RedFame

Journal of Education and Training Studies
Vol. 15, No. 1; January 2027
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Do Final Pitches Measure Fieldwork Grounding? A Multi-Source Validity Audit in a Sustainability Design-Thinking Bootcamp

Pattarada Rungruang[1], Athawit Buahong[2], Meechai Orsuwan[3], Huayi Tseng[4], Busaya Vongchavalitkul[1]

[1]Faculty of Accounting and Management Science, Southeast Bangkok University, Bangkok, Thailand

[2]Language Institute, Khon Kaen University, Khon Kaen, Thailand

[3]Faculty of Liberal Arts, Southeast Bangkok University, Bangkok, Thailand

[4]Global SDGs Alliance, Singapore

Correspondence: Meechai Orsuwan, Faculty of Liberal Arts, Southeast Bangkok University, Bangkok, Thailand.

Received: September 20, 2026 Accepted: October 30, 2026 Online Published: November 11, 2026

doi:10.11114/jets.v15i1.9490 URL: https://doi.org/10.11114/jets.v15i1.9490

Abstract

This study examined the validity of pitch-evaluation scores as indicators of fieldwork grounding in a design-thinking bootcamp. Using a qualitative, multi-source validity audit, it analyzed six international student teams participating in a four-day sustainability bootcamp held in mid-2026 in Khung Bang Kachao, Thailand. Data included team worksheets, final pitch decks, rubric scores from four evaluators, facilitator observation notes, community representative feedback, and post-forum student surveys. A two-sided coding framework assessed artifact-based evidence of insight quality, solution alignment, claim-level traceability to fieldwork, and AI-production markers, alongside rubric-derived evaluation signals. The artifacts were independently double-coded, and the two datasets were integrated in a convergence matrix. Findings revealed a marked disconnect between evaluation outcomes and documented fieldwork: a team whose pitch contained no field-traceable claims received grounding-related scores at least as high as those of all field-grounded teams. At the cohort level, the two most field-grounded teams ranked third and last on composite scores, whereas the two most detached teams ranked first and second. The findings challenge the use of final-pitch rubrics alone as valid measures of fieldwork-based design reasoning. The study recommends incorporating process artifacts, evidence-traceability checks, pre-pitch synthesis materials, formal community participation, and clearer simulation framing into assessment design.

Keywords: assessment validity, experiential learning, presentation-based scoring, rubric evaluation, simulation-based learning

1. Introduction

Design thinking has become a standard pedagogy in sustainability education. Its promise is specific: students enter a field setting, engage with community members, synthesize what they learn, and develop solutions responsive to locally grounded problems. Bootcamps, studios, and challenge-based programs around the world now use variations of this process (Kienzler & Fontanesi, 2017; Phi & Waldesten, 2021; Szymanska et al., 2020). Typically, these programs culminate in a short public pitch evaluated by a panel. The pitch often functions as the program's highlight and, in many cases, as its principal assessment point. This study examined a central but largely untested assumption of this format: that a highly rated final pitch provides valid evidence that a team's proposed solution is grounded in fieldwork. More specifically, it investigated whether evaluator scores for insight grounding and root-cause alignment corresponded to the documented fieldwork records underlying each team's pitch. The study used a qualitative, multi-source validity audit of six international student teams participating in a four-day sustainability bootcamp in Bang Kachao, Thailand.

1.1 Fieldwork Grounding and the Final Pitch

Fieldwork is not merely a preparatory activity in sustainability-oriented design thinking. It is intended to be the means through which students develop and demonstrate key competencies, including interpersonal competence, systems thinking, and integrated problem solving (Brundiers et al., 2021; Wiek et al., 2011). In the Double Diamond process used in the present bootcamp, students move from *Discover*, involving community fieldwork, to *Define*, involving synthesis into a point-of-view statement and a "*How Might We*" question, and then to *Develop* and *Deliver*, involving a solution concept and public pitch (Design Council, 2019). The pedagogical logic is that the final solution should remain traceable

to insights generated during these earlier stages. However, most programs assess this process through its final performance: the pitch. A polished presentation may communicate clarity, confidence, coherence, visual quality, and persuasive storytelling. Yet these qualities are not necessarily equivalent to evidence that a solution emerged from, or remained aligned with, field-based inquiry. This raises a measurement question: do scores assigned to grounding-related pitch criteria support the interpretation that students have demonstrated fieldwork-based design reasoning? This issue is especially consequential because pitch ratings, panel scores, and expert assessments of final presentations are commonly used as indicators of learning outcomes in design-thinking education research (Guaman-Quintanilla et al., 2022, 2023; Ingram et al., 2022; Yu et al., 2024). If such ratings do not reflect the process they purport to assess, both educational judgments and research conclusions may be based on a construct other than fieldwork grounding.

1.2 Significance of the Problem

The problem is worth investigating for practical and theoretical reasons. Practically, assessment communicates to students which forms of work matter. When a final pitch is treated as the principal evidence of learning, students may reasonably infer that persuasive presentation is more valuable than careful inquiry, evidence synthesis, or accountable engagement with community perspectives. This risk is particularly salient in sustainability education, where programs often claim to develop the capacity to understand complex real-world problems before proposing interventions. The problem also matters for the validity of educational research. Reviews of sustainability-competence assessment have found that available instruments remain ambiguous, insufficiently validated, and frequently dependent on student self-reports (Annelin & Boström, 2023). Researchers studying challenge-based sustainability education have similarly described student learning processes as a "black box" that conventional evaluation methods rarely open (Holmén et al., 2021). The field therefore has substantial theoretical agreement about the competencies students should develop, but less evidence about whether common assessment practices capture how, or whether, those competencies are demonstrated. The concern may be amplified by generative AI. Experimental evidence indicates that generative AI can increase evaluators' ratings of creative artifacts, with the largest gains occurring among less creative producers (Doshi & Hauser, 2024). Although this evidence derives from short fiction rather than student pitches, it raises a relevant possibility: if evaluators respond strongly to surface features of a presentation, and those features become easier to produce, presentation polish may become an increasingly weak proxy for the quality of the underlying process. The present study examined whether production-level patterns in a real student cohort were consistent with this broader concern. The purpose of this study was therefore to assess whether grounding-related pitch scores corresponded to the documented fieldwork evidence behind the pitches. By examining the relationship between score-based judgments and teams' process artifacts, the study evaluated whether the final-pitch format provided valid evidence for claims about fieldwork-grounded design reasoning.

