing on how to improve the quality and fairness of evaluations. We will consider some ways in which AI can enhance various aspects of assessment, from the creation of personalised tests to the provision of tailored feedback that help learners to improve.

4.2 THE CONTEXT AND RELEVANCE OF GENERATIVE AI TO SUMMATIVE ASSESSMENT

Before turning to a discussion of GenAI use cases in summative assessment, it is worthwhile to briefly discuss the importance of summative assessments and how AI could make them even better. Summative assessments are most widely applied to evaluate whether learners have achieved certain outcomes but there are several additional points of relevance. For instance, summative assessments offer way to keep learners accountable. Impending assessments at the end of a structured learning period is one of the reasons learners stay focused throughout the period. Moreover, summative assessments provide a formal certification of learners' achievements as a basis for progression to higher levels of learning or certification of final achievement.

In today's teaching and learning landscape, widely accessible GenAI tools could be applied in each area of importance as highlighted in Table 4.1. The content of the Box was generated with Microsoft Copilot. After creating the table, Copilot was quick to point out that "while AI has the potential to greatly enhance summative assessment, it's important to use it as a tool to support, rather than replace, human judgement and expertise". This is a stark reminder that, no matter how smart they become, AI tools are not likely to completely usurp the role of the human educator. This reminder is relevant today but even more so in the future for which today's education is preparing the learner. The World Economic Forum's Future of Jobs Report 2023 predicts that by 2027, 44% of workers' core skills will be disrupted due to the rapid pace of technological change, because technology is moving faster than companies can design and scale up their training programmes (Di Battista et al, 2023). The emphasis of education, and specifically of assessments, in this context should therefore transcend tests of what learners know—what we referred to as "know-what" in Kolade et al. (2024). Helping learners to acquire competence ("know-how") and demonstrate it ("show-how") should form the core of summative assessments. The rest of this chapter concentrates on illustrative cases of how educators can leverage GenAI tools to facilitate know-how and show-how.

Table 4.1 The importance of summative assessment

| <i>Importance of summative assessment</i> | <i>How AI can facilitate</i> |
|---|---|
| Evaluation of learning outcomes | AI can provide a more nuanced and comprehensive evaluation of learning outcomes by analysing a wide range of learner data and generating detailed performance reports |
| Summative assessments provide a clear and objective measure of what a learner has understood and can do at the end of a learning period. They help in determining whether the learning objectives of the course or unit have been met | Example tools: Google Forms (for creating assessments), Google Classroom (for managing and tracking student progress) |
| Accountability | AI-powered platforms can track learner progress in real time, providing immediate feedback and holding learners accountable for their learning |
| Summative assessments hold learners accountable for their learning by providing a formal structure for evaluating their knowledge and skills. This can motivate learners to review and consolidate their learning | Example tools: Kahoot! (for creating fun, engaging learning games), Quizlet (for creating study sets and quizzes) |
| Feedback for educators | AI can analyse assessment data to provide educators with insights into the effectiveness of their teaching strategies, helping them make data-driven decisions |
| Summative assessments can provide valuable feedback for educators, helping them understand the effectiveness of their teaching strategies and make necessary adjustments for future instruction | Example tools: Edmodo (for tracking student progress and providing feedback), Socrative (for instant feedback on quizzes and checks for understanding) |
| Standardisation | AI can ensure the standardisation of assessments by eliminating human bias and error, leading to fairer and more consistent evaluations |
| Summative assessments provide a standardised measure of comparison between learners, classes, or schools, which can be useful for benchmarking and policymaking | Example tools: Turnitin (for plagiarism detection and ensuring academic integrity), Grammarly (for writing assistance and checking quality of written work) |

(continued)

Table 4.1 (continued)

| <i>Importance of summative assessment</i> | <i>How AI can facilitate</i> |
|--|--|
| Certification and progression | AI can automate the grading process, making it quicker and more efficient, which is particularly beneficial for large-scale assessments often used for certification and progression |
| Summative assessments often play a crucial role in certification and progression, as they provide a formal record of learners' achievement | Example tools: Grade scope (for consistent and efficient grading), Moodle (for creating and managing online courses) |

Source: Extracted from conversation with Microsoft Copilot on 12 March 2024 (1500 GMT). Prompt: "I want to discuss assessment in higher education. Let's start with definition and importance of summative assessment"

4.3 CASE I: EXPLORING SUMMATIVE ASSESSMENT TYPES

Chapter 3 already introduces the idea of custom GPTs that users can create and train in ChatGPT. Custom GPTs are useful in the domain of summative assessments. For example, a custom GPT can be trained on the content of a specific course or module, including handouts, written notes, texts, and reference materials used in teaching the course or module. The GPT can then readily be prompted to create summative assessment tasks that the instructor can modify and use. For example, a new GPT named *Africa Survey Sage* was created and trained for the purpose of a seminar on field surveys in Africa. The GPT was trained with two publications: an article on implementing surveys in Africa (Egbetokun, 2022) and an authoritative open-access text published by the World Bank (Iarossi, 2006). Both publications were required reading for the seminar and formed the basis for most of the teaching materials. The instructor can create a summative exam with the help of *Africa Survey Sage*. Prompt 4.1 illustrates how this might be done. The output, truncated due to space restriction, is a comprehensive set of questions that test learners' understanding at different levels.

