TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Zhu, C., Buckley, J., Klapwijk, R. M., Spandaw, J. G., & de Vries, M. J. (2026). A holistic look at creativity: evaluating pupils' creative design ideation and prototypes through comparative judgment. *International Journal of Technology and Design Education*, *36*(3), 1019-1046. https://doi.org/10.1007/s10798-025-10027-w

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

International Journal of Technology and Design Education (2026) 36:1019–1046
https://doi.org/10.1007/s10798-025-10027-w

A holistic look at creativity: evaluating pupils' creative design ideation and prototypes through comparative judgment

Caiwei Zhu[1] · Jeffrey Buckley[2] · Remke Klapwijk[1] · Jeroen Spandaw[1,3] · Marc J. de Vries[1]

Accepted: 31 July 2025 / Published online: 15 September 2025
© The Author(s) 2025

Abstract
Developing creative solutions to improve our surroundings is a key 21st-century competency. Design & Technology (D&T) education presents valuable opportunities to teach creativity as a skill. However, the ill-defined and context-dependent nature of design problems often makes it challenging for educators to adequately evaluate the creativity demonstrated in pupils' solutions. Comparative judgment, which does not rely on a predetermined set of evaluative criteria, offers an alternative approach. In this study, we leveraged this method to investigate how 20 industrial design students, acting as judges, holistically assessed design ideas and prototypes produced by 201 pupils aged 10 to 14 in the Netherlands. Although creativity is acknowledged as central to design quality, it is not prioritized in many current D&T projects. To address this gap, we deliberately focused on evaluating the creativity evident in pupils' designs. We further explored how judges' evaluative considerations, coded as criteria, shifted from the beginning to the end of the comparative judgment process. Our findings from qualitative and quantitative analyses added to our understanding of the multifaceted process of evaluating creativity and provided practical insights into using comparative judgment as an assessment tool in design education.

Keywords Comparative judgment · Creativity · Children's design · Design education

Introduction

Creativity—one of the key 21st-century competencies (Voogt & Roblin, 2012)—is widely recognized as a crucial factor driving innovative designs (Christiaans, 2002; Cropley & Cropley, 2010; Dorst & Cross, 2001; Kimbell & Stables, 2007; Lewis, 2005; Sarkar &

✉ Caiwei Zhu
c.zhu-1@tudelft.nl

[1] Science and Engineering Education Section, Faculty of Applied Sciences, Delft University of Technology, Delft, Netherlands

[2] Department of Technology Education, Faculty of Engineering and Technology, Technological University of the Shannon: Midlands Midwest, Athlone, Ireland

[3] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands

Chakrabarti, 2011). The National Curriculum for Design and Technology subject in England, for instance, highlighted one of its learning goals as for pupils to "develop the creative, technical and practical expertise needed to perform everyday tasks confidently and to participate successfully in an increasingly technological world" (Department for Education, 2013). Design and Technology (D&T) classrooms offer untapped potential for exploring how pupils utilize and develop creativity (Benson & Lunt, 2011; Cropley & Cropley, 2010; Lewis, 2005, 2009; Xu et al., 2020). However, adequately evaluating creativity in design projects remains a challenge to this day, both in selecting the appropriate task for assessment and in determining how to weigh the various outputs generated from these projects (Casakin et al., 2010; Kimbell & Stables, 2007).

To assess creativity as a cognitive or psychological construct, psychologists have developed test-based measures that are context-independent, making them generalizable across large samples with established scoring guidelines that can be applied by trained raters. However, these creativity measures may not fully capture the nature of creativity within the D&T learning environment, where projects are not only open-ended but also context-specific and rooted in authentic real-world problems (Kimbell & Stables, 2007). To some extent, the design task itself influences how creative a design product can be (Cropley, 2005). Design tasks having a few abstract constraints (e.g., easy to use) seem to have elicited more creative designs than those imposing more concrete constraints (e.g., budget) (Starkey et al., 2016). Selecting or developing a task to assess design capabilities—and choosing the appropriate method for evaluating a complex construct like creativity, whether it applies to the product, the designer, or the overall design process—requires careful consideration that takes into account the progress made over the past two decades in the D&T community.

Assessing qualities of works in D&T projects and the use of comparative judgment

Evaluating responses to open-ended tasks, especially those that require creativity or divergent thinking, has been a challenging topic for both educators and researchers (Jones & Alcock, 2014; Lesterhuis et al., 2017). Traditional methods that rely on predefined scoring rubrics can often be time-consuming and costly, accompanied by complications that arise from large disagreements among judges and judge fatigue (Bejar, 2012; Forthmann et al., 2017). Evaluators may be inconsistent with how they interpret the rating scale throughout the assessment (Hoskens & Wilson, 2001; Myford & Wolfe, 2003; Pollitt, 2012); or they may choose to rely on more straightforward and intuitive criteria, such as aesthetics or crafting quality, since characterizing constructs like creativity can be more complex and intricate (Casakin et al., 2010). Assessing designs by calculating the relative frequency of attributes, such as the relative rarity of a solution, appears to be more objective (Shah et al., 2003). However, these values need to be interpreted with caution or may otherwise lead to a biased assessment of the construct (Sluis-Thiescheffer et al., 2016).

In contrast to evaluations based on predefined criteria, comparative judgment leverages judges' expertise to holistically select the 'better' piece of work through rounds of paired comparisons, thereby revealing the relative quality of works through rank order (Hartell & Buckley, 2021). Comparative judgment validly and reliably assesses aspects that judges consider central to the evaluated construct (Pollitt, 2012) and appears to be well-suited for constructs that can be hard to precisely characterize, such as creativity

(Jones & Alcock, 2014). Studies employing adaptive comparative judgment—a form of comparative judgment with an adaptive algorithm managing the creation of pairs of work and their presentation to assessors—to evaluate the qualities of works from design and technology education in secondary or higher education contexts have reported high inter-rater reliability, ranging from 0.93 to 0.97 (Bartholomew et al., 2018, 2019; Buckley et al., 2022; Seery et al., 2019).

Most studies in D&T education that use comparative judgment have focused on evaluating the overall quality of design works (e.g., Kimbell, 2012; Seery et al., 2019; Strimel et al., 2021). Since many factors can influence judges' perceptions of quality, relying on judges' expertise for holistic assessments implies that they may apply diverse criteria during evaluation (Jones & Alcock, 2014; Lesterhuis et al., 2022). Qualitative findings from Buckley and colleagues (2022) revealed the diverse criteria judges used to select a winning design portfolio from a pair. These included the quality of crafting seen in the work, the quality of the design concept, effectiveness in communication and presentation, the emotion conveyed through the design, and the amount of effort or details evident in the design. Interestingly, how creative, unique, interesting, or adventurous the designs were seemed to be less frequently mentioned as a criterion. Similarly, in the qualitative analysis reported by Bartholomew and colleagues (2018), it appeared that more emphasis was placed on the suitability and feasibility of the design, the aesthetics, and the completeness of the portfolio, with creativity and innovation comprising as little as 5% of the comments made by judges. Another study revealed cultural differences in the design values identified by judges from different backgrounds, reflecting the varying perspectives on what constitutes quality design (Bartholomew et al., 2020). For instance, while D&T experts from the U.K. frequently emphasized innovation in their adaptive comparative judgment process, experts from Sweden and the U.S. mentioned innovation less often, instead prioritizing usability, adherence to design criteria, and effective idea communication.

