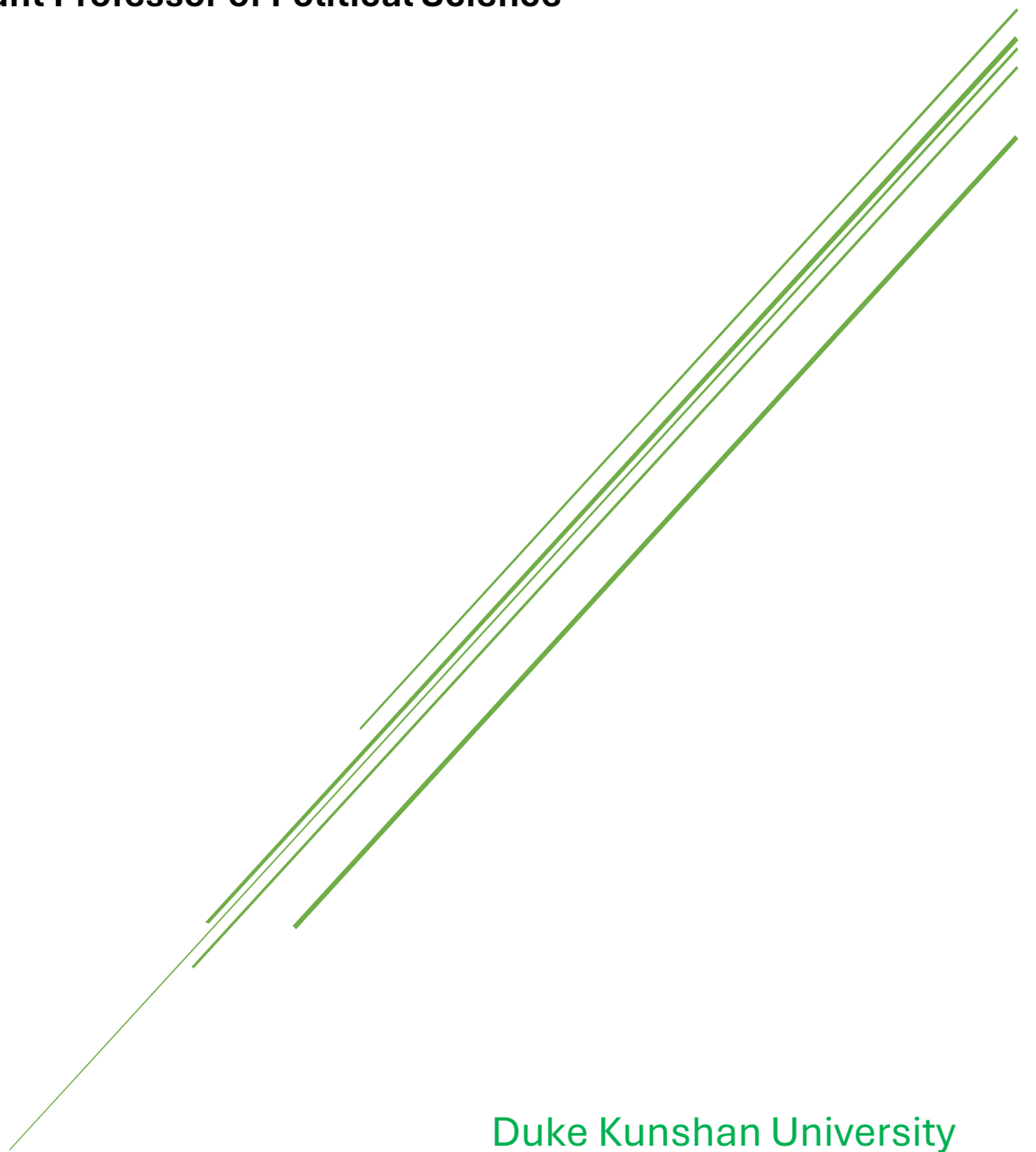


# AUTHENTIC ASSESSMENT:

## EVIDENCE, MECHANISMS, AND IMPLICATIONS FOR TEACHING PRACTICE AT DKU

**Lie Philip Santoso**

**Assistant Professor of Political Science**



Duke Kunshan University  
Institute for Global Higher Education

## Table of Contents

|                                                         |    |
|---------------------------------------------------------|----|
| Executive Summary.....                                  | 2  |
| Introduction .....                                      | 3  |
| Mechanisms: How and Why Authentic Assessment Works..... | 4  |
| Confidence in the Evidence Base .....                   | 7  |
| Actionable Recommendations for DKU Faculty .....        | 15 |
| Conclusion.....                                         | 18 |
| References .....                                        | 19 |

## Executive Summary

Over the past two decades, higher education has increasingly shifted away from assessment practices centered on the recall of isolated facts toward approaches that emphasize students' ability to use knowledge in meaningful ways. Authentic assessment (AA) has emerged from this shift as a broad family of practices designed to evaluate learning through tasks that resemble real-world, professional, civic, or social activities. Rather than asking students to demonstrate what they know in abstract or decontextualized formats, authentic assessment asks them to apply knowledge, integrate multiple skills, and exercise judgment under conditions that mirror how learning is used beyond the classroom.