Prompt 4.1: Creating Ideas for Summative Assessment (Examination)

ChatGPT Prompt

I want to create a summative exam on how to design and implement field surveys in Africa. Propose a format that is appropriate for a cohort

of international students studying in business management MSc in a UK business school.

ChatGPT Response

Exam Format: Summative Assessment

Part 1: Multiple Choice Questions (MCQs)—30%

- Objective: Assess basic theoretical knowledge and understanding of key concepts related to survey design and implementation in Africa.
- Structure: 20 MCQs, each worth 1.5 points.
- Sample Topics: Sampling techniques, challenges unique to the African context, ethical considerations, data collection methods.

Part 2: Short Answer Questions—20%

- Objective: Test students' ability to articulate insights on specific aspects of survey design and implementation.
- Structure: 4 questions, each worth 5 points.
- Sample Topics: Comparing and contrasting survey methodologies, discussing the implications of cultural diversity in survey design, strategies to enhance response rates in diverse African communities.

Part 3: Case Study Analysis—30%

- Objective: Evaluate students' ability to apply theoretical knowledge to real-world scenarios, focusing on problem-solving and strategic thinking.
- Structure: 1 comprehensive case study involving a scenario where students must design a survey for an African country, considering logistical, cultural, and ethical challenges. Students will be asked to outline their survey design, sampling strategy, and data collection plan.
- Assessment Criteria: Clarity of the survey objective, appropriateness of the methodology, consideration of ethical issues, and feasibility of the implementation plan.

Part 4: Essay Question—20%

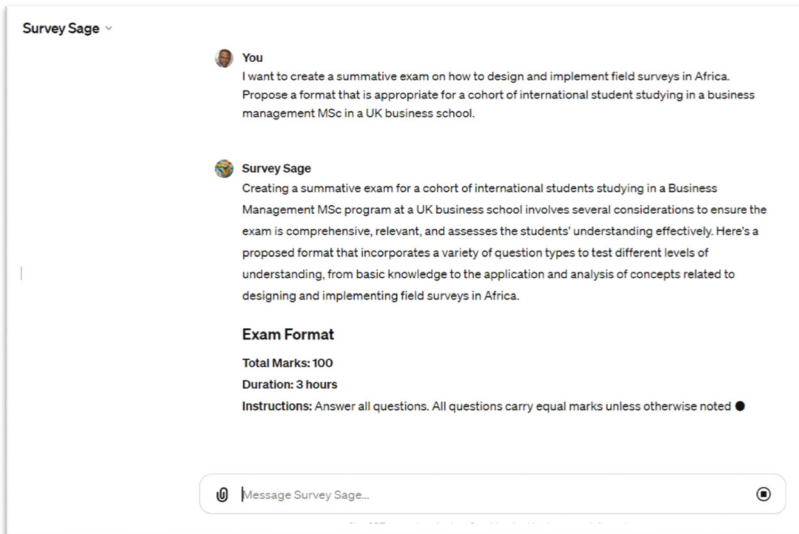
- Objective: Measure students' critical thinking and ability to engage with complex issues related to conducting surveys in Africa.

- **Structure:** 1 essay question requiring a well-structured argument, supported by examples from literature or case studies covered in the course.
- **Sample Topics:** Discuss the impact of digital technology on survey methodologies in Africa, or evaluate the role of international organizations in supporting survey research in African countries.

Part 5: Reflection on Learning—Optional Extra Credit (up to 5%)

- **Objective:** Encourage students to reflect on their learning experience, highlighting insights gained and how they can apply this knowledge in their future careers.
- **Structure:** A short reflective essay on key takeaways from the course and how the student plans to incorporate this knowledge into their professional life.

ChatGPT Screen.



4.4 CASE 2: CREATING MULTIPLE-CHOICE SUMMATIVE ASSESSMENTS

AI tools can support the instructor to speed up the rate at which tests, and assessment tasks are generated. For example, we could take our exercise from the preceding section forward by asking *Africa Survey Sage* to provide a variety of tasks within the summative assessment. The output generated from doing this is shown in Prompt 4.2. It includes a range of tasks in five parts that the instructor could modify to create a comprehensive assessment (Prompt 4.2). The collection of assessment tasks addresses theoretical knowledge, practical application, and critical thinking skills. The suggestions also include a clear objective and structure for each task as well as sample topics and assessment criteria for some of them. It is noteworthy that ChatGPT 4 provides a similar level of detail, but with less specific content compared to *Africa Survey Sage*.

Each of the assessment aspects in Prompt 4.2 can be expanded by following up with appropriate prompts. For example, we could ask it, or any other GenAI chatbot for that matter, to propose a set of multiple-choice questions (MCQs) for Part 1. For example, if we prompt Google Gemini with the specifications of Part 1 of our test, we get the output reproduced in Prompt 4.2.