These findings are somewhat surprising, given that creativity is widely recognized as a vital, even central, component of design quality (Christiaans, 2002; Goldschmidt & Tatsa, 2005; Kimbell & Stables, 2007; Sarkar & Chakrabarti, 2011). Lewis (2005, 2009) emphasized that while creativity is central to design and technology, it remains insufficiently addressed in teaching and learning, and thus demands more explicit curricular and pedagogical attention. Cropley & Cropley (2010) further highlighted assessing the creativity of design solutions as a key strategy for fostering creativity in technological design education. As recent comparative judgment studies (e.g., Bartholomew et al., 2018; Buckley et al., 2022) revealed that creativity-related criteria accounted for only a small portion of design quality evaluation, the question of what shapes our perception of creativity in today's D&T projects remains to be answered. To address this gap and gain deeper insight into how creativity is expressed in young pupils' design work, we proposed using comparative judgment specifically to assess creativity in design. This focus is novel, as comparative judgment studies in D&T education have primarily concentrated on overall design quality rather than creativity as a distinct construct. Furthermore, as comparative judgment reflects judges' conceptualizations of complex constructs (Lesterhuis et al., 2022) and has been successfully applied to assess solutions to various types of open-ended problems (Kimbell, 2012; Steedle & Ferrara, 2016; Strimel et al., 2021),

this method can be particularly well suited to unpacking the multifaceted nature of creativity in design.

This investigation is part of a larger research project examining the link between creativity in design and spatial ability. Sections regarding the selection, adaptation, and implementation of the design task were previously presented at a conference (Zhu & Klapwijk, 2024). In this article, we first examined the evaluative considerations used by judges when applying comparative judgment to assess creativity in pupils' design ideation and prototypes. Design ideation refers to the generation of preliminary concepts that may be further developed or discarded in subsequent stages of the design process (Lawson, 2005). Prototyping takes ideas one step further by embodying their role, implementation, as well as look-and-feel in visual or tangible forms (Houde & Hill, 1997). The focus on these early stages of design was because quality ideas are foundational to creative design outcomes (Goldschmidt & Tasta, 2005), and prototypes serve a range of meaningful roles throughout the innovation process (BenMahmoud-Jouini & Midler, 2020). We then examined how judges' evaluative considerations—coded as criteria—evolved from the beginning to the end of the comparative judgment process. Specifically, we investigated whether emphasizing creativity as the main assessment goal allowed judges to evaluate this construct through comparative judgment validly. In addition, we explored whether comparative judgment analysis may offer complementary insights to traditional grading methods by capturing nuanced differences in design ratings, enriching our understanding of design creativity, and revealing the unique challenges judges encounter during the comparative judgment process.

Methodology

Participants and setting

The design project was conducted in four international schools in the Netherlands. A total of 201 pupils (M_{age} = 12.52, SD_{age} = 1.14, 108 girls, 93 boys) completed all parts of a design task and other relevant tasks. These pupils came from 33 different nationalities, and 46.3% of them reported English as one of their native languages. On average, they had been enrolled in one or more international schools where English was the main language of instruction for 5.18 years (SD = 2.80). All design activities were conducted in English, and comprehension of the design brief and task prompts was essential for meaningful participation. While participants with limited English proficiency were supported by classroom teachers and allowed to take part, their data were excluded from the analysis.

Twenty master's students from the Department of Industrial Design Engineering at Delft University of Technology were recruited as judges. Participation opportunities were advertised within the department and via online student groups. Design students who expressed interest were invited to a brief interview and a 30-min training session led by the research team. This session included an overview of the design assignment and evaluation objectives, along with five practice trials using the comparative judgment platform to ensure that they understood the evaluation process and developed a basic familiarity with the pupils' design levels. Judges were selected based on their demonstrated understanding of the comparative judgment procedure. Of the final panel of 20 judges, five had recently completed

their master's degrees, while the remaining were current master's students. Thirteen were enrolled in the Design for Interaction track, five in Strategic Product Design, and two in Integrated Product Design. Twelve judges had prior experience with coursework or design projects related to designing for or with school-aged children. All judges received hourly compensation for their participation.

Selecting and adapting the design task

The *SnackSafe on the Beach* design task used in this study was inspired by a real-world problem—annoyances caused by seagulls. This authentic and engaging design challenge was developed by education researchers at the Science Hub at Delft University of Technology,[1] originally instructed in Dutch and has been carried out in multiple Dutch primary and secondary schools. During the design task selection process, we piloted this activity alongside other potential design tasks. Our aim was to identify a task that could stimulate creative design thinking by encouraging diverse interpretations and a variety of problem-solving strategies, thereby fostering divergent thinking (Klapwijk et al., 2021). Several alternative tasks were excluded because they imposed complex constraints that were difficult for pupils to address within the allotted time, or they tended to elicit similar types of solutions, thereby limiting creative variation.

For research purposes, we made several adaptations to the seagull annoyances design brief and procedure. First, we focused the design task on helping people enjoy their fries on the beach without being bothered by seagulls (Fig. 1). This specific design problem has previously been tested by educators at the Science Hub and was found to be personally relevant to pupils, feasible to complete individually at their age level, and effective in eliciting a wide range of creative responses. Second, we identified key steps in problem exploration, idea generation, and concept development that could be done individually. These steps were incorporated into an A3-size foldable design booklet to capture pupils' thought processes when solving the design task. To assess individual pupils' performance, each pupil completed all steps independently without collaboration. The design of the booklet was inspired by the design portfolio developed in the TERU project (Kimbell & Stables, 2007; Stables & Kimbell, 2000), which aimed to document pupils' idea progression, intermediate products, and reflections to illustrate the dynamic design process. A copy of the design booklet used in this study can be found in the supplementary materials.

Organizing the design task and the comparative judging sessions

The energizer and the design task

All design sessions began with a five-minute energizer activity intended to stimulate pupils' thinking and engagement. This practice aimed to help participants feel comfortable expressing ideas and actively thinking in different directions (Klapwijk et al., 2021). Pupils were asked to list as many fruit names as they could think of within one minute and stand up when the time was up. Those who had either 'apple', 'banana', or 'strawberry' among their

[1] https://www.ontwerpenindeklas.nl/les/overlastmeeuwen/



Fig. 1 The *SnackSafe on the Beach* design task

first three responses were then asked to sit down. Following this activity, pupils were told: "Oftentimes when we are asked to come up with many ideas, our first few ideas are not very uncommon or unique. However, by continuing to generate more ideas, you may come up with more unusual, uncommon, and creative ideas, or you might be able to make links between ideas to create unique combinations."

Next, the pupils were guided through the design booklet. The researcher read the design problem aloud and instructed them to spend three minutes writing down keywords for the features they believed their design would need to have to solve the design problem. Pupils were then asked to brainstorm as many ideas as possible on an A4 sheet with six empty boxes. They were encouraged with the prompt: "All ideas are welcome. Don't hesitate to sketch anything that comes to your mind. Every new idea can be a valuable addition." This prompt drew inspiration from known brainstorming rules advised by experienced designers, design educators, and researchers (IDEO.org, 2015; Klapwijk et al., 2021).