1.3 Pitch Assessment, Authentic Assessment, and Validity

The final pitch has migrated from entrepreneurship education into sustainability and design-thinking programs, often with limited examination of what pitch performance actually measures. Research on business-plan competitions suggests that performance in these formats may not be equivalent to the intended learning outcomes. Watson et al. (2018), for example, found that competencies developed by nascent entrepreneurs were situated within the competition context and identified a gap between intended outcomes and actual learning. Similarly, Overwien et al. (2024) found no advantage of assessment-based competition formats over non-assessed formats for entrepreneurial intention, suggesting that assessment pressure may shape how participants engage. Evidence from investor decision-making raises a related concern. Clark (2008) found that business angels' screening decisions were strongly influenced by presentational factors, although evaluators explained their decisions primarily in terms of substantive criteria. This discrepancy is relevant to educational pitch assessment: evaluators may sincerely believe they are judging the quality of evidence or problem-solution alignment while responding substantially to features of presentation performance.

The issue also relates to debates about authentic assessment. Authentic assessment seeks to engage students in tasks resembling real-world professional practice and is often discussed in terms of realism, cognitive challenge, and evaluative judgment (Villarroel et al., 2018). However, recent critiques caution that authentic assessment is frequently justified by the resemblance of a task to professional practice rather than by evidence that the task produces valid inferences about learning (Ajjawi et al., 2024; Fawns et al., 2025). A field-based sustainability pitch is authentic in the sense that it resembles a public proposal or stakeholder presentation. Whether it provides valid evidence of the fieldwork and reasoning that supposedly produced it is a separate empirical question. Rubrics are commonly used to score these performances. Analytic rubrics and rater training can improve scoring consistency, but they do not independently establish the validity of score interpretations (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). Studies have further shown that well-developed rubrics do not necessarily guarantee reliable judgments and that raters and criteria remain potential sources of error in project-based assessment (Menéndez-Varela & Gregori-Giralt, 2018; Stellmack et al., 2009). Accordingly, a high score on a grounding-related criterion cannot itself establish that a team's work was grounded in field evidence. The present study adopted an argument-based approach to validation (Kane, 2013; Messick, 1989). From this

perspective, validation concerns the chain of inferences that connects an observed performance to a score-based claim. The present study examined the claim that high insight-grounding or root-cause scores indicated field-evidenced solutions. If a pitch lacks observable evidence of grounding, refining rubrics or calibrating raters cannot fix the validity issue, as the flaw lies in the assessment's underlying evidentiary basis.

1.4 Research Questions

This study used a qualitative, multi-source validity audit to compare evaluation signals with independently documented evidence of fieldwork grounding. The setting was a four-day sustainability bootcamp in Bang Kachao, Thailand, in which six international student teams progressed from desk research and community fieldwork to a final staged presentation. The final pitches were evaluated by four assessors using a rubric whose first two criteria were intended to measure insight grounding and root-cause alignment. The primary research question was whether evaluation scores converged with the fieldwork evidence documented in each team's artifacts. Three research questions guided the analysis:

RQ1: How grounded in field evidence was each team's pitched solution, as audited from the artifact record?

RQ2: Did the evaluation signal converge with the audited record, or did it diverge from it?

RQ3: Where the two diverged, what did the evaluation signal appear to track instead, and what features of the format made the divergence possible?

To address these questions, the study combined six data sources for each team: group worksheets, final pitch decks, rubric scores from four evaluators, facilitator observation notes, community representative feedback, and a post-forum student survey. A two-sided codebook coded the artifacts and pitches on one side and the rubric scores on the other. The artifacts were independently double-coded, and the two datasets were then compared in a convergence matrix. This design permits a direct test of the score interpretation rather than an indirect estimate based on evaluator agreement or mean-score reliability. It examines whether the claims presented and rewarded in final pitches can be traced to teams' recorded fieldwork and synthesis processes. The findings therefore contribute both a validity audit of one assessment format and practical recommendations for making fieldwork grounding more visible in sustainability design education.

2. Method

2.1 Research Design

This study used a qualitative, multi-source validity audit guided by the argument-based approach to validation (Kane, 2013). In this approach, a score interpretation is considered valid only to the extent that supporting evidence justifies the inferences drawn from that score. The focal score interpretation was that a high evaluator score for insight grounding or root-cause alignment indicated that a team's proposed solution was grounded in field evidence. The study used a single-case embedded design. The four-day sustainability bootcamp constituted the case, and the six student teams constituted embedded units of analysis. Rather than testing population-level associations or making statistical inferences, the study conducted a transparent, team-level audit of convergence between two forms of evidence: (a) evaluator scores assigned during final pitches and (b) independently coded documentary evidence from each team's fieldwork and synthesis process. Its analytical strength lay in the depth and triangulation of evidence available for each team across the full design process, from fieldwork preparation through final presentation.

2.2 Setting and Participants

The study was conducted during a four-day sustainability bootcamp held in mid-2026 in Khung Bang Kachao, Thailand. Bang Kachao is a protected green area located within a bend of the Chao Phraya River near Bangkok. The bootcamp involved 30 international undergraduate and graduate students, who were organized into six teams of approximately five students each. Teams are anonymized in this article as Teams A–F. Students worked across three sustainability themes: circular economy, climate change, and biodiversity. All teams followed the Double Diamond design process (Design Council, 2019). They conducted desk research, planned fieldwork, interviewed community stakeholders in Bang Nam Phueng, synthesized findings into a point-of-view statement and a "*How Might We*" question, and developed a solution concept.

The bootcamp concluded with a simulated international sustainability forum held at a Bangkok conference venue. Teams delivered role-played final pitches, and community representatives from Khung Bang Kachao attended as audience members and provided written feedback. Four evaluators scored each team's pitch using the study rubric. Eligibility for the student sample was determined by enrollment in the bootcamp. All 30 enrolled students participated in the team-based learning activities and final presentations. The post-forum survey was completed by 15 students. The study did not seek to estimate effects for a broader population; therefore, conclusions were limited to this case and should be interpreted as an instrument-validity audit rather than as generalizable estimates of student learning.

2.3 Sampling Procedures

The study used a purposive, bounded-case sampling strategy. The bootcamp was selected because it provided a complete documentary trail for each team, including materials created before, during, and after fieldwork, alongside evaluator ratings of final presentations. This combination enabled direct comparison between the evidence underlying each pitch and the scores assigned to it. All six teams in the bootcamp were included; no teams were excluded from the artifact audit. All four evaluators who scored the final presentations were included in the analysis of rubric data. Community feedback was available from two respondents, and post-forum survey responses were available from 15 students. Facilitator observation notes varied in completeness: notes were rich for two teams, brief for two teams, and unavailable for two teams. Students were not paid to participate in the bootcamp. The final forum involved no rankings, prizes, awards, or grade consequences for students. The rubric was used exclusively as a research instrument in this setting. This low-stakes condition was important because it enabled examination of the rubric without direct consequences for the students being assessed.