Prompt 4.2: Ideas for a Multiple-Choice Test - Google Gemini

Google Gemini Prompt

I want to generate a summative test with the following specifications.

Type: Multiple Choice Questions (MCQs).

Objective: Assess basic theoretical knowledge and understanding of key concepts related to survey design and implementation in Africa.

Structure: 20 MCQs, each work 1.5points.

Sample Topics: Sampling techniques, challenges unique to the African context, ethical considerations, data collection methods.

Number of options per question: 4

Indicate the correct answer with an explanation.

Google Gemini Response

Sample Questions

1. Which sampling technique would be most appropriate for a researcher seeking to gather data from a widely dispersed rural population in an African country with limited infrastructure?

- (A) Simple random sampling
- (B) Cluster sampling
- (C) Systematic sampling
- (D) Quota sampling

Answer: B (Cluster sampling) Explanation: Cluster sampling allows for representative sampling when a complete list of the population is unavailable or logistically difficult to work from. It groups the population geographically for more efficient data collection.

2. A major challenge when conducting survey research in many African contexts is:
- (A) High levels of literacy
 - (B) Reliable internet access across the population
 - (C) Difficulty obtaining consent due to cultural norms
 - (D) Uniformity of languages spoken

Answer: B (Reliable internet access across the population) Explanation: Limited internet infrastructure can create significant hurdles for online surveys and can impact data collection methods in general.

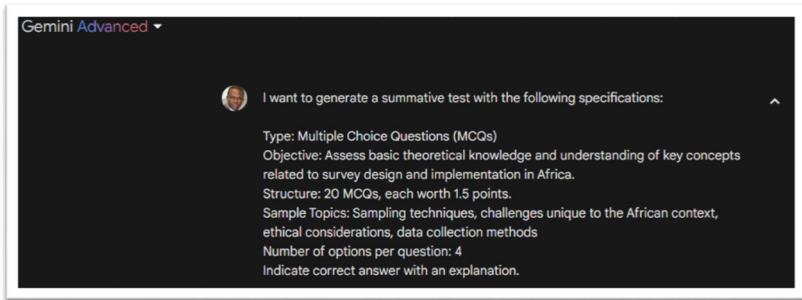
3. Which of these is a primary ethical consideration when implementing surveys in Africa?
- (A) Anonymizing collected data to protect participant identities
 - (B) Ensuring the survey tool is offered in the dominant local language
 - (C) Providing participants with a financial incentive for their time
 - (D) Using open-ended questions to avoid limiting the participants' responses

Answer: A (Anonymizing collected data to protect participant identities) Explanation: While other options are important considerations, the protection of participant confidentiality is a foundational ethical principle of research.

4. A researcher aims to collect in-depth qualitative data about community perceptions of a new public health initiative in a sub-Saharan African country. Which data collection method is most suitable?
- (A) Online questionnaire
 - (B) Telephone survey
 - (C) Focus group discussions
 - (D) Sampling of government records

Answer: C (Focus group discussions) Explanation: Focus groups facilitate the exploration of perspectives and experiences in a rich, nuanced way, well-suited to understanding community perceptions.

Google Gemini Screen



There are two important issues here. First, the human instructor needs to review both the proposed items and the options with a view to making changes as necessary. For example, the clarity of item 2 in Prompt 4.2 could be improved by modifying the options. The wording of option B, the correct answer, does not read like a challenge. Secondly, the tool obviously did not generate all 20 requested questions the first time. This issue is frequently encountered but it is not hard to circumvent. The user could readily expand upon the initial sample of items by asking for more. More generally, it may be more efficient to request for a large number of MCQs in small batches of, say, five items.

Interestingly, Google Gemini provided some unsolicited important considerations and additional tips which represent good advice. This reflects the expanding capabilities of GenAI tools to provide meaningful output. The important considerations include different aspects, namely:

- (i) Distribution: How will you administer the test (paper-based, online platform, etc.)?
- (ii) Question variety: Include questions assessing different levels of knowledge (recall, application, analysis).
- (iii) Clarity: Ensure your questions and explanations are concise and unambiguous.
- (iv) Alignment with learning outcomes: Make certain the test questions reflect the specific content covered in the survey design course.

For additional tips, Google Gemini recommended as follows:

- (i) Start with a solid understanding of Bloom's Taxonomy to design questions of varying difficulty.
- (ii) Get feedback from an expert familiar with survey design and African contexts to validate your test.
- (iii) Pilot test your questions with a small group to ensure clarity.

4.5 CASE 3: CREATION OF PERSONALISED SUMMATIVE ASSESSMENTS

The above example, like typical assessments, does not take the diversity of learners into account. However, due to learner diversity, the creation of personalised summative tests should be seen as a crucial aspect of modern education. This is because today's learning environment is often internationalised and diverse. Especially in higher education, learners often come from diverse cultural backgrounds and possess varying abilities. In this context, we can think of several purposes that personalised tests serve. Firstly, they help to ensure that each learner's unique needs and circumstances are considered. Secondly, personalised tests can help to level the playing field such that all learners are assessed fairly and accurately. Thirdly, being tailored to match the learner's proficiency level, learning style, and cultural context, personalised assessment tasks provide a more accurate measure of the learner's understanding and skills. Finally, personalised tests can enhance learner engagement and motivation. When learners see that the assessments are tailored to their needs and reflect their learning journey, they are likely to become more invested in the

learning process. This facilitates the transition from mechanistic learning to competence and application.