Interest in authentic assessment has grown alongside broader shifts in higher education toward applied learning, transferable skills, and preparation for life beyond the classroom. Advocates argue that traditional examinations often fail to capture these goals, whereas authentic tasks—such as projects, simulations, or research-based assignments—better reflect how knowledge is used in practice and align more closely with institutional learning objectives.

The literature suggests that authentic assessment is most consistently associated with higher student engagement and motivation. When assessment tasks are clearly connected to real-world problems or social purposes, students are more likely to see their work as meaningful and to invest sustained effort. Authentic tasks may also promote deeper cognitive engagement by requiring students to synthesize information, evaluate evidence, and justify decisions, particularly when they are supported by feedback and opportunities for reflection.

At the same time, the empirical evidence supporting authentic assessment is mixed. Much of the existing research relies on small-scale studies, student self-reports, and instructor-developed evaluation criteria, with limited use of validated learning measures or strong causal designs. While students often report positive experiences with authentic assessment, there is less consistent evidence that it reliably produces superior or more durable learning outcomes across contexts.

A central finding in the literature is that the effectiveness of authentic assessment depends heavily on implementation. The impact of AA is shaped less by the surface realism of tasks than by the alignment between learning goals, task design, evaluation criteria, and feedback. When authentic tasks are evaluated using narrow or procedural criteria, their pedagogical potential is weakened. When evaluation criteria explicitly reward reasoning, judgment, and reflection, authentic assessment is more likely to influence how students engage with learning.

Overall, the evidence supports a cautious and context-sensitive approach to authentic assessment. Rather than treating AA as a universal or superior replacement for traditional assessment, it is best understood as one tool among many, particularly well-suited to courses that emphasize application, synthesis, and interpretation. For institutions like DKU, encouraging selective, well-aligned, and reflective use of authentic assessment is most consistent with the current evidence.

## Introduction

Educational assessment has undergone a significant transformation in recent decades. Where assessment once focused primarily on standardized tests designed to measure discrete units of knowledge, contemporary approaches increasingly emphasize students' ability to apply knowledge in meaningful contexts (Shepard, 2000; Darling-Hammond & Adamson, 2014). Authentic assessment (AA) emerges from this shift by redefining what it means to demonstrate learning: it refers to assessment practices that require students to demonstrate knowledge and skills through tasks that resemble real-world, professional, civic, or social activities. Unlike traditional assessments that emphasize the recall of discrete facts or the execution of standardized procedures, AA asks students to apply knowledge in context, integrate multiple competencies, and exercise judgment in conditions that reflect how learning is used beyond the classroom (Wiggins, 1990; Lombardi, 2008).

Research suggests that AA is associated with several educational benefits that extend beyond short-term academic performance. First, AA is widely understood to promote deeper learning. By engaging students in realistic and often complex tasks, authentic assessment encourages conceptual understanding, synthesis, and transfer, rather than surface-level memorization. Studies argue that such context-rich tasks support more durable and flexible learning outcomes (Herrington & Herrington, 2006; McArthur, 2023).

Second, AA is linked to higher levels of student motivation and engagement. When assessment tasks are clearly connected to real-world problems or social purposes, students are more likely to perceive their learning as meaningful and worthwhile. This relevance can increase persistence, effort, and intellectual curiosity, particularly when students understand how their work might be used or evaluated beyond the classroom setting (Darling-Hammond & Adamson, 2014; Lombardi, 2008).

Third, authentic assessment is often credited with strengthening students' understanding of the broader significance of their learning. By embedding assessment within real or simulated social, professional, or civic contexts, AA encourages reflection on both process and consequence. This reflective component is associated with the

development of critical thinking, ethical reasoning, and self-awareness—capacities that are increasingly emphasized in higher-education learning outcomes frameworks (Gulikers et al., 2004).

Finally, AA is commonly framed as a tool for improving post-graduation readiness. The literature highlights its role in developing transferable skills such as communication, collaboration, problem-solving, and project management, as well as adaptive capacities needed in changing professional and social environments (Lombardi, 2008; McArthur, 2023). Importantly, proponents argue that the value of AA lies not only in simulating workplace tasks, but in helping students see themselves as capable contributors to society with responsibilities beyond academic success (McArthur, 2023).