2.3.1 Sample Size, Power, and Precision

The analytic sample comprised six teams, 30 student participants, four evaluators, two community feedback respondents, and 15 student survey respondents. Sample size was determined by the full cohort of the bounded bootcamp case; a power analysis was unnecessary because the study did not test population parameters or intervention effects. The small number of teams also precluded meaningful inferential statistical testing. Instead, the study emphasized within-case precision through claim-level evidence tracing, independent double-coding, multiple data sources, and transparent reporting of convergence and divergence across sources. The findings should not be interpreted as estimates of the prevalence of assessment invalidity across sustainability education programs. Rather, they show whether the intended interpretation of this specific rubric's grounding-related scores was supported by the available evidence in this case.

2.4 Measures and Data Sources

Six data sources were collected across the design lifecycle. Table 1 presents these sources, their sample sizes, and analytic functions.

Table 1. Data sources, available samples, and primary analytic purposes

Data source	Available sample	Primary analytic purpose
Group worksheets	6 teams	Insight Quality, Solution Alignment, and fieldwork trace targets
Final pitch decks	6 teams	Solution Alignment, Evidence Traceability, and AI-production markers
Evaluator rubric scores	4 evaluators × 6 teams	Evaluation signal for grounding-related criteria
Facilitator observation notes	4 teams with variable coverage	Qualitative context after primary coding
Community feedback forms	2 respondents	Community-insider contrast
Post-forum student survey	15 students	Student self-assessment and open-text reflections

To measure artifact-based groundedness, a two-sided codebook was developed for the audit. Side 1 assessed fieldwork groundedness through four artifact-based codes. *Insight Quality* (IQ) was coded on a three-level scale (Surface, Identified Issue, and Systemic) based on the self-assessment levels in the student field guide, and a level was assigned only when documentary evidence supported it. *Solution Alignment* (SA) measured the internal coherence between the pitched solution and the team's documented root cause on a three-level scale (Misaligned, Partial, or High); it was not treated as evidence of field grounding or solution quality. *Evidence Traceability* (ET), the primary measure of fieldwork grounding, extracted each substantive claim in the final pitch deck and classified it as Traceable (linked to specific fieldwork records), *Imported* (derived from desk research or outside models), or *Unsourced* (unsupported by any artifact). Claims that contradicted the team's own field data received an additional flag, and the proportion of traceable claims was used to categorize each team as field-grounded, hybrid, or detached. Finally, *AI-Production Markers* (AIPM) categorized visible production-level indicators in pitch decks as *Definitive*, *Probable*, or *Indeterminate*. This code described artifact production only and was not used to infer reasoning processes, authorship, or actual AI use.

Side 2 drew the evaluation signal from the final-pitch rubric, on which four evaluators scored all six teams on a 1-to-4 scale across seven criteria (insight grounding, root-cause alignment, sustainability impact, business viability, feasibility and scalability, creativity and innovation, and pitch communication). The primary outcomes focused on *Criterion 1* (Insight Grounding) and *Criterion 2* (Root-Cause Alignment) because their labels explicitly represented the constructs examined through independent artifact coding. For each team and criterion, mean scores, score ranges, and team rankings across *Criterion 1*, *Criterion 2*, and total rubric score were calculated. A non-discriminating scoring instance was defined as an evaluator range of two or more points for the same team and criterion. Finally, the analysis examined whether Criteria 1 and 2 aligned more closely with *Criterion 7* (Pitch Communication), the evaluators' own measure of presentation quality, than with audited fieldwork groundedness.

2.5 Data Analysis

The analysis compared two independently derived evidence streams. The first stream consisted of artifact-based codes representing documented fieldwork groundedness. The second consisted of evaluator rubric scores representing the assessment signal produced by the final-pitch format. The coding process followed a fixed sequence to reduce contamination between the two streams. First, Side 1 was coded solely from team worksheets and final pitch decks. During this stage, coders did not consult evaluator scores, facilitator notes, community feedback, or student surveys. Second, Side 2 was computed after Side 1 codes were finalized. The two streams were then joined in a convergence matrix. Each team received one of three convergence verdicts: *Convergent*, if the evaluation signal positioned the team approximately where the artifact audit positioned it; *Inverted*, if the evaluation signal materially exceeded the team's audited groundedness; or *Discounted*, if the team's audited groundedness materially exceeded its evaluation signal. Facilitator notes, community feedback, and survey responses were used after the primary verdicts had been assigned. These sources provided explanatory context and a comparison of assessor, student, and community interpretations; they did not determine the primary artifact-based groundedness codes.

2.6 Experimental Manipulations or Interventions

No experimental manipulation, intervention, control condition, or random assignment was used. All teams participated in the same four-day bootcamp process and delivered a final pitch in the same simulated forum format. The low-stakes deployment of the rubric was a contextual feature rather than an experimental manipulation. Students received no rankings, prizes, awards, or academic consequences based on their rubric scores. This condition reduced the likelihood that direct assessment consequences would shape student behavior during the closing event and allowed the study to focus on the validity of score interpretations.

2.7 Coding Procedures and Intercoder Agreement

The primary coder applied the artifact-based codebook to all six teams. A second coder independently coded *Insight Quality*, *Solution Alignment*, and *Evidence Traceability* for all teams. The second coder received the codebook, six team worksheets, and six final pitch decks, but did not receive evaluator scores, facilitator notes, survey responses, or community feedback. Coding proceeded team by team in a fixed sequence (Insight Quality, then Solution Alignment, then claim extraction and classification for Evidence Traceability), and each code was recorded with a one-line justification quoting or pointing to the relevant artifact. Artifact coding was completed and locked before the primary coder computed any evaluator-score summaries, so that score information could not inform the process codes. For *Evidence Traceability*, coders independently extracted claim lists from each pitch deck before reconciling the list of substantive claims to be classified. A countable claim was defined as a declarative statement about the problem, its causes, or the solution's expected effect that could in principle be sourced; slogans, team or venue framing, agenda slides, and closing slides were excluded. Claims were then categorized as traceable, imported, or unsourced according to the codebook. Decision rules governed borderline cases: an Insight Quality judgment falling between two levels was assigned the lower level and flagged as borderline; a solution rationale that followed logically from a traceable insight was coded as traceable (derived); numerical claims with no source in any team artifact were coded as unsourced even when plausible; and claims that contradicted the team's own field data were coded by source and additionally flagged. Intercoder agreement was established through independent double-coding followed by code-by-code reconciliation, rather than through a summary reliability coefficient. Because the team-level sample was limited to six teams, chance-corrected reliability coefficients were not calculated at the team level. After reconciliation, the convergence verdicts (convergent, inverted, or discounted) were assigned jointly by both coders. The codebook deliberately contained no researcher-rated polish code, because such a scale would have been subjective and unanchored; perceived polish was instead represented by the evaluators' own Pitch Communication criterion (C7) in the halo analysis. All disagreements were resolved through discussion before the convergence matrix was finalized. The first-round coding sheets were not retained in a form that allows agreement statistics to be computed retrospectively; agreement is therefore described procedurally rather than numerically. Facilitator notes were deliberately excluded from primary coding because coverage varied substantially across teams and because facilitators were participant-observers rather than independent assessors. One uniformly positive facilitator report that lacked specific incidents was treated as low-weight contextual evidence.