Following the initial brainstorming, pupils were explicitly instructed to be creative and generate four additional ideas, which could either be entirely new concepts or improvements on previous ideas. Pupils were further advised: "Try to think of unusual, original, exciting ideas that others won't easily think of. Think from different viewpoints and explore different directions. You can also jot down improved versions of your previous ideas or make unique combinations between your ideas." This instruction, adapted from Butler (1987) and drawn from the YourTurn Design Tool, developed through a collaboration between Delft University of Technology and Goldsmiths, University of London (Klapwijk et al., 2021), aimed to clarify the task objective, sustain engagement, encourage divergent thinking, and emphasize the importance of exploring different directions while postponing judgment.

After generating ideas, pupils were instructed to individually select the idea they considered the most original in addressing the design task. They then created detailed sketches of their chosen ideas, adding annotations to help others understand the features and intended functions of their design prototypes. Each pupil had 40 min to complete both ideation and prototype sketching, using only pencils for all drawings and notes. All pupils received equivalent instructions, materials, and resources. The above-mentioned prompts were intended to facilitate their idea development and reflect typical practices in design projects.

The comparative judging process

All ideas and design prototypes were photographed, anonymized, and uploaded to an online comparative judgment platform, No More Marking (n.d.), to create two comparative judgment sessions. When pupils' handwritten annotations were hard to distinguish, we transcribed them and placed the typed text alongside the original handwriting for clarity. Figures 2 and 3 illustrate the interfaces presented to the judges, where pupils' designs were randomly paired across multiple rounds for judges to determine the more creative one in each pair.

Judges were told to use their own understanding of creativity—which can be criteria developed from their personal or professional experiences—to evaluate the designs. The prompts displayed above each pair of works—either 'Which set of brainstormed ideas do you think is more creative?' or 'Which design prototype do you think is more creative?'—were made to maintain a holistic focus on assessing creativity.

Each judge completed 101 pairwise comparisons ([201 pieces of work × 10]/20 judges) separately for pupils' ideation and design prototypes. Judges were required to leave comments on the first and final 15 judgments they made to explain the criteria they used in making their decisions. They were also encouraged to leave additional comments on other comparisons if they felt that it would help the researchers better understand their decision-

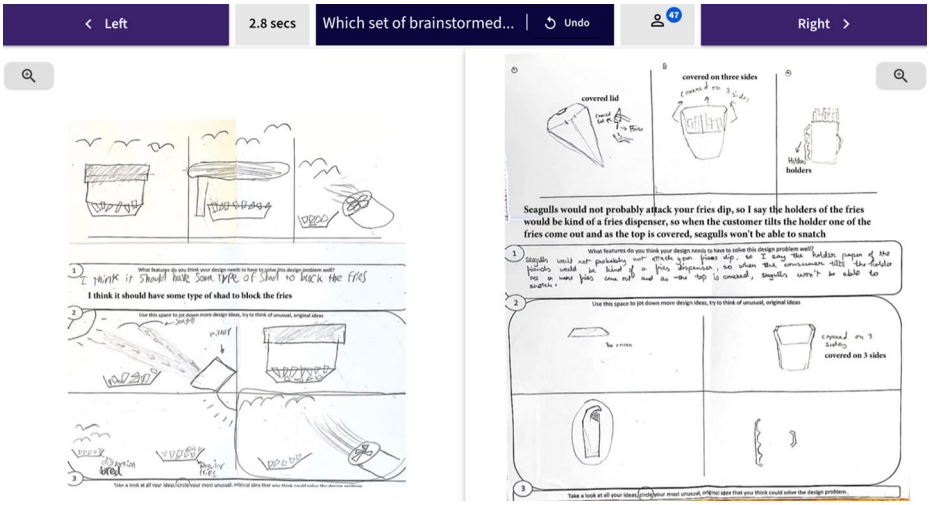


Fig. 2 A pair of design ideation sketches displayed side by side on the comparative judgment interface

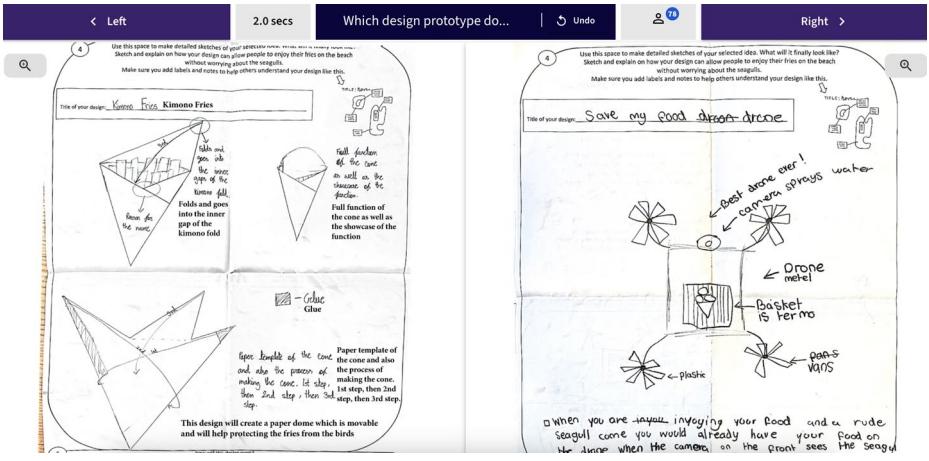


Fig. 3 A pair of design prototypes displayed side by side on the comparative judgment interface

making rationale. All judges first completed the judging session on design prototypes, followed by the session on design ideation (brainstormed ideas). This sequence was intentional. During the ideation phase, pupils were not required to provide annotations and were encouraged to produce only simple sketches. As a result, many of the brainstormed ideas were difficult to interpret on their own. Presenting the more detailed, annotated design prototypes first was intended to provide judges with a clearer sense of the pupils' design intentions and solution scope, thereby helping them better comprehend the ideation sketches evaluated afterward.

Data analysis

Data were analyzed across three aspects, combining qualitative interpretation with quantitative analysis. An overview of the analysis procedure is presented in Table 1. Given the open-ended nature of the design task, we first categorized the designs and documented the frequency of each type. We then examined whether certain types of designs were statistically more likely to be favored by judges.

Second, we analyzed judges' comments made during comparative judgment to understand the criteria they used to justify their decisions. The comments were imported into MAXQDA 2022 for qualitative coding. An inductive, data-driven approach was used to generate an initial coding scheme. A word combination frequency analysis (minimum 2 to maximum 5 words) was conducted to identify frequently referenced phrases, which informed the initial categorization of evaluation criteria. Each sentence in a piece of comment served as a unit of analysis and could be assigned multiple codes. The coding system was refined iteratively: codes were renamed, merged, or split into sub-codes to more accurately reflect the range of criteria and rationales expressed in the comments, with careful consideration of the context in which the comments were made. For example, the code 'user experience' was split into two sub-codes, 'judges' consideration of user experience' and 'pupils' consideration of user experience.' This distinction was intended to separate what judges inferred about user experience based on pupils' designs from pupils' own attention to user-related affordances reflected in their designs. Once the initial coding system was developed, a second coder reviewed and refined it using a set of randomly selected comments in which all initial codes appeared at least once. A third coder then independently coded 5% of the dataset, which was randomly selected using a random number generator. This step aimed to assess inter-coder agreement and further refine the codebook. Any discrepancies were discussed until full consensus was reached on code interpretations. The full dataset was then recoded using the finalized codebook.