### **Mechanisms: How and Why Authentic Assessment Works**

Researchers often highlight three key features of authenticity (Gulikers et al., 2004; Herrington & Herrington, 2006):

- *Realism*: Tasks should mirror the complexity and uncertainty of real-life problems, where there is rarely a single “correct” answer.
- *Contextualization*: Assessments are embedded in real or simulated environments, so that students see a clear connection between classroom learning and the situations they may encounter in their professional or civic lives.
- *Problematization*: Students are challenged to think critically about open-ended or ill-defined problems, weighing evidence, considering multiple perspectives, and justifying their choices.

These features are not merely descriptive characteristics of authentic tasks; they are the channels through which authentic assessment is expected to operate. Each activates one or more of the mechanisms outlined below.

At a theoretical level, authentic assessment is expected to influence learning through four interrelated mechanisms: (1) motivational engagement, (2) cognitive processing, (3) metacognitive regulation, and (4) normative signaling about what counts as valued

knowledge. These mechanisms are activated through specific features of authentic assessment design.

#### *Motivational Mechanism: Perceived Relevance and Ownership*

A central mechanism underlying authentic assessment is increased motivation through perceived relevance. Contextualization plays a key role here. When assessment tasks are embedded in real or simulated social, professional, or civic contexts, students are more likely to see them as purposeful rather than arbitrary. This perception of relevance increases intrinsic motivation and encourages students to invest greater effort and persist through difficulty (Darling-Hammond & Adamson, 2014; Lombardi, 2008).

The mechanism operates as follows: contextualized tasks signal that the goal of learning is not merely to satisfy course requirements, but to engage with problems that matter beyond the classroom. As a result, students experience greater ownership over their work, which in turn leads to higher engagement and sustained effort. This mechanism affects *how much* and *how persistently* students engage with learning, rather than guaranteeing deeper learning on its own.

#### *Cognitive Mechanism: Deeper Engagement Through Problem Complexity*

Authentic assessment is also theorized to affect learning through cognitive mechanisms related to task structure. Realism and problematization are central here. Authentic tasks are typically open-ended, ill-defined, and characterized by uncertainty—features that require students to integrate knowledge, weigh evidence, and make justified choices rather than apply fixed procedures (Gulikens et al., 2004; Herrington & Herrington, 2006).

This structure promotes deeper cognitive processing by forcing students to move beyond surface-level strategies such as memorization or algorithmic application. Instead, students must organize information, connect concepts across domains, and adapt knowledge to novel situations. In principle, this form of cognitive engagement is more likely to support conceptual understanding and transfer than tasks with predetermined correct answers.

However, this mechanism depends critically on task design. If tasks appear realistic but remain cognitively routinized in execution, the opportunity for deep processing is diminished. Authenticity, therefore, operates not simply through resemblance to real-world activity, but through the cognitive demands imposed by the task.

#### *Metacognitive Mechanism: Reflection, Feedback, and Self-Regulation*

A third mechanism involves metacognitive regulation. Problematization, when paired with structured opportunities for feedback and reflection, encourages students to monitor their understanding, evaluate the quality of their reasoning, and consider alternative approaches.

Authentic assessment often unfolds over extended periods and incorporates revision, peer review, or self-assessment. Through these processes, students are prompted to articulate not only *what* they did, but *why* they made particular choices and how their work could be improved. This reflective dimension supports the development of self-regulated learning skills, which are essential for long-term learning and adaptability. Importantly, these benefits depend on whether reflection is explicitly scaffolded rather than assumed to emerge automatically.

#### *Normative Mechanism: Signaling What Counts as Learning*

Finally, authentic assessment operates through a normative mechanism by signaling what forms of knowledge and performance are valued. Traditional assessments often communicate that success lies in reproducing information or following procedures correctly. In contrast, authentic assessment—particularly when it emphasizes *realism and problematization*—signals that reasoning, judgment, communication, and the ability to justify decisions are central to what it means to “know” something.

This signaling effect matters because students adapt their learning strategies to what they believe will be rewarded. When assessment criteria align with authentic goals, students are more likely to allocate effort toward conceptual understanding and transferable skills. Conversely, when authentic tasks are evaluated using narrow or procedural criteria, this signal is weakened or reversed, leading students to prioritize surface-level performance.

### *Take-Away Points*

In sum, the discussion above shows why authentic assessment does not reliably produce uniform learning gains across contexts. The effectiveness of authentic assessment depends not on the surface realism of tasks alone, but on whether realism, contextualization, and problematization jointly activate motivational, cognitive, metacognitive, and normative mechanisms in a coherent way. When these features are well aligned with evaluation criteria and feedback structures, authentic assessment can meaningfully alter how students engage with learning and what they perceive as valued knowledge. When this alignment is absent, authenticity risks becoming cosmetic rather than pedagogically consequential. This framework helps explain the mixed empirical findings in the literature and underscores the importance of implementation rather than authenticity per se.