Artificial Intelligence (AI) can play a key role in the creation of personalised tests. For example, AI algorithms can analyse background data about a wide range of learners, including their past performance, learning preferences, and engagement levels. This data can then be used to generate assessments that are tailored to the learners' specific needs and abilities. For instance, if a learner struggles with a particular topic, the AI could generate additional questions on that topic to help the learner improve. If a learner excels in a certain area, the AI could provide more challenging questions to further stretch their abilities. The approach described here is not fundamentally different from the well-known computerised adaptive tests (CAT) used in standardised and certification tests, such as the Graduate Management Admission Test (GMAT), a globally recognised aptitude test. However, to calibrate them reliably, CATs require a considerable amount of “pilot” data, that is, a large pool of questions pre-tested on typical test takers (Wainer & Mislevy, 2000). For this reason, they are expensive to develop and are subject to major logistical, reliability and ethical concerns. The unparalleled ability of AI systems to analyse exceptionally large and complex data can help to overcome these problems. Admittedly, the implementation of such systems will require a sizeable initial investment, preferably at institutional level.

On a less sophisticated level, instructors can use commonly available GenAI tools to create personalised tests that take into account their students' cultural background, learning styles, and varied abilities. Such personalisation can take place on different scales. For example, the instructor may consider the different clusters within their student cohorts and develop summative assessments that reflect this differentiation. While this approach does not personalise the test at individual level (this is taken up in Chapter 5), it helps to make the assessment responsive to learners' group characteristics. To illustrate this approach, we will ask ChatGPT to generate an essay question on a topic, with variations based on students' language characteristics. Let us return to the subject of field surveys to illustrate how this may be done. An example prompt (Prompt 4.3) is as follows: *“I have a student cohort that comprises a group of 50 students. 30% of them are non-native English speakers. However, they are*

all required to take a final assessment comprising a 1000-word essay in English. The topic that has been taught is how to design and implement field surveys in resource-constrained contexts. Propose an essay topic for the student cohort, with two variations based on the students' characteristics". This prompt (Prompt 4.3) generated an essay topic for each student group and the motivation for the topic. Instructors can leverage their expertise to structure the assessment and frame the topics appropriately.

Two observations can be made about the output of Prompt 4.3: One, the two topic variants are comparable in terms of how much effort the learners have to expend. Two, each topic takes advantage of the unique characteristics of the group of learners and challenges them accordingly. While the native speakers are challenged to think broadly about creative approaches to field surveys when resources are limited, the non-native speakers are challenged to think about the same problem but with a focus on language barriers with which they are likely to be more familiar. As illustrated in previous chapters, we may ask ChatGPT to produce the output in different languages or for learners with a learning disability, say, dyslexia.

Prompt 4.3: Ideas for Creating Personalised Exam

ChatGPT Prompt

I have a student cohort that comprises of a group of 50 students. 30% of them are non-native English speakers. However, they are all required to take a final assessment comprising a 1000-word essay in English. The topic that has been taught is how to design and implement field surveys in resource-constrained context. Propose an essay topic for the student cohort, with two variations based on the student characteristics.

ChatGPT Response

Given the diverse composition of your student cohort and the requirement to write an essay in English on the topic of designing and implementing field surveys in resource-constrained contexts, it's essential to consider essay topics that are accessible to all students while also allowing them to explore the subject matter deeply. Here are two variations of an essay topic designed to cater to the different characteristics of your students:

General Essay Topic:

"Challenges and Strategies in Conducting Field Surveys in Resource-Constrained Environments"

Variation 1: For Non-Native English Speakers

"Understanding and Overcoming Language Barriers in Field Surveys within Resource-Limited Settings"

This variation focuses on language barriers, which is particularly relevant for non-native English speakers. It encourages these students to explore and articulate their understanding of the complexities involved in conducting field research in environments where resources are scarce, and communication challenges might compound the difficulties faced. This topic allows them to draw from personal experiences or concerns related to language and communication, which can enrich their analysis and make the essay writing process more relatable and engaging.