2.8 Ethics, Anonymization, and Reflexivity

The study followed the ethics protocol approved for the bootcamp research program (Southeast Bangkok University, Protocol No. 035/2568, approved May 19, 2026).The evaluator rubric was included in the approved instrument set, and participant consent covered secondary analysis of evaluator scores and other research materials produced during the bootcamp. All participants were anonymized in reporting. Student teams are identified as Teams A–F, evaluators as Evaluators 1–4, and facilitators by team designation. Because the cohort was small, the event date and venue are reported in coarsened form. The analysis was directed toward the assessment instrument and format, rather than toward judging

the performance of individual students or teams. Community respondents were cited according to their consent preferences: one permitted anonymous quotation, and one permitted attribution by role only. Several dual roles were considered in interpretation. First, facilitators were members of the teams they observed; their notes were therefore treated as participant accounts rather than independent observations. Second, one community feedback respondent had also served as a fieldwork informant for one team. Third, the lead researcher designed the bootcamp, student field guide, and rubric under audit. This insider position provided detailed familiarity with the instructional and assessment context but also created a risk of confirmation bias. To address this risk, the research team separated primary artifact coding from access to evaluation scores and contextual accounts, used an independent second coder, retained a transparent audit trail, and distinguished documentary evidence from researcher interpretation throughout the analysis.

3. Results

3.1 Overview: The Convergence Matrix

Table 2 presents the full convergence matrix. It joins the artifact audit (Side 1) with the evaluation signal (Side 2) for all six teams. Three summary observations frame the sections that follow. The first and most important observation was that the central finding of the audit was categorical rather than correlational: the team whose pitch contained no field-traceable claims (Team B) received grounding-criterion scores at or above those of every field-grounded team. Second, audited groundedness did not predict the evaluation signal more broadly: the two field-grounded teams (C and D) ranked third and last on the composite score. Third, four of the six teams received a non-convergent verdict.

Table 2. Convergence matrix for the six teams

Team	IQ	SA	ET (T:I:U)	Grounding verdict	AIPM	C1 mean	C2 mean	Total (rank)	Verdict
B	Surface	High	0:5:2	Detached	Probable	3.25	3.25	21.25 (1)	Inverted
F	Identified to Systemic	Partial	1:5:3	Detached	Definitive	3.00	2.50	20.25 (2)	Inverted
C	Systemic	High	6:1:1	Field-grounded	Indeterminate	3.25	2.75	19.25 (3)	Convergent-high
E	Identified	Misaligned	0:2:4	Detached	Probable	2.75	3.00	18.75 (4)	Inverted
A	Identified	Partial	1:4:2	Detached	Probable	2.50	2.50	18.25 (5)	Convergent-low
D	Systemic	High	8:1:2	Field-grounded	Indeterminate	3.00	2.50	18.00 (6)	Discounted

Figure 1 shows the central pattern in raw values, using only the two criteria whose labels claimed to measure grounding. Each point represents one team. The horizontal axis shows the percentage of field-traceable claims in the pitch deck, and the vertical axis shows the mean of Criterion 1 (Insight Grounding) and Criterion 2 (Root-Cause Alignment) across the four evaluators. If these criteria tracked groundedness, the points would be expected to rise from the lower left to the upper right. However, the observed pattern was close to the opposite. Team B sat in the upper left, with 0 of 7 claims traceable and the highest grounding score, whereas Team D sat in the lower right, with 8 of 11 claims traceable and a grounding score near the lowest. No trend line was fitted, because the sample contained only six teams.

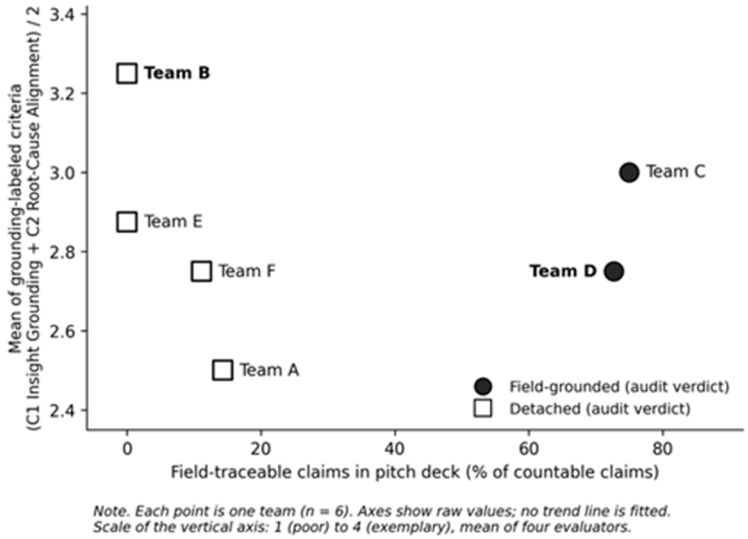


Figure 1. Field-traceable claims versus the mean of the grounding-labeled criteria

Description: Filled circles indicate teams audited as field-grounded; open squares indicate teams audited as detached. Axes show raw values; no trend line was fitted.

3.2 The Decisive Case: Scored Highest Without Grounding (Team B)

The argument of this paper rests mainly on one team, so this case is presented first. Team B produced the thinnest documented fieldwork in the cohort, yet it received the highest scores on the two criteria that claim to measure grounding. This combination was the decisive result of the audit, because it did not require any judgment about which team was the most grounded; it required only the negative demonstration that Team B was not grounded. First, the pitch itself contained no field-traceable claims (0 of 7). The deck presented a complete and internally coherent solution package, consisting of salt-tolerant crop varieties modeled on Bangladesh, a four-step rainwater irrigation scheme, and an inter-district coordination committee with key performance indicators. Every substantive claim was either imported or unsourced. The only traceable item in the team's records, an informant's account of a friend's death from heat stroke, did not appear in the pitch at all.