Third, we examined how judges' use of evaluation criteria evolved as they proceeded through a large number of comparative judgments. Specifically, we analyzed whether the types of criteria applied shifted from early to later stages of the comparative judgment process. To explore this, we visualized changes in judges' use of criteria across different categories by comparing the first and last 15 comments made by each judge. We then tested for statistical differences in their tendency to apply different types of criteria over time.

Table 1 Overview of the data analysis procedure

	Analysis 1	Analysis 2	Analysis 3
Subject of analysis	Pupils' design prototypes	Judges' comments generated during comparative judgment	Judges' use of evaluation criteria over time
Outcomes	Frequencies of design types and statistical comparison of design rankings by types	Inductive coding of comments and iterative refinement of the codebook	Visualization and statistical comparison of criteria used in judges' first versus last 15 comments

Results

Comparative judgment outcomes and reliability

The average median time spent by the 20 judges for making a judgment was 68.6 s for judging design ideation and 73.0 s for judging design prototypes. The scale separation reliability (SSR), which is a Cronbach's alpha equivalent in comparative judgment, was.834 and.753 for judging design ideation and judging design prototypes respectively, both indicating good internal consistency. No critical misfit was found among judges, meaning that each judge's decisions consistently fit with the overall consensus (Pollitt, 2012).

The ideation and prototype scores were generated by uploading the judgment decisions to the Adaptive Comparative Judgement App (Buckley, 2024a, b), where the parameter values were determined by fitting the Bradley-Terry-Luce model using the Supplementary Item Response Theory Models (sirt) package (Robitzsch, 2021) in R. The parameter values indicate the relative ranking of each pupil's design ideation and design prototype works (see Figs. 4 and 5), and will be referred to as design ideation

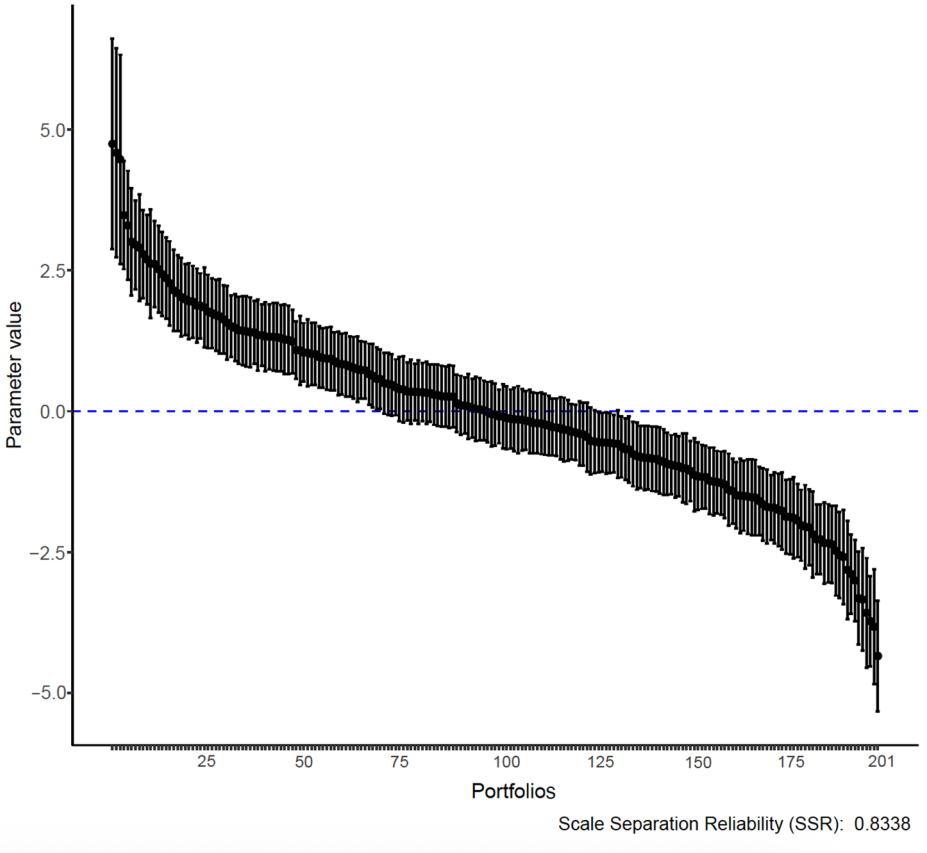


Fig. 4 Comparative judgment ranks for pupils' design ideation

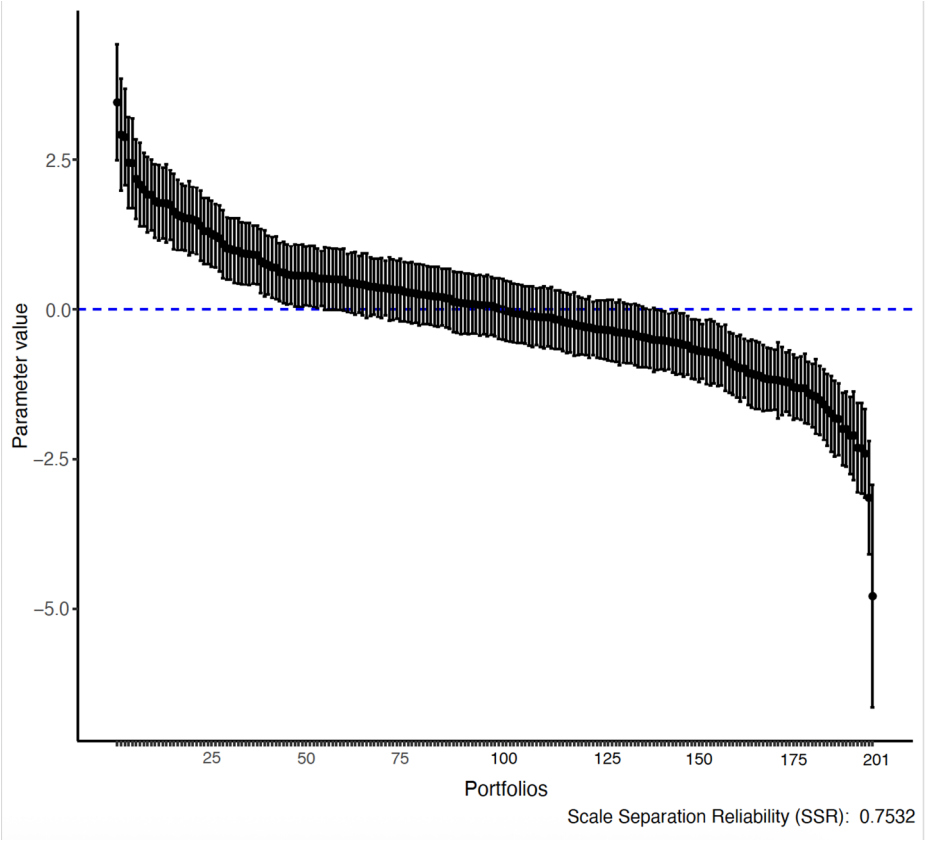


Fig. 5 Comparative judgment ranks for pupils' design prototypes

and design prototype scores. Spearman correlation between the parameter values pupils received for design ideation and design prototype was moderate, $r(199) = .371$, $p < .01$, 95% CI [.244,.486].