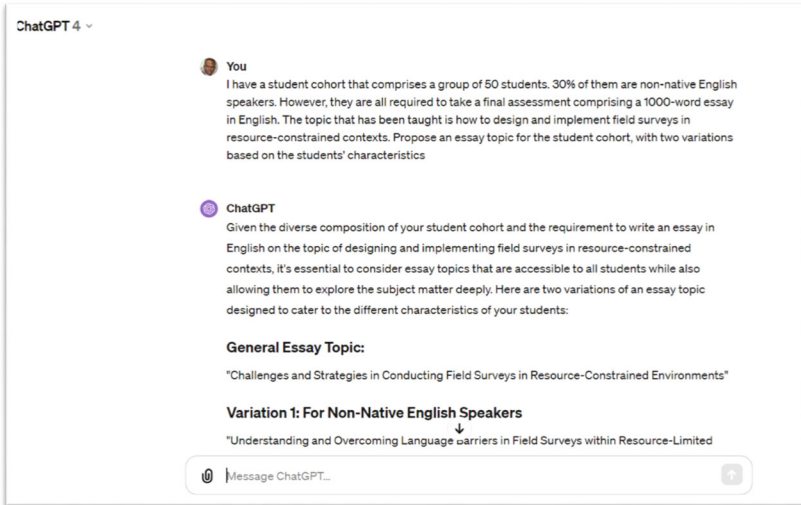
Variation 2: For Native English Speakers

"Innovative Approaches to Data Collection and Analysis in Field Surveys Under Resource Constraints"

This version emphasizes the innovation and creativity required to effectively carry out field surveys in environments where resources are limited. It prompts students to think critically about methodological adaptability, technological solutions, and strategic planning to overcome such constraints. This topic is designed to push native English speakers to delve into more complex analytical and problem-solving aspects of the subject, encouraging a deeper engagement with the material.

By tailoring the essay topics to address specific aspects of the main theme, students can leverage their unique backgrounds and perspectives, ensuring that all students, regardless of their native language, can engage with the assignment in a meaningful way. This approach also fosters inclusivity and recognizes the diverse strengths and challenges of the student body.

ChatGPT Screen



4.6 CASE 4: CREATION OF ASSESSMENT RUBRICS

An assessment rubric is a structured guide used to evaluate student performance against a set standard. Assessment rubrics provide a way to standardise the grading of assessments and make them transparent, fair, and consistent. Moreover, the use of rubrics helps to streamline the grading process, and to communicate clear and concise expectations for students. This clarity helps students understand what is expected of them and how they can achieve their learning goals. Thus, rubrics help to enhance the quality and impact of assessments. Also, rubrics help to provide feedback to both students and teachers, the former in terms of where their learning is deficient and the latter in terms of where teaching content and delivery could be improved. For these reasons, assessment rubrics have become a staple in the context of teaching and learning, especially in higher education (Chowdhury, 2019; Ragupathi & Lee, 2020; Reddy & Andrade, 2010).

Developing a reliable rubric is an involving process, especially the first time it is done for a particular course or module. From a practical point of view, there are several steps and considerations for the instructor (Table 4.2). A rubric must, first and foremost, match the learning outcomes that is being assessed. In Chapter 2 we discussed how ChatGPT could be used to support the development of learning outcomes. At the same time, the tool could be used to support the development of a rubric to assess students in relation to the learning outcomes. Notwithstanding, the expertise of the instructor is required in each step of the process. For example, to review and refine the rubric, the instructor may choose to pre-test it on a random set of students' work or ask for feedback from colleagues. These activities cannot be completely outsourced to AI.

In the remainder of this section, we will focus on illustrating how educators can save time by leveraging AI tools like ChatGPT in developing rubrics for summative assessment. However, beyond helping to develop rubrics, AI tools make it easy to implement ongoing updates and modifications to the rubric based on student performance and feedback. This ensures the assessment process stays relevant to students' changing needs and the overall learning goals in today's rapidly evolving educational context.

Table 4.2 Steps and considerations for developing an assessment rubric

| <i>Activity</i> | <i>Description and purpose</i> |
|------------------------------|--|
| Identify learning objectives | Determine what skills or knowledge the assessment should measure. These should align with the overall learning objectives of the course or unit |
| Define performance criteria | Identify the specific criteria that will be used to evaluate student performance. These could include aspects like understanding of the material, application of knowledge, critical thinking skills, etc |
| Establish performance levels | Define different levels of performance for each criterion. These levels could range from "excellent" to "needs improvement" and should provide a clear distinction between different levels of student achievement |
| Assign points or grades | Allocate points or grades to each performance level. This provides a quantitative measure for each level of performance |
| Review and refine | Review the rubric to ensure it aligns with the learning objectives and provides a fair and comprehensive assessment of student performance. Make any necessary refinements before implementing the rubric |

For our illustration, let us return to the personalised essay questions for native and non-native English speakers Prompt 4.3. A simple prompt could be passed to ChatGPT as follows: *“Generate a rubric for grading the essays. Use the grade range 0–39%, 40–49%, 50–59%, 60–69% 70–89% and > 90%”*. Note that this prompt does not include any information on learning objectives as this is expected to be embedded in the design of the assessment. Similarly, the prompt does not specify any performance criteria. However, in the output, part of which is pictured in Fig. 4.4, ChatGPT starts by defining six performance criteria upon which the rest of the output, the detailed rubric reproduced in Prompt 4.3 is based. It also attached a weight to each criterion.