One plausible objection is that deck structure reflects rhetorical choices rather than gaps in knowledge: a team could know its field research well but simply choose to foreground its solution. If that were the case here, traces of field knowledge should appear in the artifacts that carried no time limit. They did not. The group worksheet, which was completed without any presentation constraints, recorded only single-line reflections, and its documented synthesis did not move beyond surface observation. Similarly, the post-forum survey responses of Team B members contained no field-specific content. Detachment thus appeared consistently across three artifacts of different genres, which is difficult to reconcile with the rhetorical-choice explanation. Even so, the package aligned cleanly with the team's stated focus: salinity affecting crops. This internal coherence was exactly what the rubric credited: Team B received the highest Criterion 1 and Criterion 2 mean in the cohort, together with the highest composite score. This case shows that solution alignment can be manufactured from imported material without any field grounding and that the evaluation instrument could not tell the difference. Figure 2 places this case in the cohort context. It shows the claim-level composition of each deck, ordered by composite score. As the figure indicates, the two highest-scoring decks contained almost no field-traceable claims, while the lowest-scoring deck consisted mostly of them.

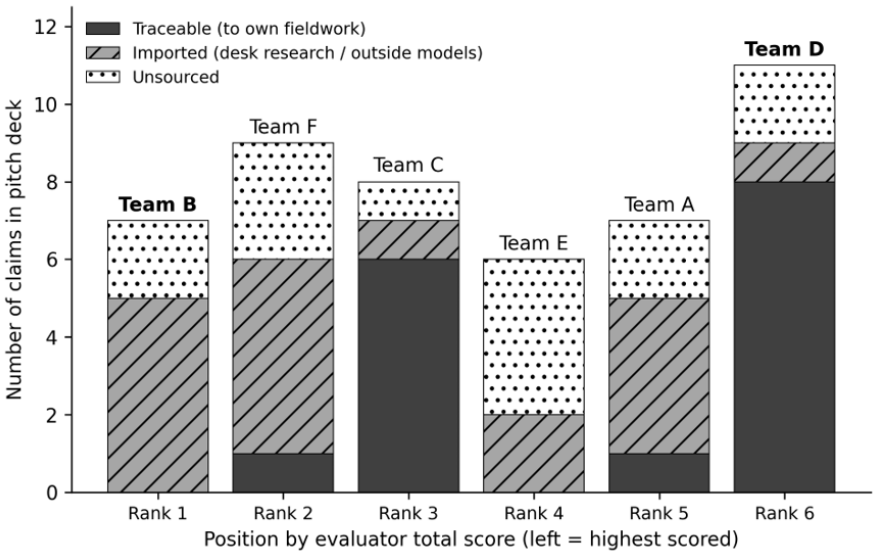


Figure 2. Evidence-traceability composition of each pitch deck

Description: Ordered from the highest-scoring team (left) to the lowest-scoring team (right). Dark segments indicate claims traceable to the team's own fieldwork; hatched segments indicate imported claims; dotted segments indicate unsourced claims.

3.3 Two Further Routes Away From the Field

Two other teams received evaluation signals that exceeded their audited groundedness, each by a different route. The research team labeled these two routes substitution and escalation. Together with Team B's borrowed coherence, these routes form three mechanisms by which a pitch can detach from the field without any visible cost. Substitution (Team E). Team E's field synthesis identified a power asymmetry, namely that the community suffered from problems generated by larger external actors. Its pitch, however, presented algae carbon-capture tanks, and the documented root cause never appeared in it. The facilitator's notes explained the route from the inside: the team's *How Might We* question could not mature into a business idea, so the team pivoted, and when it then received critical feedback, the team was unsure how to revise and left the plan unchanged. The evaluators gave Team E the highest creativity mean in the cohort (3.25). The

novelty of the substituted solution was credited, while its disconnection from the field was not registered at all.

Escalation (Team F). Team F's fieldwork produced genuine mid-level insights, including a contradiction that the team noticed itself (the community expresses environmental care yet uses plastic heavily) and a specific infrastructure grievance (water access closed for eight months). Its pitch was produced end to end with an identifiable AI tool; a NotebookLM watermark appeared on every slide. The pitch escalated these insights into a three-pillar governance framework (payments for ecosystem services, a community data mandate, and dynamic carrying capacity), which was presented as a simulated address to a UN biodiversity forum in 2030. The local insights survived only as background statistics, whereas the centerpiece figures of the framework (an 85 percent ecosystem stress gauge and a 60 percent capacity throttle) were unsourced. Team F ranked second by composite score.

3.4 The Grounded Contrast Cases (Teams C and D)

Two teams produced field-grounded pitches, and how the evaluation signal treated them completes the picture. Team C's pitch traced six of eight claims to the team's own fieldwork. Its worksheet documented a systemic synthesis, including an assumption that was corrected in the field. The team had expected tourists to be the main source of waste, but found that daily community activities contributed heavily as well. Its Criterion 1 mean (3.25) was the highest in the cohort. However, Team C shared this highest score with Team B, whose traceable claim count was zero. Team D's record was the densest in the cohort. Its worksheet connected vendor margin economics (a 6-baht profit on a 20-baht snack), packaging cost economics (eco-bags costing 1 baht each, against standard plastic bags priced at 90-100 baht per kilogram), and canal infrastructure (grease-contaminated orphaned plastic in open tidal canals) into one explanatory system, and two initial assumptions were explicitly corrected by field data. Eight of eleven pitch claims traced to named informants, including three verbatim quotes, and the solution engaged the documented root cause directly. Nevertheless, Team D received the lowest composite score in the cohort (18.0). At the same time, the audit recorded lapses inside this grounded profile: one unsourced statistic (single-use plastic as 40 percent of total waste) and one claim that contradicted the team's own field data (no sorting system for the waste, although the team's informants had described the sorting system in detail). Groundedness in this cohort was thus a matter of degree, and the audit registered lapses even in the strongest record. Team D's facilitator reported actively assisting with late-night presentation restructuring and slide organization. While the underlying evidence originated entirely from the students' prior fieldwork, the deck's final evidence visibility was co-produced with the facilitator. Because such intensive support was non-uniform across the cohort, Team D was treated as a suggestive contrast rather than a primary case. The study's central argument therefore rests on Team B, whose verdict remained independent of external facilitator effects.