### **Confidence in the Evidence Base**

This section evaluates how robust the existing research on AA is, with particular attention to methodological limitations common in education research. While the theoretical appeal of authentic assessment is widely acknowledged, confidence in the empirical evidence supporting its effects on student learning remains limited. This is not unusual for education research, which often faces structural challenges such as small samples, difficulty with randomization, contextual variation across courses and disciplines, and heavy reliance on self-reported outcomes. Nevertheless, these constraints shape what can be concluded about the effectiveness of authentic assessment and where caution is warranted.

### *Overall Strength of the Evidence*

Although many studies report positive outcomes associated with authentic assessment, much of the evidence rests on student self-reports and instructor-developed rubrics rather than on independent or validated measures of learning (Gulikers et al., 2008; Segers et al., 2008). Students frequently report feeling more engaged, motivated, or confident when working on authentic tasks, and instructors often observe richer classroom discussions or more creative outputs. These findings

are encouraging, but they capture perceptions rather than direct evidence of improved knowledge or skill acquisition. As Brookhart (2013) cautions, positive experiences with assessment should not be conflated with demonstrable gains in conceptual understanding or long-term competence.

Scholars have repeatedly noted that the field lacks large-scale, systematic studies that clearly link authentic assessment to measurable improvements in learning outcomes (Jonsson & Svingby, 2007; Martínez-Argüelles et al., 2023). Many studies remain small in scale, situated within a single course or institution, and rely on locally developed evaluation tools that raise concerns about validity and generalizability. Even when quasi-experimental designs are attempted, reported gains tend to emphasize perceived employability skills, confidence, or engagement rather than objective indicators of learning (Li et al., 2023).<sup>1</sup> Systematic reviews similarly conclude that, while authentic assessment is consistently associated with higher levels of student enthusiasm, there is little causal evidence that it produces more durable or transferable learning outcomes.

Taken together, the literature suggests that authentic assessment enjoys strong intuitive and experiential support but rests on a comparatively weak evidentiary foundation when evaluated against stricter standards of measurement and causal inference.

### *Measurement Challenges: What Is Being Measured?*

A central limitation in the study of authentic assessment is the lack of agreement on how it should be measured. While there is broad consensus about AA's theoretical goals—such as promoting deep learning, critical thinking, and transferable skills—scholars differ markedly on what constitutes an authentic task, how such tasks should be evaluated, and whether existing measures capture what they purport to measure.

---

<sup>1</sup> A quasi-experimental study is a type of research design that looks a lot like an experiment but does not use random assignment of participants to groups.

- In a true experiment, researchers randomly assign participants to treatment and control groups, which helps ensure the groups are comparable.
- In a quasi-experiment, participants are placed in groups based on pre-existing conditions (like classes, schools, or programs), not by random assignment.

These disagreements stem from divergent theoretical commitments and pedagogical aims, and they have contributed to a fragmented and inconsistent evidence base.

Disagreement first arises over what counts as an “authentic” task. Some scholars, drawing on situated learning theory, emphasize performance-based assessments that closely mirror professional practice, such as laboratory investigations, policy briefs, or simulations of civic deliberation (Mueller, 2005). Others, working from a constructivist perspective, emphasize student agency, reflection, and identity formation, treating reflective journals, peer assessments, or self-evaluations as equally authentic (Torrance, 2007; Andrade, 2010). These competing definitions mean that studies labeled as investigating authentic assessment may in fact examine substantively different instructional practices, undermining comparability across studies.

Even when there is agreement on task design, there is often a disconnect between the authenticity of the task and the authenticity of the assessment criteria. Gulikers et al. (2004) outline multiple dimensions of authenticity, including not only the task itself but also the assessment form and evaluation criteria. In practice, however, many studies focus on designing realistic tasks while retaining grading schemes that reward procedural correctness or reproduction of established methods. This misalignment weakens construct validity<sup>2</sup> and limits the extent to which observed outcomes can be attributed to authentic assessment as theorized.

Although this literature devotes substantial attention to the design of authentic tasks, far less consensus exists about how authenticity should be operationalized and assessed once such tasks are in place.

#### *How Is Authenticity Evaluated?*

Gulikers et al. (2004) provide a useful framework for understanding how authenticity can break down at the level of evaluation. They distinguish among five dimensions of authenticity—the task, physical context, social context, assessment form, and criteria—highlighting that authenticity is not a property of task design alone but also of

---

<sup>2</sup> Construct validity refers to how well an assessment actually measures the underlying concept or “construct” it is intended to measure.

how student performance is assessed. In practice, however, many studies emphasize realism in task design while retaining evaluation criteria that reward procedural correctness or reproduction of established methods. This creates a disconnect between the authenticity of the task and the authenticity of the assessment itself, weakening construct validity and limiting the extent to which observed outcomes can be attributed to authentic assessment as theorized.