This rubric contains the basic elements for comprehensive evaluation and feedback to students. From this point, the instructor has only two tasks: revising or adapting the performance criteria and editing the rubric as desired. For instance, the instructor may want to provide additional details under each grade range to offer more specific feedback to students regarding their strengths and weaknesses. It may be easier to do this in a table which we can generate by prompting ChatGPT to *“Bring the criteria and grade ranges together in tabular form. Put the grade range in columns and remember to indicate the weight of each criterion”*. The output of this prompt is reproduced in part in Prompt 4.3.

Prompt 4.4: Creating an assessment rubric

ChatGPT Prompt

Generate a rubric for grading the essays, use the grade range 0–39%, 40–49%, 50–59%, 60–69%, 70–89% and > 90%

ChatGPT Response

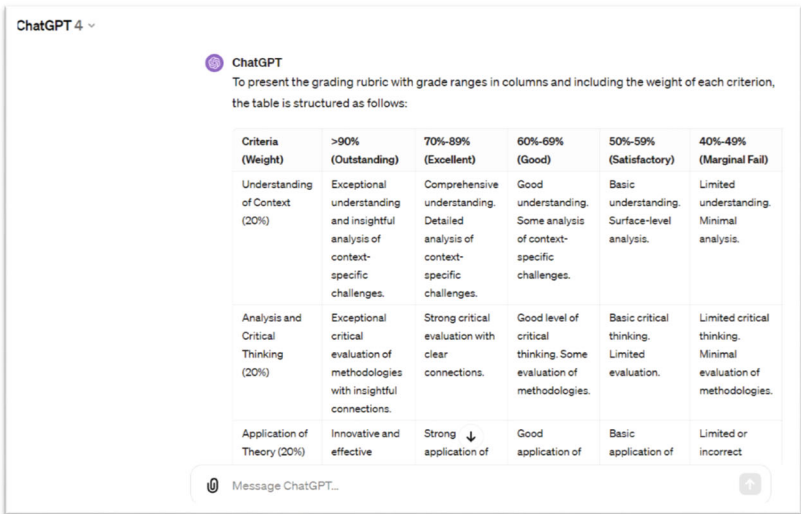
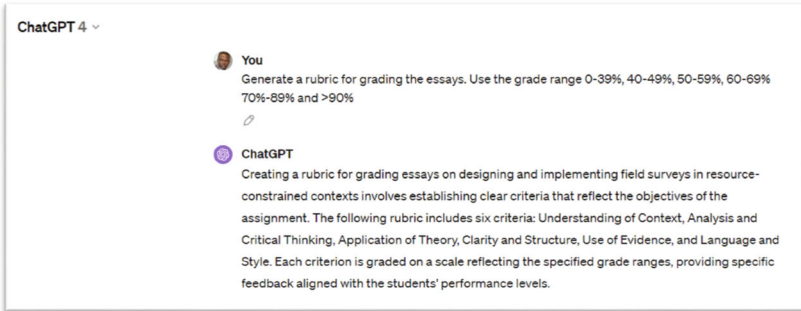
Grade Ranges

- > 90% (Outstanding)
 - Exceeds expectations in all criteria, demonstrating exceptional understanding, analysis, and application. Essays are

well-structured, with sophisticated language use and extensive evidence.

- 70–89% (Excellent)
 - Meets all criteria with high standards. Shows in-depth understanding and critical analysis, with strong application of theory. Essays are clearly structured and well-supported by evidence.
- 60–69% (Good)
 - Meets most criteria at a good standard. Demonstrates solid understanding and analysis with some minor gaps. Structure and use of evidence are adequate, with some room for improvement in language.
- 50–59% (Satisfactory)
 - Satisfactorily meets the basic requirements of the criteria. Shows understanding and some analysis but lacks depth. Structure and argumentation are adequate but could be enhanced.
- 40–49% (Marginal Fail)
 - Meets some but not all criteria, showing limited understanding and analysis. Structure is weak, and use of evidence is minimal. Language and style need significant improvement.
- 0–39% (Fail)
 - Does not meet the criteria. Lacks understanding of the context, with little to no analysis or application. Poor structure, minimal evidence, and numerous language errors.

ChatGPT Screen*



*The second screenshot showed the output when ChatGPT was asked to present the rubric in a more tabular form.

4.7 CASE 5: GRADING AND FEEDBACK TO LEARNERS

Summative assessments are not an end in themselves. Rather, they form part of the overall teaching and learning experience the aim of which is to enhance the learner's knowledge, skills, and competencies. As part of the

overall experience, learners will benefit from specific and actionable feedback that clearly highlights their strengths and weaknesses. In this sense, an educator can consider the learner's submission in response to summative assessment tasks as input into the pedagogy process. The concomitant output should be feedback on the submission. This is how learners will know what they are doing well and where they need to improve.

To provide feedback, students' work must first be scored. This is a non-trivial task especially where it involves written essays and large student cohorts. Apart from the widely discussed challenge that students sometimes submit AI-generated content, it takes a significant amount of time and effort to grade a large number of essays. It is even more demanding to provide personalised feedback that address each learner's needs. This is where automated essay scoring (AES) can be especially useful. As we argued in Kolade et al. (2024), it is "appealing to deploy AI tools in automatically scoring and providing feedback on assessment tasks such as essays and computer codes". This argument is consistent with previous research findings that automated assessments are indistinguishable from human grading and offer a potentially useful complement to the human teacher (Fathi et al, 2024; Vittorini et al, 2021). Even the more critical analyses (e.g., del Gobbo et al., 2023; Gardner et al, 2021) admit that AES systems are promising even if they are not yet sufficiently advanced for reliable assessment of higher-order learning as in written essays and oral presentations. This underscores the need for educators to complement rather than substitute key teaching activities with AI tools.