3.5 The Criterion Could Not Discriminate

The central validity test of the audit was a direct one: did Criterion 1 (Insight Grounding) separate the grounded teams from the detached teams? It did not. Team B (0 of 7 claims traceable) and Team C (6 of 8 claims traceable) received identical Criterion 1 means of 3.25. Rater dispersion reinforces this conclusion. On a four-point scale, the four evaluators spanned a range of 3 points for Team A on Criterion 1 (scores of 1 to 4) and a range of 2 points on Criterion 2 for Team E (scores of 2 to 4). Under the threshold used in this study (a range of 2 or more), non-discriminating instances occurred for every team on at least one criterion. Disagreement of this size was arguably itself evidence, because it suggested that the criterion gave the raters nothing stable to score, which is exactly what the argument-based framework predicts when the construct was absent from the observed performance. Prior work on rubric-based assessment of student projects has likewise identified raters and criteria as substantial sources of score variance, even when the instrument itself was well constructed (Menéndez-Varela & Gregori-Giralt, 2018). The findings pointed to a distinct halo pattern. Team B coupled the cohort's most fluent pitch with the weakest field evidence, yet secured top marks for grounding. In contrast, Team D paired the most extensive field trail with a brief delivery and earned the lowest overall score. Because the data cover only six teams and four evaluators, we report this dynamic as a qualitative pattern rather than a statistical correlation.

3.6 Three Evaluation Signals, One Shared Blind Spot

The rubric was not the only signal that failed to register groundedness. The students' own post-forum self-assessments (n = 15, five teams represented) showed a ceiling effect on the item asking whether the team's presentation responded to the needs or pain points of the local field. Members of Team B and Team E, whose pitches contained no traceable field claims, rated themselves 4 or 5 out of 5, and their ratings could not be distinguished from those of the members of Teams C and D. Self-perception, like the external rating, therefore did not track the audited record. A different perspective came from the community audience. The research team treated community feedback (n = 2) as an illustrative counterpoint rather than a systematic comparison, given its aggregate nature and lack of team attribution. Community representatives observed that the bootcamp participants generally understood waste management but overlooked legal contexts and local community responsibilities, both of which appeared in the field data but not in the final pitches. Student post-forum reflections, however, confirmed that learning went beyond what the pitches conveyed; for instance, a Team D member

articulated economic insights in open-text responses that were heavily compressed in the team's pitch. Thus, the documented discrepancy reflected a gap between student learning and presentation visibility rather than a deficit in learning itself.

4. Discussion

4.1 What the Evaluation Signal Appeared to Capture

The audit began from a simple validity question in Kane's (2013) sense: did a high score on Insight Grounding support the claim that a team's work was grounded in field evidence? In this cohort, the answer was no. The criterion produced identical scores for the team with the richest evidence trail and for the team with none, and the total score ordered the cohort almost in reverse of the audited record. Following the argument-based approach, the failure occurred at the first inference in the chain: the scoring inference. The evaluators scored a timed presentation, but a presentation is not a record of fieldwork; it is a performance about fieldwork. What the scores did track can be seen in the profiles of the highest-scoring teams. Team B offered a complete and internally coherent solution narrative, Team F offered a polished and professionally produced policy framework, and Team E offered a novel technology. Coherence, production quality, and novelty are all legitimately impressive properties of a presentation. However, none of them required field grounding, and in this cohort all three were achieved without it. The results point not to evaluator error but to a construct that the format left unobservable: evaluators could only score the presentation evidence in front of them. As Team B illustrated by assembling alignment from external sources, internal coherence can easily simulate authentic grounding. This finding was consistent with evidence from investor settings. Clark (2008) reported that prospective investors were most influenced by presentational factors while citing substance-oriented reasons for their investment decisions. Likewise, Tsay (2021) demonstrated across twelve studies that observers who watched silent videos of pitches, with no access to the content at all, identified the winning entrepreneurs more accurately than those who read the transcripts. Both studies show that evaluators often mistake surface performance for substance. The bootcamp evaluators showed the same pattern, and the pitch format reinforced it by offering little other evidence to evaluate.

One rival explanation is that evaluators simply rewarded genuine solution quality on the non-grounding criteria (e.g., viability, impact). While this could explain overall score rankings (such as Team B's high viability marks for a proven, imported solution), it does not account for the grounding-specific criteria. A criterion assessing field grounding should not award top marks to a pitch with zero field-traceable claims. Isolating the grounding criteria confirmed that the inversion persisted independently of solution quality. Furthermore, evaluating solution quality presupposes an accurate problem diagnosis, which pitches cannot validate. A team can competently design a solution to a misframed problem without the flaw being detected in a six-minute presentation. Team E, for example, pitched a solution detached from its documented root cause, Team D included claims contradicted by its own informants, and community feedback highlighted critical missing field contexts (e.g., legal frameworks). Diagnostic accuracy is therefore a distinct construct, separate from narrative coherence and solution design, and standard pitches do not make it observable. A second rival explanation holds that the problem lies in a flawed rubric rather than in the observation point itself. An argument-based validation framework answers this: when a construct is absent from the observed performance, no amount of rubric refinement or rater calibration can yield valid scores for unobservable evidence. This is consistent with the broader rubric literature, which notes that well-designed instruments improve scoring consistency but cannot guarantee validity (Jonsson & Svingby, 2007; Stellmack et al., 2009). Revising the rubric to require inline claim sourcing would not undermine this thesis; it would confirm it, because such a revision changes what evaluators are able to observe.

4.2 Three Routes Away From the Field and the Forces Behind Them

The three inversion mechanisms (borrowed coherence, substitution, and escalation) were descriptively distinct, but they may have responded to the same underlying pull. Two features of the setting were candidates for that pull. The first feature was the format itself. A six-minute staged presentation favors a self-contained story. Prior work on business plan competitions reached a compatible conclusion: what students developed in competition formats was situated within the competition context, opening a gap between the intended and the actual learning outcomes (Watson et al., 2018). The present claim-level audit adds a mechanism to that literature: the competition artifact can detach from its evidentiary base without any visible cost, because the format never asks for the base. The second feature was the simulation frame. The closing forum was staged as an international UN-style event, and this frame plausibly signaled that the appropriate register was global, scalable, and policy-shaped. Team F's escalation into a 2030 governance framework and Team E's carbon-technology substitution both fit that register, whereas Team D's one-baht bag economics did not. As one participant observed, the setup effectively simulated presenting research before international judges at a high-level forum. Yet the more prestigious the stage appears, the stronger the pull away from small-scale, field-evidenced solutions. Generative AI markers were also definitive in one inverted team's deck and probable in the other two, but absent from both grounded teams' decks. Because chat logs were not collected, we cannot establish a causal link to team reasoning. The pattern is,

however, consistent with Doshi and Hauser's (2024) finding that generative AI can raise evaluators' ratings, with the largest gains for weaker producers. Where presentation polish already masks a lack of field grounding, tools that render polish effortless will only widen the divide.