This can create subtle but important misalignments. For example, students might be asked to conduct a small-scale scientific inquiry project—designing an experiment, collecting data, and reporting their findings. While the task itself is authentic in design, the grading may emphasize whether students applied the “correct” method or reproduced established practices, rather than how well they reasoned through their choices or interpreted their results. In such cases, students may learn that compliance with a procedure matters more than engaging with the underlying concepts, which limits the task’s potential to foster critical thinking or adaptability.

A more appropriate, or “model,” evaluation would align grading criteria with the intended learning outcomes of the task. In the case of a scientific inquiry project, this means assessing whether the research question is clearly stated and appropriately scoped, whether the methods chosen are reasonable and applied correctly within the resources available, and whether the data are presented accurately and used to support claims. Attention would also be given to the students’ reasoning: are the conclusions logically connected to the evidence, and do they show an understanding of the underlying concepts rather than just the ability to reproduce established procedures? Finally, a strong evaluation would include some recognition of the project’s limitations, with students encouraged to reflect on how their approach could be improved. Taken together, these criteria ensure that assessment captures not only the technical execution of the task but also the higher-order skills of interpretation, judgment, and reflection that authentic assessment is designed to foster.

While such model evaluation criteria allow instructors to capture complex reasoning and reflection within the bounds of a course, they do not by themselves guarantee that these demonstrated competencies will persist once the scaffolding of the classroom is removed. In other words, even when authentic tasks are well designed and supported

by aligned evaluation criteria, an important question remains: to what extent do the skills and understandings demonstrated in these settings transfer beyond the classroom? A student may succeed in completing a laboratory project, policy brief, or design exercise within a course, but this does not necessarily mean they will be able to reproduce the same quality of work in the less structured, higher-stakes contexts they encounter after graduation. Some scholars note that transfer is one of the most difficult outcomes to achieve, as it requires not only mastery of knowledge and skills, but also the ability to recognize when and how to apply them in unfamiliar situations (Perkins & Salomon, 1992; Barnett & Ceci, 2002).<sup>3</sup> Bransford and Schwartz (1999) similarly argue that transfer depends on cultivating flexible knowledge that can be adapted across contexts rather than applied mechanically. From this perspective, authentic assessment fulfills its promise only if it helps students *generalize their learning* by continuing to use evidence-based reasoning, problem-solving, and reflection when they face novel challenges after graduation. Addressing this issue will likely require longitudinal studies and real-world follow-up to evaluate whether the gains from authentic assessments persist and translate into professional, civic, or personal settings.

Moreover, many empirical studies treat authenticity as a binary input into course design (i.e., whether an assignment is labeled “authentic”) rather than examining how authenticity is recognized and rewarded in practice. Tools developed by Kearney (2013) and Vu & Dall’Alba (2007), for example, offer frameworks for instructors to reflect on the presence of AA features in their teaching. For instance, Kearney’s framework focuses on

---

<sup>3</sup> Barnett and Ceci taxonomy (2002) proposes transfers in 2 dimensions: content and context. Content refers to *what* is being transferred—whether it is specific facts, procedures, problem-solving strategies, or higher-order principles. Context refers to *where* the transfer occurs, which they break down into multiple domains such as knowledge domains (e.g., science vs. everyday reasoning), physical settings (classroom vs. workplace), temporal settings (immediate vs. delayed), functional contexts (academic vs. professional), and social contexts (individual vs. collaborative). The taxonomy shows that transfer is not a simple yes-or-no outcome but a spectrum: students may reproduce a skill in a very similar situation (near transfer) while struggling to adapt it in a novel domain or setting (far transfer).

Applied to authentic assessment, some questions are pertinent: are students merely able to demonstrate competence within the classroom version of the task, or can they carry that knowledge into new contexts after graduation? For example, writing a policy brief for a course may represent near transfer if students apply theories learned in class to a familiar policy issue, but far transfer would require them to adapt those same skills to a professional workplace with different expectations, audiences, and constraints.

dimensions such as student ownership, relevance, and self- and peer-assessment, encouraging instructors to check whether their tasks give learners real choice, connect with meaningful contexts, and involve multiple perspectives. Similarly, Vu and Dall’Alba (2007), drawing on higher education case studies, emphasize alignment between assessment, learning activities, and professional practice, providing guiding questions rather than rigid criteria.

While useful for raising awareness, these frameworks remain largely descriptive. They help instructors identify whether authenticity-like elements are present, but they do not provide a means of evaluating quality or consistency across contexts. More importantly, few studies have tested whether the presence of these features actually correlates with improved student outcomes. This points to a deeper theoretical issue: does the mere inclusion of authentic tasks lead to learning, or does learning depend on how students are assessed and rewarded for demonstrating deeper understanding, creativity, and independent thinking? Without clearer evaluative tools, it is difficult to determine whether authentic assessment succeeds because of the design of the task itself, or because of the incentives and criteria attached to it.