Types of designs developed

We observed that several judges attempted to categorize the design ideas during evaluation—either to facilitate their comparative assessment process or because comparing designs across different conceptual categories made decision-making more challenging. To illustrate the breadth and diversity of pupils' designs, we categorized them and further examined whether design scores varied across types. Pupils' design prototypes were classified into six categories: packaging, repellent, shielding, disguise, multi-element sys-

tem, and distraction. Table 2 presents the percentage of designs in each category, along with representative examples. Designs that clearly demonstrated one or two key features were categorized based on their most salient characteristics. Designs that integrated three or more distinct features spanning multiple categories were classified as multi-element system designs.

Based on the six types, we conducted another comparative judgment analysis by computing judges' decision data based on the types of design that were being compared. For example, instead of having 'prototype 1' versus 'prototype 2' in the decision data, where 'prototype 2' was chosen as the winner, we reformatted the data as 'Packaging' versus 'Distraction', where 'Distraction' was chosen as the winner. Results are presented in Fig. 6.

Table 2 Overview of types of design prototypes developed by pupils ($n = 201$)

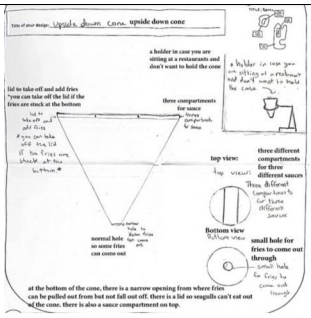
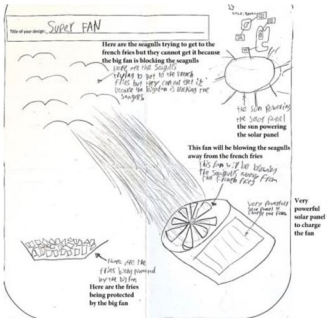
Types of design	Percentage	Design examples with typed annotations
Packaging	41.29 %	 An upside-down fries packaging that is covered on the top and has users pull fries from the narrow opening at its bottom
Repellent (physical/chemical)	24.38 %	 A solar-powered fan to blow seagulls away from the fires

Table 2 (continued)

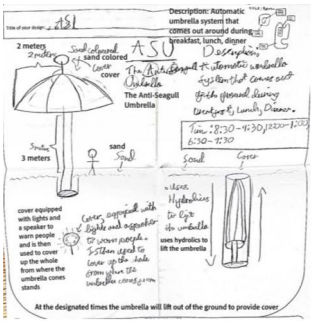
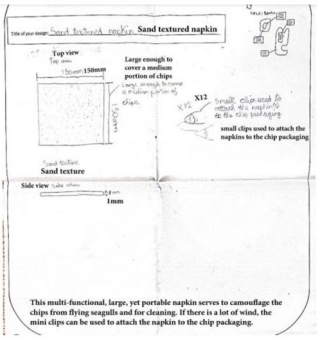
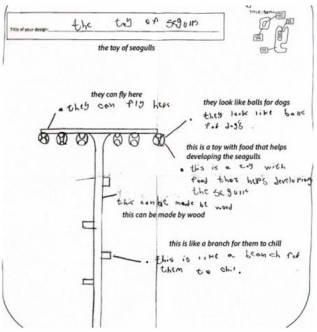
Shielding	13.43%	 An automatic umbrella system that is lifted out of the ground during mealtimes to shield people from seagulls
Disguise	7.96%	 An attachable, sand-textured napkin that camouflages fries from seagulls and can be used for cleaning after eating
Multi-element system	7.46%	 A fries delivery service that delivers fries to users by using seagull-shaped drones to deter real seagulls

Table 2 (continued)

Distraction	5.47%	 A toy for seagulls with food inside that attracts seagulls there

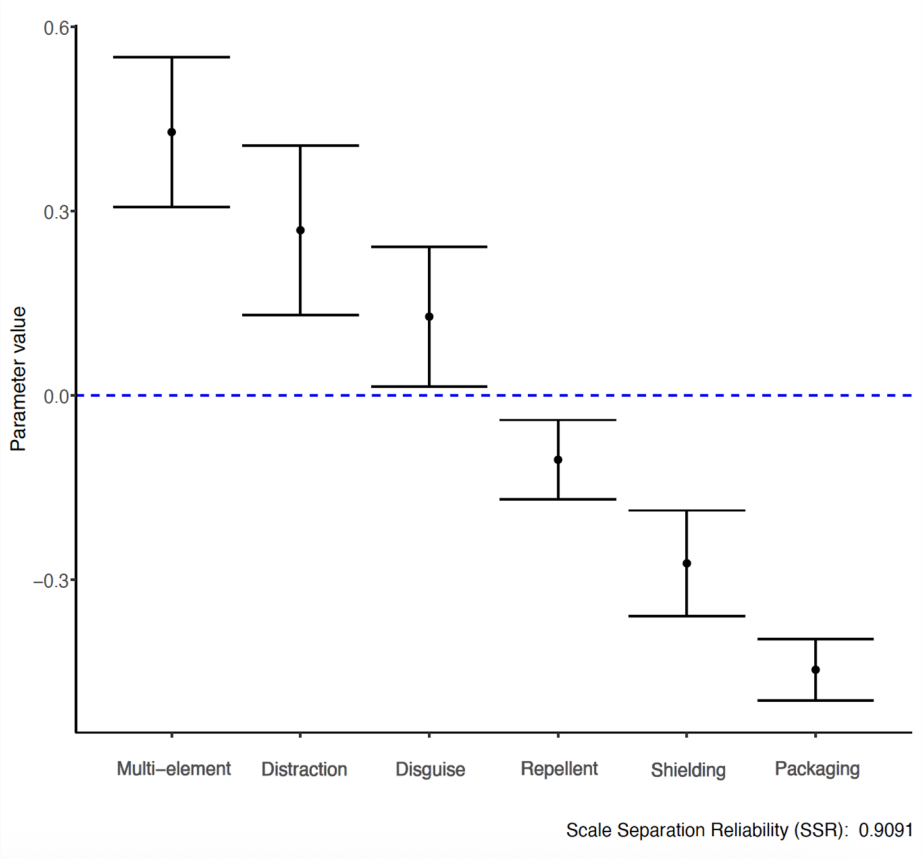


Fig. 6 Comparative judgment ranks by types of design

There appeared to be differences in how certain types of designs were ranked, with some types receiving higher ranks than others. To examine whether such differences were statistically significant, a Kruskal–Wallis H test was conducted in SPSS. A significant difference was found across design types, $H(5) = 13.67$, $p < 0.05$, providing evidence that certain types of designs were considered to be more or less creative than others. Follow-up post hoc pairwise comparisons (unadjusted p-values) revealed that packaging designs were rated significantly lower than both multi-element system designs and distraction designs ($p < 0.05$) and that repellent designs were also rated significantly lower than multi-element designs and distraction designs ($p < 0.05$). To account for potential inflation of Type I error due to multiple comparisons, significance levels were adjusted using Holm-Bonferroni correction (Holm, 1979) with an alpha level of 0.05. After adjustment, none of the pairwise differences remained statistically significant. These results suggest that while overall group differences exist, the differences may be subtle and distributed across design types rather than driven by a single pairwise comparison difference. It is also possible that the effect sizes of these differences were relatively small (Buckley, 2024c), resulting in limited statistical power to detect differences at the pairwise level.