4.3 Rethinking How Bootcamp Outcomes Are Evaluated

The most consequential audience for these findings may be researchers rather than program designers. A large share of the empirical literature on design-thinking and sustainability bootcamps measures outcomes at the pitch: rubric scores, jury rankings, and expert ratings of final presentations (Guaman-Quintanilla et al., 2022; Yu et al., 2024). If the present cohort is at all typical, such measures capture presentation performance, and presentation performance can invert the underlying process record. Studies that report high-quality insights or well-aligned solutions on the basis of pitch-stage ratings may be reporting the wrong construct. The remedy is not a better rubric at the same observation point. The remedy is a different observation point: the process artifacts that teams produce anyway, such as worksheets, synthesis documents, and evidence tables. This study showed that these artifacts can be audited at the level of individual claims. Reviews of hackathon research reached the same diagnosis from the other direction: most studies relied on perceived learning, which cannot substitute for evidence of the process itself (Medina Angarita & Nolte, 2020; Schulten & Chounta, 2024).

4.4 The Insiders Saw It

Two signals in this study did register the gap, and both came from proximity to the process rather than from expertise in evaluation. The facilitators embedded in Teams D and E placed their teams accurately, and one facilitator documented her own team's pivot away from its field data. The community representatives, with only one open-ended question, named the exact categories of field material that had not survived into the pitches. The pattern was consistent with a simple interpretation: in this cohort, visibility of groundedness depended on access to the process, and neither evaluative expertise nor participant sincerity could substitute for that access. The most authentic audience, the community itself, could name what was missing, but the format gave that judgment no channel into the formal record; the rubric simply had no cell for it.

4.5 What This Means for the Fieldwork-to-Design Promise

Field-based design thinking rests on a promise: participants go to the field so that what they design responds to what is actually there. The cohort in this study showed that the promise was realizable, since Teams C and D fulfilled it. At the same time, the cohort showed that the standard end-of-program checkpoint cannot tell whether it was fulfilled. Yet visibility carries pedagogical weight: when a program's terminal assessment credits superficial coherence, it implicitly teaches students that field research is optional.

4.6 Final Considerations

In a design-thinking bootcamp built explicitly around community fieldwork, the final evaluation could not see the fieldwork. Scores on the criteria named for grounding and alignment ordered the six teams almost in reverse of their documented evidence trails: the pitch with no field-traceable claims scored highest, while the pitch with the densest trail scored lowest overall. The failure was structural rather than personal. A six-minute staged performance simply did not contain the evidence that the criteria asked evaluators to score. Internal coherence can be built from borrowed models, substituted technologies, or escalated abstractions, and at the observation point that the format provided, internal coherence cannot be distinguished from grounding. None of this argues against the pitch as pedagogy, and none of it suggests that the learning was absent; the process artifacts and the student reflections showed clearly that it was there. The conclusion is narrower and more actionable than that. The pitch cannot fully serve as the measurement instrument for field groundedness, either for programs or for research. Valid alternatives exist in the process artifacts students generate and the community insiders who observe them, both of which are usually set aside when assessment begins. If sustainability education expects students to privilege evidence over performance, its evaluation methods must do the same.

Based on the audit, this study proposes five practical design recommendations. First, programs should score the trail, incorporating existing process artifacts (e.g., group worksheets) into formal evaluations. Second, an evidence-traceability step requiring teams to map core claims to field sources would make detachment visible and enforce empirical rigor. Third, providing field artifacts to evaluators in advance (for example, a one-page synthesis of pain points and root causes) would re-anchor the criteria to observable data. Fourth, formalizing insider evaluation by giving community representatives a scored vote on solution relevance would bring the most valid diagnostic signal into the formal record. Fifth, educators should calibrate the simulation frame to counter prestige bias, explicitly instructing judges to favor locally scaled, well-evidenced solutions or pairing pitches with community-sited presentations. These recommendations retain the pitch for synthesis and communication while ensuring it no longer serves as the sole instrument for measuring field grounding.

This study had four primary limitations. First, examining a single six-team cohort yields an illustrative case record rather than a generalizable distribution, and the facilitator's involvement in restructuring Team D's deck blends team and

facilitator output, although the core conclusions do not rely on this team. Second, data boundaries (the absence of recorded Q&A sessions and AI chat logs) risk undercounting spoken evidence and restrict AI-related findings to observed artifacts rather than causal mechanisms (Doshi & Hauser, 2024), a gap currently being addressed in an ethics-reviewed follow-up study. In addition, because the first-round coding sheets were not retained, intercoder agreement could be described only procedurally, and no per-code agreement statistics are reported. Third, the evaluation signal itself carries potential evaluator bias. The four evaluators scored live pitches within the program rather than as blind external benchmarks, so their ratings may have been shaped by presentation order, delivery style, impressions formed during the event, and the halo pattern documented in Section 3.5. Evaluator characteristics such as disciplinary background, familiarity with the field site, and prior rubric experience were not analyzed, and with four raters and six teams the study could not separate rater-specific leniency or severity from criterion-level non-discrimination. Researcher bias is a related concern, because the lead researcher designed the rubric under audit; this risk was mitigated, though not eliminated, through artifact-only coding, independent double-coding, fixed coding protocols, and locking artifact codes before scores were examined. Fourth, the findings are bound to contextual factors specific to this cohort: a four-day timeline, a six-minute presentation limit, a simulated international forum frame that may have favored global, policy-shaped solutions (Section 4.2), uneven facilitator support across teams, uneven use of generative AI tools, and community feedback limited to two respondents. Different time allocations, audiences, or presentation formats might narrow or widen the gap between pitch scores and fieldwork evidence. Future research can apply this portable claim-level audit across diverse program formats (e.g., competitive vs. non-competitive) and across multiple cohorts or institutions to test whether the pattern generalizes. A multi-site validation study would be the natural next step. Applying the same codebook to several bootcamps, run by different institutions and evaluated by different panels, would yield enough teams and claims to estimate chance-corrected intercoder reliability, to test the association between grounding-labeled scores and audited traceability quantitatively, and to model rater effects directly. Such a design could also vary the features this study could only describe, such as presentation length, the simulation frame, advance access to field artifacts, and formal community scoring, to establish which format changes restore the link between pitch scores and fieldwork evidence.