#### *Validity, Reliability, and Generalizability*

Concerns about validity and generalizability further limit confidence in the evidence base. Much of the empirical literature relies on small-n case studies conducted within single institutions or classrooms (Darling-Hammond & Snyder, 2000; Herrington & Herrington, 2006). These studies frequently use instructor-developed rubrics to assess student performance. While such rubrics are pedagogically useful, they raise serious concerns about construct validity and reliability<sup>4</sup> when used as research instruments.

At the core of this issue is the assumption that individual instructors are capable of designing evaluation criteria that accurately reflect complex learning constructs such as critical thinking, collaboration, or real-world problem-solving. However, these constructs are often under-theorized or inconsistently defined across contexts. As

---

<sup>4</sup> Reliability refers to the consistency of the assessment results. An assessment is reliable if different graders would score the same student’s work similarly, or if a student would get a comparable result if they repeated the task under the same conditions.

Sadler (2009) notes, many instructors operate without an explicit framework of what constitutes quality performance, relying instead on tacit knowledge or subjective judgment. This creates a gap between the rubric and the theoretical model it purports to represent, leading to questionable construct validity.

Moreover, these rubrics often reflect the instructor's own disciplinary biases or pedagogical preferences, which may or may not align with broader educational goals or learning taxonomies (Jonsson & Svingby, 2007). For example, an instructor might place heavy weight on creativity or originality in an open-ended assignment, while neglecting domain-specific accuracy or procedural rigor. Without grounding in validated models, such as Bloom's taxonomy (Anderson & Krathwohl, 2001) or frameworks like the AAC&U VALUE rubrics (Rhodes & Finley, 2014), such assessments risk becoming idiosyncratic and non-generalizable.

Inter-rater reliability is also frequently overlooked. Because rubrics are often applied by a single instructor within a single course, it is unclear whether the same student performance would be evaluated similarly by other raters or in other contexts. When grading is unblinded and conducted by those who designed and implemented the intervention, confirmation bias becomes a serious concern (Rezaei & Lovorn, 2010). The absence of external validation or cross-context testing further limits the external validity of findings based on such measures.

#### *Reliance on Self-Reported Outcomes*

A second major limitation is the widespread reliance on self-reported perceptions of authenticity, engagement, or learning. Many studies treat students' subjective experiences as proxies for cognitive development, despite longstanding concerns about the validity of self-reports as indicators of learning. Students may report higher engagement or satisfaction because authentic tasks feel novel, flexible, or empowering, even when actual learning gains are modest or absent.

Self-reported measures have well-known limitations. Students may respond in socially desirable ways, reinterpret their experiences after the fact, or confuse enjoyment and confidence with actual learning (Podsakoff et al., 2003; Nisbett & Wilson, 1977). Recent reviews reinforce this concern. For example, in a review conducted by Papageorgiou et

al. (2024), they found that although students often value the flexibility and perceived relevance of authentic tasks, these positive perceptions do not always provide causal link to improved learning outcomes. Similarly, measures of engagement tend to show only weak relationships with more objective indicators such as grades or standardized assessments (Evans & Zhu., 2023).

As a result, much of the AA literature risks operating on an implicit assumption that positive perceptions equal educational effectiveness. Without triangulation using objective measures—such as concept inventories<sup>5</sup>, standardized performance tasks, or external evaluations—claims about learning gains remain uncertain.

### *Limits to Causal Inference*

Finally, confidence in the evidence base is constrained by limited capacity for causal inference. Most studies rely on observational designs, single-group pre–post comparisons, or uncontrolled classroom implementations. Even quasi-experimental studies often bundle authentic assessment with other instructional changes, such as new technologies, revised curricula, or increased instructor involvement, making it difficult to isolate the effect of authenticity itself.

Although some quasi-experiments report positive outcomes—particularly in healthcare or blended-learning contexts—these studies typically focus on perceived employability or engagement rather than validated learning outcomes (Martínez-Argüelles et al., 2023; Li et al., 2023)<sup>6</sup>. Randomized controlled trials remain rare, and those that do exist tend

---

<sup>5</sup> A concept inventory is a type of standardized test designed to measure students' understanding of key concepts in a subject, rather than their ability to memorize facts or follow procedures.

- It usually consists of multiple-choice questions where all answer options are based on common misconceptions as well as the scientifically correct idea.
- By analyzing which options students choose, instructors can identify not just whether students are correct, but also how they are thinking and where misunderstandings lie.
- Famous examples include the Force Concept Inventory (FCI) in physics, which reveals whether students truly grasp Newtonian mechanic using everyday language.

In the context of authentic assessment, concept inventories can be used before and after an intervention to see if students' conceptual understanding (not just their perceptions) has actually improved.

<sup>6</sup> For example, Li et al., 2023 employs a quasi-experimental mixed-methods design comparing blended learning with traditional instruction among nursing students. It reports improvements in academic performance (e.g., final examination scores) and learning engagement, as well as higher student satisfaction. However, while some objective outcomes show gains, the study shows not all skill-based measures are statistically significant or systematically validated.

to focus on narrow forms of simulation rather than generalized authentic assessment across disciplines (Chernikova et al., 2024).