Qualitative coding of judgment criteria

Each judge provided at least 30 pieces of comments separately for evaluating design ideation and design prototypes, resulting in a total of 600 pieces of comments for each design phase. While all judges were instructed to leave comments for the first and last 15 judgments they made, some judges provided additional comments throughout the process. To ensure that we analyzed a comparable number of comments across judges, we included only the first and last 15 pieces of comments provided by each judge in this analysis. This approach was intended to capture judges' insights from both the beginning and the end of the judging process.

On average, each comment on design ideation was 34 words in length, and each comment on design prototypes averaged 44 words. Two coders achieved good inter-rater agreement when assigning codes to selected comments regarding design ideation (77.9%) and design prototypes (79.9%). Any discrepancies were discussed until full agreement on code interpretations was reached.

Through iterative coding, a total of 39 codes were developed to capture judges' evaluative criteria. These codes were grouped into six overarching categories: *Novelty, Idea qualities*, *Usability*, *Feasibility*, *Presentation*, and *Problem-solving*. An additional category, *Idea generation*, was developed to capture aspects specific to evaluating design ideation. Table 3 presents the definitions for each category and lists the codes included under *Novelty*, which was of particular interest in this study. A complete description of the remaining codes can be found in the codebook provided in the supplementary materials.

The frequencies of the main code categories are shown in Figs. 7 and 8. Codes under each category are listed below in Table 4, along with their frequencies relative to the total number of code appearances ($n_{\text{Ideation}} = 1720$, $n_{\text{Prototype}} = 1935$).

Table 3 Qualitative coding of judges' evaluative considerations: categories and codes

Code categories	**Definition**	
Novelty	This category refers to how uncommon, unique, or unconventional the design (or idea) is	
Usability	This category refers to how well the design (or idea) supports users' needs and ease of interaction	
Presentation	This category refers to how well the design (or idea) is communicated and visually represented	
Idea qualities	This category involves a series of subjective attributes of the design (or idea)	
Feasibility	This category refers to the practicality of bringing the design (or idea) to life	
Problem-solving	This category refers to how comprehensively the design (idea) addresses the design brief	
Idea generation	This category refers to the variety and quantity of ideas produced during the ideation phase	
Novelty codes	**Definition**	**Examples**
Original/unique	Judges commenting on a design being innovative, new, original, unique, rare, fresh, not often seen, or proposing a new concept; also on a design being an obvious idea, similar to other frequently seen or existing solutions	"It stands out as an original idea compared to the more straightforward container solution" "Ideas on the left are a bit boring because it is obvious"
Uncommon mechanism (mechanical/structural)	Judges commenting on an uncommon mechanism—a specific structure or mechanical component—in the design, including the shape, the structure, the compartments and components, the way of assembling/attaching parts of the design, or the motion or interaction generated due to these compartments, that are unusual, novel, or not seen before	"The idea of linking all these elements and make them activate in sequence is really unique" "left design is more innovative using rubber mouth-like openings so it can keep original state automatically"
Creatively combining different ideas	Judges commenting on a design being a creative and meaningful combination of different concepts that may have otherwise been common ideas	"I like that the left one uses play predators in combination with the sound of playing kids" "I think the left design is a bit more out-of-the-box thinking with the use of water and sound to repel the seagulls"
Different starting point	Judges commenting on a design taking a different direction or perspective, changing the context, or reframing the scope of the problem; could be expressed in these forms:"instead of (a common way) the child did (new and different) way""it's not about a (common idea)…but about…."	"the child did not think about the fries or the packaging but rather what the seagulls really want which in the end is just food" "The child focuses on the delivery of the fries and not the packaging"
Modifying an otherwise usual idea	Judges commenting on the core design idea being common, but noting that the creative features added to the design made it distinct from the otherwise conventional types of usage	"they also thought of the bottle spraying based on a timer, which would make it a bit more innovative than already existing scent sprays" "While the core lies in covering the fries the creativity lies in the shape and the deception through a baby crib"

Table 3 (continued)

Creative/out-of-the-box (general)	Judges commenting on a design being 'creative' or 'out-of-the-box' in a general way that does not fall into the above-mentioned categories, that is, without mentioning that it is new and rare, exhibiting uncommon mechanisms, combining different ideas, taking a different direction, or modifying a usual idea	"in the end I think the left one is a bit more outside of the box thinking" "the drawing of the seagull is a creative idea"

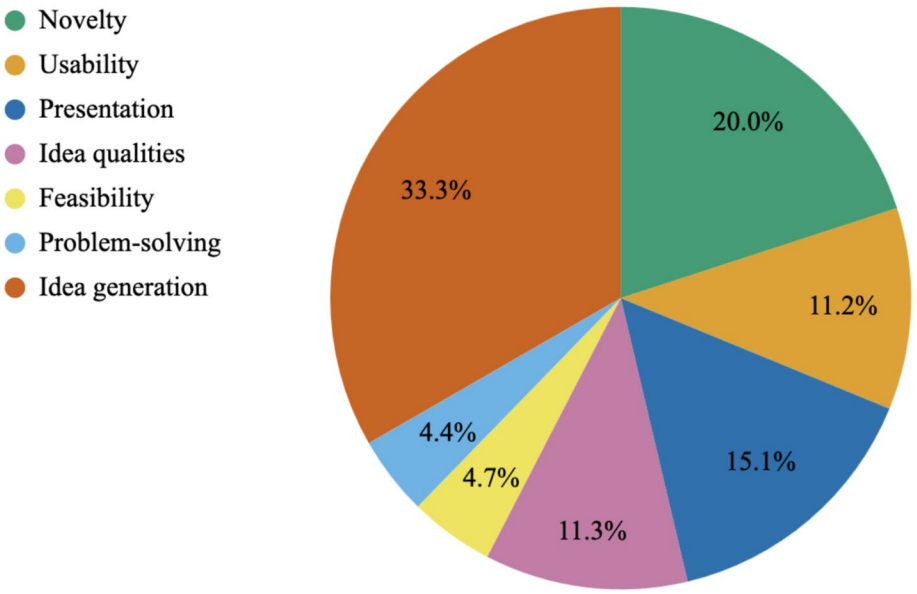


Fig. 7 Percentage of main code categories identified in design ideation comments

Shifts in criteria use from the beginning to the end of the judging process

The third analysis examined whether, and how, the frequency of judges' use of different criteria changed between the start and the end of the judging process. For each judge, we calculated the relative frequencies of criteria used within each code category based on their first 15 and last 15 recorded comments. Figures 9 and 10 visualize the shifts in judges' use of different evaluative criteria over time.