Acknowledgments

The authors wish to thank UN ESCAP, Southeast Bangkok University, and the Global SDGs Alliance for facilitating global collaboration on this project. We also express our sincere appreciation to the community of Bang Nam Phueng for their essential support and participation.

Author contributions

Pattarada Rungruang developed the study conceptualization, established the methodology, managed project administration, and contributed to the critical review and editing of the manuscript. Athawit Buahong wrote the original draft and led the investigation and formal data analysis. Huayi Tseng facilitated the focus groups as part of the investigation and participated in the formal analysis. Busaya Vongchavalitkul contributed to the methodology, conducted formal data analysis and interpretation, and wrote the original draft as well as reviewed and edited the Findings and Discussion sections. Meechai Orsuwan provided overall supervision, edited the final manuscript and served as the corresponding author. All authors read and approved the final manuscript.

Funding

Not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Redfame Publishing.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

References

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education, 49*(4), 499-510. https://doi.org/10.1080/02602938.2023.2271193

Annelin, A., & Boström, G. (2023). An assessment of key sustainability competencies: A review of scales and propositions for validation. *International Journal of Sustainability in Higher Education, 24*(9), 53-69. https://doi.org/10.1108/IJSHE-05-2022-0166

Brundiers, K., Barth, M., Cebrián, G., Cohen, M., Diaz, L., Doucette-Remington, S., … Zint, M. (2021). Key competencies in sustainability in higher education: Toward an agreed-upon reference framework. *Sustainability Science, 16*(1), 13-29. https://doi.org/10.1007/s11625-020-00838-2

Clark, C. (2008). The impact of entrepreneurs' oral 'pitch' presentation skills on business angels' initial screening investment decisions. *Venture Capital, 10*(3), 257-279. https://doi.org/10.1080/13691060802151945

Design Council. (2019). *Framework for innovation: Design Council's evolved Double Diamond.* Design Council.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Fawns, T., Bearman, M., Dawson, P., Nieminen, J. H., Ashford-Rowe, K., Willey, K., ... Press, N. (2025). Authentic assessment: From panacea to criticality. *Assessment & Evaluation in Higher Education, 50*(3), 396-408. https://doi.org/10.1080/02602938.2024.2404634

Guaman-Quintanilla, S., Chiluiza, K., Bravo Matamoros, A., Everaert, P., & Valcke, M. (2022). What is the state of the art regarding the application of design thinking in higher education? A scoping review. *Aula Abierta, 51*(4), 319-328. https://doi.org/10.17811/rifie.51.4.2022.319-328

Guaman-Quintanilla, S., Everaert, P., Chiluiza, K., & Valcke, M. (2023). Impact of design thinking in higher education: A multi-actor perspective on problem solving and creativity. *International Journal of Technology and Design Education, 33*, 217-240. https://doi.org/10.1007/s10798-021-09724-z

Holmén, J., Adawi, T., & Holmberg, J. (2021). Student-led sustainability transformations: Employing realist evaluation to open the black box of learning in a Challenge Lab curriculum. *International Journal of Sustainability in Higher Education, 22*(8), 1-24. https://doi.org/10.1108/IJSHE-06-2020-0230

Ingram, C., Langhans, T., & Perrotta, C. (2022). Teaching design thinking as a tool to address complex public health challenges in public health students: A case study. *BMC Medical Education, 22*, 270. https://doi.org/10.1186/s12909-022-03334-6

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kienzler, H., & Fontanesi, C. (2017). Learning through inquiry: A Global Health Hackathon. *Teaching in Higher Education, 22*(2), 129-142. https://doi.org/10.1080/13562517.2016.1221805

Medina Angarita, M. A., & Nolte, A. (2020). What do we know about hackathon outcomes and how to support them? A systematic literature review. In A. Nolte, C. Alvarez, R. Hishiyama, I.-A. Chounta, M. Rodríguez-Triana, & T. Inoue (Eds.), Collaboration technologies and social computing: CollabTech 2020 (Lecture Notes in Computer Science, Vol. 12324). Springer. https://doi.org/10.1007/978-3-030-58157-2_4

Menéndez-Varela, J. L., & Gregori-Giralt, E. (2018). The reliability and sources of error of using rubrics-based assessment for student projects. *Assessment & Evaluation in Higher Education, 43*(3), 488-499.

https://doi.org/10.1080/02602938.2017.1360838

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

Overwien, A., Jahnke, L., & Leker, J. (2024). Can entrepreneurship education activities promote students' entrepreneurial intention? *The International Journal of Management Education, 22*(1), 100928. https://doi.org/10.1016/j.ijme.2023.100928

Phi, G. T., & Waldesten, T. (2021). Educating sustainability through hackathons in the hospitality industry: A case study of Scandic hotels. *Scandinavian Journal of Hospitality and Tourism, 21*(2), 212-228. https://doi.org/10.1080/15022250.2021.1879669

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Schulten, C., & Chounta, I.-A. (2024). How do we learn in and from hackathons? A systematic literature review. *Education and Information Technologies, 29*, 20103-20134. https://doi.org/10.1007/s10639-024-12668-1

Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology, 36*(2), 102-107. https://doi.org/10.1080/00986280902739776

Szymanska, I., Sesti, T., Motley, H., & Puia, G. (2020). The effects of hackathons on the entrepreneurial skillset and perceived self-efficacy as factors shaping entrepreneurial intentions. *Administrative Sciences, 10*(3), 73. https://doi.org/10.3390/admsci10030073

Tsay, C.-J. (2021). Visuals dominate investor decisions about entrepreneurial pitches. *Academy of Management Discoveries, 7*(3), 343-366. https://doi.org/10.5465/amd.2019.0234

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840-854. https://doi.org/10.1080/02602938.2017.1412396

Watson, K., McGowan, P., & Cunningham, J. A. (2018). An exploration of the Business Plan Competition as a methodology for effective nascent entrepreneurial learning. *International Journal of Entrepreneurial Behavior & Research, 24*(1), 121-146. https://doi.org/10.1108/IJEBR-05-2017-0158

Wiek, A., Withycombe, L., & Redman, C. L. (2011). Key competencies in sustainability: A reference framework for academic program development. *Sustainability Science, 6*(2), 203-218. https://doi.org/10.1007/s11625-011-0132-6

Yu, Q., Yu, K., & Lin, R. (2024). A meta-analysis of the effects of design thinking on student learning. *Humanities and Social Sciences Communications, 11*, 742. https://doi.org/10.1057/s41599-024-03237-5