Without clearer causal models specifying how authentic assessment produces learning gains, and without designs capable of isolating those mechanisms, causal claims remain tentative.

### *Summary Assessment*

In sum, while authentic assessment is supported by strong theoretical arguments and consistently positive student perceptions, the empirical evidence supporting its causal impact on learning outcomes remains limited. Much of the existing research relies on small samples, self-reported measures, instructor-developed rubrics, and weak causal designs. These limitations do not invalidate authentic assessment as a pedagogical practice, but they do warrant caution in interpreting its effects and underscore the need for more rigorous, multi-institutional, and methodologically diverse research.

### **Actionable Recommendations for DKU Faculty**

The evidence reviewed above does not support a simple endorsement of authentic assessment as a universally superior form of evaluation. Instead, it suggests a set of conditional, implementation-sensitive recommendations for DKU faculty who are considering incorporating authentic assessment into their courses. These recommendations emphasize alignment, transparency, and appropriate expectations, rather than wholesale replacement of traditional assessment practices.

#### *Prioritize Alignment Between Tasks, Criteria, and Learning Goals*

A central implication of the evidence is that the effectiveness of authentic assessment depends less on the surface realism of tasks than on the alignment between task design, evaluation criteria, and intended learning outcomes. DKU instructors who adopt authentic assessments should therefore begin by articulating clearly what kinds of knowledge and skills they aim to assess—such as conceptual understanding, reasoning, or judgment—and ensure that grading criteria explicitly reward those outcomes.

In practice, this means avoiding situations in which students are asked to complete open-ended or realistic tasks but are evaluated primarily on procedural correctness or reproduction of established methods. Instead, evaluation criteria should foreground reasoning, justification of choices, and interpretation of evidence. Making these criteria explicit to students can also help reinforce the normative signal that understanding and judgment, rather than compliance alone, are valued forms of learning.

### *Use Authentic Assessment Strategically, Not Universally*

The mixed and context-dependent nature of the evidence suggests that authentic assessment is best used selectively rather than as a default assessment mode. It is likely to be most effective in courses where learning objectives emphasize synthesis, application, and interpretation—such as project-based courses, capstone seminars, applied methods courses, or interdisciplinary offerings.

In contrast, in large introductory courses or courses with highly standardized curricula, authentic assessment may require substantial scaffolding to avoid cognitive overload or ambiguity. In such settings, authentic tasks may function best as supplements to, rather than replacements for, more structured forms of assessment. Recognizing these contextual constraints can help instructors deploy authentic assessment where it is most pedagogically appropriate.

### *Combine Authentic Tasks with Structured Feedback and Reflection*

The discussion above highlights that many of the proposed benefits of authentic assessment operate through metacognitive mechanisms that depend on feedback, reflection, and opportunities for revision. Accordingly, DKU faculty should not assume that authentic tasks will automatically foster deeper learning. Instead, they should intentionally incorporate structured feedback cycles, peer review, or reflective components that prompt students to evaluate their own reasoning and learning strategies.

Without such scaffolding, authentic tasks risk becoming time-intensive exercises that improve engagement without reliably improving learning. Reflection and feedback

should therefore be treated as integral components of authentic assessment, not optional add-ons.

### *Be Cautious in Interpreting Student Perceptions as Evidence of Learning*

Given the heavy reliance on self-reported outcomes in the existing literature, DKU instructors should be cautious about treating positive student feedback as evidence of learning gains. While student perceptions of relevance and engagement are important indicators of educational experience, they should not be conflated with improvements in conceptual understanding or transferable skills.

Where feasible, instructors may consider triangulating evidence by combining student reflections with more objective indicators, such as concept inventories, standardized rubrics (e.g., AAC&U VALUE rubrics), or externally reviewed performance tasks. Even modest efforts at triangulation can improve confidence in claims about learning outcomes.

### *Institutional Support and Faculty Development*

Finally, the review suggests that effective implementation of authentic assessment requires institutional support. Faculty may benefit from shared rubrics, calibration exercises, and professional development opportunities focused on aligning tasks with evaluation criteria. Communities of practice or teaching workshops could provide spaces for instructors to share examples, discuss challenges, and reflect on how authenticity is being operationalized across courses.

At the institutional level, support structures should emphasize experimentation and reflection rather than uniform adoption. Given the current state of the evidence, encouraging careful, evaluative use of authentic assessment is likely to be more productive than promoting it as a one-size-fits-all solution.

## **Conclusion**

Taken together, the evidence supports a measured and reflective approach to authentic assessment at DKU. Authentic assessment can enhance engagement and signal the value of applied learning when it is well aligned with course goals and evaluation criteria. However, its effectiveness is contingent on implementation, context, and expectations. Treating authentic assessment as one tool among many, rather than as a one size fits all solution, is most consistent with the current evidence.