To examine whether the 20 judges'use of evaluation criteria changed significantly between their first and last 15 comments left during comparative judgment, we conducted non-parametric Wilcoxon Signed-Rank Tests in SPSS. No significant differences were found in the frequency of applying criteria from each category between their first and last 15 comments regarding design ideation. For design prototypes, judges appeared to use *Nov-*

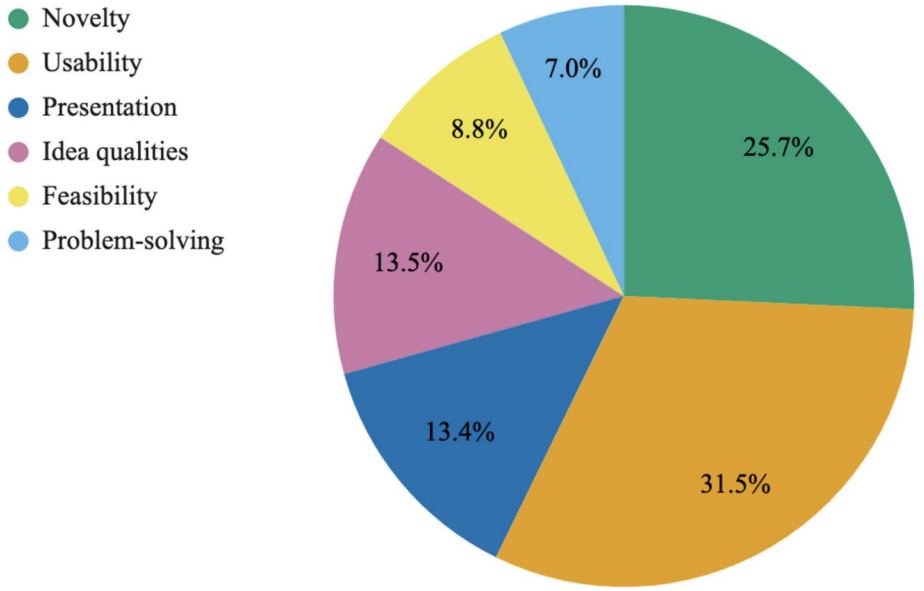


Fig. 8 Percentage of main code categories identified in design prototype comments

elty criteria ($Z=-2.092$, $p=0.036$, $p_{adjusted}=0.180$) more frequently and *Feasibility* criteria less frequently *(*$Z=-2.620$, $p=0.009$, $p_{adjusted}=0.054$) in their last 15 comments. However, the adjusted p-values (Holm-Bonferroni correction) suggested that these differences were no longer significant. Overall, the judges applied the coded criteria with largely comparable frequency at both the beginning and the end of the judging processes for both design ideation and design prototypes.

Discussion of main findings

This study illustrated how 201 pupils aged 10 to 14 generated a range of creative designs—packaging, shielding, physical or chemical repellent, disguise, distraction, and multi-element systems—to address a given design task. We explored the decisions and evaluative considerations provided by 20 industrial design students acting as judges, who used comparative judgment to assess both pupils'design ideation outcomes and design prototypes. The main findings, outlined below, are discussed by merging qualitative and quantitative results (Fetters et al., 2013).

Judges' decisions and types of designs

Our analysis of the design types produced by pupils revealed overall differences in how these types were ranked by the judges. This finding is not unexpected, as several judges noted that certain types of design were more creative than others. For example, Judge 3 remarked, "I like that it is not a packaging design but something to distract the seagulls

Table 4 Qualitative coding of judges' evaluative criteria: categories and codes

Category	Codes	Code frequency in design ideation comments (total code appearances = 1720)	Code frequency in design prototype comments (total code appearances = 1935)
Novelty	Original/unique	.099	.094
	Creative/out-of-the-box (general)	.062	.057
	Uncommon mechanism	.018	.027
	Creatively combining different ideas	.013	.027
	Different starting point	.008	.037
	Modifying an otherwise usual idea	—	.015
Usability	Useful & functional	.039	.102
	Simple & intuitive	.016	.045
	Pupils' consideration of user experience	.024	.045
	Judges' consideration of user experience	.010	.039
	Multiple elements/features	.013	.038
	Multiple functions/purposes	.006	.022
	Tailored for the context	.004	.013
	Customization	—	.009
Presentation	Clarity in explanation	.062	.075
	Elaborated details	.080	.044
	Quality of drawing	.008	.014
	Storytelling	.002	—
Idea qualities	Underdeveloped ideas	.035	.013
	Interesting	.026	.017
	Fun & playful	.011	.026
	Good (general)	.008	.019
	Smart	.008	.018
	Aesthetics & desirability	.008	.016
	Sustainability	.004	.012
	Idea potential	.008	.005
	Imagination	.003	—
	Considerate	—	.005
	Surprising	—	.003
	Explorative	—	.002
Feasibility	Realistic to make	.017	.035
	Involving technology	.018	.020
	Considering materials	.009	.019
	Cost-effectiveness	.003	.010
Problem-solving	Thought-through solutions	.026	.045
	Meeting the design brief	.019	.025
Idea generation	Diverse directions	.243	—
	Quantities of ideas	.052	—
	Variations of a key idea	.038	—

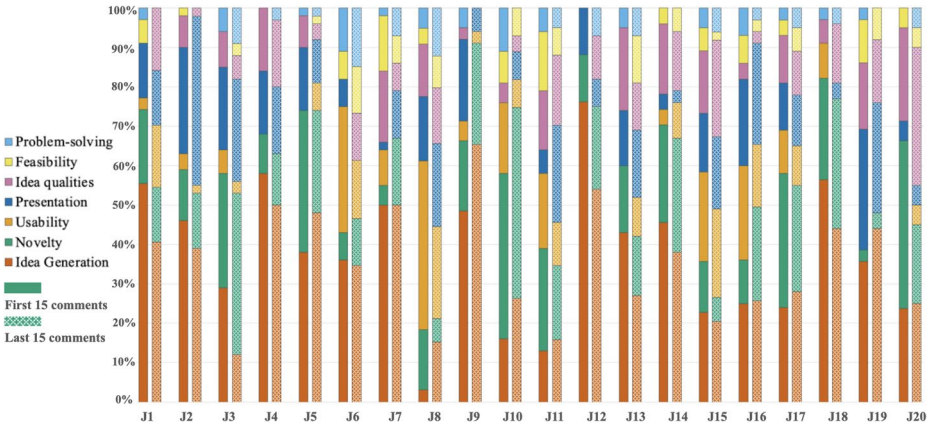


Fig. 9 Relative frequencies of criteria from seven code categories in the first and last 15 design ideation comments from 20 judges

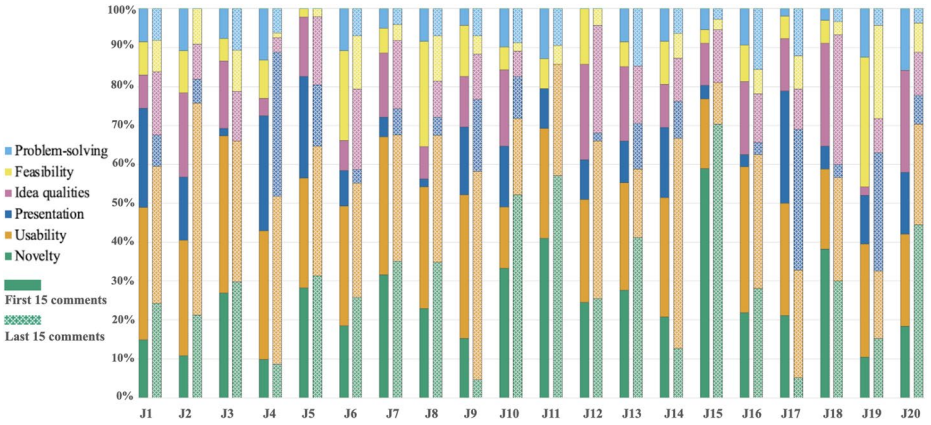


Fig. 10 Relative frequencies of criteria from six code categories in the first and last 15 design prototype comments from 20 judges

and keep them away due to the movement," and Judge 7 mentioned, "the designer thought out-of-the-box when designing a hat instead of a fry packaging which is good!" Packaging design was the most frequently attempted solution (41.29%), whereas other types, such as distraction designs, were relatively rare (5.47%). Since judges were explicitly asked to focus on creativity—which implies the need for rare and uncommon solutions—their evaluative considerations appeared to align with their comparative judgment decisions. Statistical analysis using Holm-Bonferroni–adjusted p-values revealed no significant differences in creativity rankings between design types. This suggests that although certain types of design (e.g., packaging) were more prevalent, they were not necessarily viewed as less creative overall by the judges.