While current GenAI tools may not necessarily have all the capabilities of specialised AES systems, they have significant potentials that educators may already exploit. Thus, it is useful to reflect on the capabilities of existing AI tools to support grading and feedback to learners. For example, GenAI models like ChatGPT can be trained (see Chapter 3 on how to create custom GPTs) to understand the criteria of what makes an essay good or bad based on previous examples. This will not replace human judgement, but it will speed up the grading process by providing initial scores. It can also help in identifying strengths and areas that need improvement. Based on this, it can generate personalised feedback for each student. For example, if a student has not provided enough evidence to support their argument, the tool will point this out and suggest ways to incorporate more evidence. The remarkable ability of AI tools to consistently apply the same criteria is useful for ensuring the fairness of assessment and feedback. Instructors can take these as a basis for more

detailed feedback based on their nuanced understanding, expertise, and critical thinking.

To illustrate, consider an essay topic that requires students to write on the relationship between human capital and economic growth. We start by asking ChatGPT to generate a rubric, using the exact same criteria, criteria weights, and grade ranges applied in the previous section. We then upload¹ an essay produced by one of the authors² and ask ChatGPT to grade it against the rubric and to generate detailed feedback (Fig. 4.6). The full grading and feedback are reproduced in Box 4.6. Clearly, the level of detail is extensive and beyond what the instructor wishes to manage, especially with large student cohorts. We could reduce the amount of detail generated by revising the prompt accordingly, limiting, for example to the overall grade and feedback.

It is also possible to use GenAI to support feedback to learners in a more interactive manner. For example, the instructor could provide a set of key evaluative statements and ask ChatGPT to generate a more elaborate write-up. Notwithstanding, as we have repeatedly emphasised throughout the book, the use of ChatGPT or any other GenAI tool cannot substitute for the expertise and nuanced knowledge of the human instructor. AI-generated feedback as typified in Prompt 4.4 should be taken as a starting point by the human instructor. The output itself creates the room for this, as the quantitative scores are provided in ranges and approximations rather than the single number that the human instructor needs to specify.

Prompt 4.5: Essay scoring with ChatGPT

ChatGPT Prompt

Apply the rubric to grade the attached essay. In addition to the overall score, provide detailed feedback on what is done well and what needs to be improved.

ChatGPT Response

Based on the detailed essay provided, here is the grading using the specified rubric:

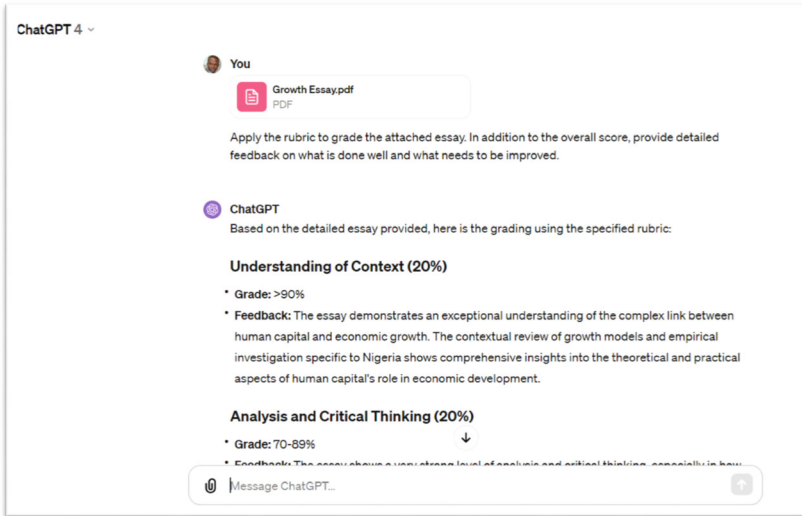
¹ At the time of writing, the capability to accept uploaded files was not available in the free version, ChatGPT-3.5.

² See <http://dx.doi.org/10.13140/RG.2.2.16248.87046>. The essay was written more than a decade ago and is similar to a typical student essay.

- Understanding of Context (20%)
 - Grade: > 90%
 - Feedback: The essay demonstrates an exceptional understanding of the complex link between human capital and economic growth. The contextual review of growth models and empirical investigation specific to Nigeria shows comprehensive insights into the theoretical and practical aspects of human capital's role in economic development.
- Analysis and Critical Thinking (20%)
 - Grade: 70–89%
 - Feedback: The essay shows a very strong level of analysis and critical thinking, especially in how it critiques and compares various models of economic growth with human capital. However, some areas, such as the exploration of negative findings related to human capital accumulation, could benefit from deeper analysis to challenge or support existing theories.
- Application of Theory (20%)
 - Grade: 70–89%
 - Feedback: There is a very strong application of theoretical models to the empirical context of Nigeria. The essay skillfully navigates through complex theories, applying them to real-world data. Nonetheless, a more nuanced discussion on the applicability of these theories to diverse economic contexts could enhance the essay.
- Clarity and Structure (15%)
 - Grade: > 90%
 - Feedback: The essay is exceptionally well-structured, with a clear introduction, literature review, methodology, analysis, and conclusion. The logical flow and organization of content significantly enhance the reader's understanding and engagement with the topic.