## References

- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition*. New York: Longman.
- Andrade HL (2010) Students as the definitive source of formative assessment: academic self-assessment and the self-regulation of learning. In: Andrade HL, Cizek GJ (eds) *Handbook of Formative Assessment*. New York: Routledge, 90–105.
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking Transfer: A Simple Proposal With Multiple Implications. *Review of Research in Education*, 24(1), 61–100. <https://doi.org/10.3102/0091732X024001061>
- Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.
- Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-Based Learning in Higher Education: A Meta-Analysis. *Review of Educational Research*, 90(4), 499–541. <https://doi.org/10.3102/0034654320933544>
- Darling-Hammond, L. (with Adamson, F.). (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. Jossey-Bass.
- Evans, C., & Zhu, X. (2023). The development and validation of the assessment engagement scale. *Frontiers in Psychology*, 14, 1136878. <https://doi.org/10.3389/fpsyg.2023.1136878>
- Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67–86. <https://doi.org/10.1007/BF02504676>
- Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2008). Authenticity is in the eye of the beholder: Student and teacher perceptions of assessment authenticity. *Journal of Vocational Education & Training*, 60(4), 401–412. <https://doi.org/10.1080/13636820802591830>
- Herrington, A., Herrington, J., & IGI Global (Eds.). (2006). *Authentic learning environments in higher education*. IGI Global (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA). <https://doi.org/10.4018/978-1-59140-594-8>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kearney, S. (2013). Improving engagement: The use of 'Authentic self-and peer-assessment for learning' to enhance the student learning experience. *Assessment & Evaluation in Higher Education*, 38(7), 875–891. <https://doi.org/10.1080/02602938.2012.751963>

- Li, Y. R., Zhang, Z. H., Li, W., Wang, P., Li, S. W., Su, D., & Zhang, T. (2023). Effectiveness and learning experience from undergraduate nursing students in surgical nursing skills course: A quasi- experimental study about blended learning. *BMC Nursing*, 22(1), 396. <https://doi.org/10.1186/s12912-023-01537-w>
- Lombardi, M. (2008). *Making the Grade: The Role of Assessment in Authentic Learning*. <https://www.semanticscholar.org/paper/Making-the-Grade%3A-The-Role-of-Assessment-in-Lombardi/9ecdbdff06b4024145368ff53a4b6bbc1bfba6c9>
- Martínez-Argüelles, M.-J., Plana-Erta, D., & Fitó-Bertran, À. (2023). Impact of using authentic online learning environments on students' perceived employability. *Educational Technology Research and Development*, 71(2), 605–627. <https://doi.org/10.1007/s11423-022-10171-3>
- McArthur, J. (2023). Rethinking authentic assessment: Work, well-being, and society. *Higher Education*, 85(1), 85–101. <https://doi.org/10.1007/s10734-022-00822-y>
- Mueller, J. 2005. "The Authentic Assessment Toolbox: Enhancing Student Learning Through Online Faculty Development." *Journal of Online Learning and Teaching*.
- Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250–256. <https://doi.org/10.1037/0022-3514.35.4.250>
- Papageorgiou, E., Wong, J., Liu, Q., Khalil, M., & Cabo, A. J. (2025). A Systematic Review on Student Engagement in Undergraduate Mathematics: Conceptualization, Measurement, and Learning Outcomes. *Educational Psychology Review*, 37(3), 66. <https://doi.org/10.1007/s10648-025-10046-y>
- Perkins, D. N., & Salomon, G. (1992). Transfer of Learning. In T. Husén, & T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (2nd ed., pp. 425-441). Oxford: Pergamon.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Rhodes, T., & Finley, A. (2014). Using the VALUE rubrics for improvement of learning and authentic assessment. *Peer Review*, 3, 32
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159–179. <https://doi.org/10.1080/02602930801956059>
- Segers, M., Gijbels, D., & Thurlings, M. (2008). The relationship between students' perceptions of portfolio assessment practice and their approaches to learning. *Educational Studies*, 34(1), 35–44. <https://doi.org/10.1080/03055690701785269>

Shepard, L. A. (2000). The Role of Assessment in a Learning Culture. *Educational Researcher*, 29(7), 4–14. <https://doi.org/10.3102/0013189X029007004>

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning.<sup>1</sup>. *Assessment in Education: Principles, Policy & Practice*, 14(3), 281–294. <https://doi.org/10.1080/09695940701591867>

Vu, T. T., & Dall’Alba, G. (2007). Students’ experience of peer assessment in a professional course. *Assessment & Evaluation in Higher Education*, 32(5), 541–556. <https://doi.org/10.1080/02602930601116896>

Wiggins, G. (1990). *The Case for Authentic Assessment*. <https://doi.org/10.7275/FFB1-MM19>