- Use of Evidence (15%)
- Grade: 70–89%
 - Feedback: The use of empirical data, especially from Nigerian sources, is very strong and well-integrated into the analysis. The essay effectively supports its arguments with evidence. Further diversification of sources and evidence, particularly from comparative international contexts, could provide additional depth.
 - Language and Style (10%)
- Grade: > 90%
 - Feedback: The essay is written in an engaging, clear, and precise academic style, virtually free of errors. The language used effectively communicates complex ideas and contributes positively to the essay's arguments.
 - Overall Score and Feedback
- Overall Score: 85%
- General Feedback: This essay stands out for its exceptional understanding and presentation of the link between human capital and economic growth, particularly within the Nigerian context. It excels in clarity, structure, and language use. To elevate the essay further, more critical engagement with the negative findings related to human capital accumulation and broader application of theory could provide additional depth and insight. Overall, this is an impressive analysis that contributes meaningfully to the academic discussion on economic development and human capital.

ChatGPT Screen



4.8 CONCLUSION

In this chapter, we have examined how the creation and grading of summative assessments could be enriched with GenAI. Indeed, the capacity of AI to automate routine evaluative tasks enables educators to dedicate more time to nuanced pedagogical engagements, thereby elevating the learning experience. Furthermore, custom GPTs enable personalised learning and assessment which help to optimise educational outcomes. However, our exploration of the application of GenAI in summative assessment does not ignore the challenges associated with these technologies, such as the risk of academic dishonesty and overreliance that could ultimately enfeeble the students. We acknowledge these challenges and emphasise the need for the human educator to take GenAI for what it is, an assistant rather than a substitute.

REFERENCES

- Chowdhury, F. (2019). Application of rubrics in the classroom: A vital tool for improvement in assessment, feedback and learning. *International Education Studies*, 12(1), 61–68.
- del Gobbo, E., Guarino, A., Cafarelli, B., Grilli, L., & Limone, P. (2023). Automatic evaluation of open-ended questions for online learning. A systematic mapping. *Studies in Educational Evaluation*, 77, 101258. <https://doi.org/10.1016/j.stueduc.2023.101258>
- Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., & Zahidi, S. (2023). *Future of jobs report 2023*. Geneva: World Economic Forum. <https://www.weforum.org/reports/the-future-of-jobs-report-2023>
- Dixon, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory Practice*, 55(2), 153–159. <https://doi.org/10.1080/00405841.2016.1148989>
- Egbetokun, A. (2022). Conducting surveys in Africa: Reflections from national surveys in Nigeria. In O. Kolade, D. Rae, D. Obembe, & K. Woldesenbet Beta (Eds.), *The Palgrave handbook of African entrepreneurship* (pp. 573–594). Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-75894-3_23
- Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners’ speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, 121, 103254.
- Gardner, J., O’Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: ‘Breakthrough? Or buncombe and ballyhoo?’ *Journal of Computer Assisted Learning*, 37(5), 1207–1216. <https://doi.org/10.1111/jcal.12577>
- Iarossi, G. (2006). *The power of survey design: a user’s guide for managing surveys, interpreting results, and influencing respondents*. Washington, DC: World Bank. <http://hdl.handle.net/10986/6975>
- Kolade, O., Owoseni, A., & Egbetokun, A. (2024). Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework. *Heliyon*, 10(4). <https://doi.org/10.1016/j.heliyon.2024.e25953>
- Man-Sze-Lau, A. (2016). ‘Formative good, summative bad?’—A review of the dichotomy in assessment literature. *Journal of Further and Higher Education*, 40(4), 509–525. <https://doi.org/10.1080/0309877X.2014.984600>
- Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. In S. S. Shanger & N. W. Gleason (Eds.), *Diversity and inclusion in global higher education: Lessons from across Asia* (pp. 73–95). London: Palgrave Macmillan. <https://doi.org/10.1007/978-981-15-1628-3>

- Rawlusk, P. E. (2018). Assessment in higher education and student learning. *Journal of Instructional Pedagogies*, 21, 1. <http://www.aabri.com/copyright.html>
- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, 35(4), 435–448. <https://doi.org/10.1080/02602930902862859>
- Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education*, 33(1), 57–69. <https://doi.org/10.1080/03098770802638671>
- Vittorini, P., Menini, S., & Tonelli, S. (2021). An AI-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education*, 31(2), 159–185. <https://doi.org/10.1007/s40593-020-00230-2>
- Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing: A primer* (pp. 61–100). Milton Park: Routledge.