Coding scheme and its alignment with existing frameworks

The qualitative coding scheme developed inductively from judges' comments on pupils' designs aligns closely with established frameworks for evaluating product creativity. Specifically, it reflects all three dimensions outlined in the Creative Product Semantic Scale (O'Quin & Besemer, 1989): novelty, resolution (usefulness and effectiveness in solving the design problem), and elaboration and synthesis. It also corresponds with the key criteria identified by Casakin & Kreitler (2005) for evaluating creative design solutions, including innovation, usefulness and functionality, fulfilling specified design requirements, elaborations, consideration of context, aesthetics, and the number of relevant solutions. Furthermore, our codes align with several core components of the Creative Solution Diagnosis Scale (Cropley & Kaufman, 2012), such as relevance ('solution fits within task constraints'), effectiveness ('the solution does what it is supposed to do'), multiple aspects of novelty ('the solution indicates a radically new approach'; 'the solution offers a fundamentally new perspective on possible solutions'; 'the solution makes use of new mixtures of existing elements'; 'the solution helps the beholder see new and different ways of using the solution') (p. 124), as well as how complete, pleasing, and sustainable the solution is. Lastly, the additional coding category developed for judges' comments on design ideation aligned with the key criteria for effective design ideation—variety and quantity of ideas—identified by Shah and colleagues (2003). Collectively, these alignments support previous findings that leveraging expert judgment can enrich the conceptualization of the assessed construct (Lesterhuis et al., 2022; Whitehouse, 2012).

Compared to previous studies that reported creativity or innovation as low-frequency criteria in the comparative judgment of design quality, our study took a different approach. Through explicitly instructing pupils to make creative designs and asking judges to prioritize creativity in their evaluations, *Novelty* was positioned as a central consideration in the assessment process. Nevertheless, considerations beyond *Novelty*—such as *Usability*, *Presentation*, and *Idea Qualities*—still played an important role in judges' decisions. In the evaluation of design ideation, judges' comments were most frequently coded under *Novelty* and *Idea Generation*, which together accounted for more than half of all codes. In contrast, *Usability* and *Feasibility* were referenced less frequently, aligning with the primary aim of design ideation—to encourage thinking outside the box without placing too much emphasis on practicality.

In the evaluation of design prototypes, judges referred more often to *Usability* ($f = 0.315$) than to *Novelty* ($f = 0.257$). This aligns with the well-established view that creative products need to be both novel and effective (Cropley & Kaufman, 2012; Horn & Salvendy, 2006; Runco & Charles, 1993; Sarkar & Chakrabarti, 2011). Technological design, in particular, requires relevant and workable solutions, as novelty alone is insufficient (Cropley & Cropley, 2010). Interestingly, our coding of judges' comments revealed that clearly distinguishing between what is considered'creative'and what is'useful'was not always possible—or even necessary. Several judges highlighted that a design's affordances contributed to its perceived creativity. For example, Judge 13 remarked, "It seems multifunctional and has utilities which consider the environment like being able to stand in the sand and the context-aware solution is interesting and creative." Similarly, Judge 18 noted, "The right one is a bit more creative to me because it does more than blocking the fries from the seagulls it scares them away as well." These comments illustrate how judges engaged in holistic comparisons

between solutions, blending multiple criteria in their assessment of creativity. This underscores a key advantage of comparative judgment over rubric-based scoring: judges are not burdened with the need to precisely differentiate between the desired qualities (Bejar, 2012; Pollitt, 2012) or assign discrete scores to separate criteria. Instead, they can draw on multiple considerations in a natural and intuitive way.

In summary, whereas conventional scoring methods rely on predefined criteria and emphasize fine-grained distinctions between criteria—even when the differences are subtle—comparative judgment allows evaluators to integrate various considerations into a cohesive judgment. Furthermore, we propose that comparative judgment may be particularly well suited to capturing insights that emerge from judges weighing multiple evaluation criteria—insights that, as Sadler (2008) noted, traditional assessment rubrics might overlook.

Comparison of the first and last 15 comments

As in design problem-solving, where tentative solutions and problem context co-evolve through ongoing reinterpretation and redefinition, resulting in a dynamic understanding of the design context (Dorst & Cross, 2001), judges' conceptualization of design quality can also shift as they make successive comparisons. Rubric-based assessment is susceptible to rater drift, where raters deviate from their original scoring standards over time, leading to inconsistencies and reduced reliability across the assessment process (Hoskens & Wilson, 2001; Myford & Wolfe, 2003). It is therefore important to examine whether judges' evaluative criteria in comparative judgment may likewise change over the course of the evaluation process.

Due to the random pairing of pupils' work, the designs that judges encountered at the beginning and end of the judging process likely differed, which may explain any observed changes in their rationale. However, with 20 judges each providing 30 pieces of comments—detailing their rationales for nearly one-third of their judgments—the sample size was sufficiently large to mitigate these variations and allowed us to analyze shifts in the criteria used, an aspect rarely reported in previous comparative judgment studies.

After Holm-Bonferroni correction, none of the Wilcoxon tests comparing the criteria judges used in the first and last 15 comments reached significance for either the ideation or prototype phase. These results indicate that judges' reliance on the various evaluation rationales remained largely stable from the beginning to the end of the assessment. Nevertheless, non-significant results do not confirm the absence of change. Both the visualizations and statistical analyses indicated that, by the end of their evaluation of design prototypes, judges referred more frequently to *Novelty* and less to *Feasibility* compared to the beginning. This may reflect a growing focus on the primary evaluation goal—creativity—as judges became more acquainted with the comparative judgment process. For instance, Judge 13 commented, "Irrespective of which is the better solution overall (it might be Left) the Right one feels more creative." Holistically evaluating the design means constantly considering trade-offs between criteria, such as between novelty and feasibility. On one hand, comparative judgment offers an integrated perspective that traditional methods of grading may not fully reflect. On the other hand, synthesizing multiple, and sometimes competing, evaluative considerations may be cognitively demanding for some judges (Forthmann et al., 2017).

It is also worth noting that the relative frequency of judges using different criteria (as shown in Figs. 9 and 10) does not necessarily indicate that they made different decisions. Judges could favor the same design but justify their choices using different rationales. For example, Judge 1 praised a design's creativity due to its details and that "all steps and the goal of the design are explained." Meanwhile, Judge 20 commented on its functionality and novelty—"Although there is a scope that this could get attacked by other seagulls to get the fries while it is delivering, the idea looks very distinct." Despite the differing justifications, both judges—and many others—agreed on the overall quality of this design. This highlights how, in comparative judgment, judges can provide different but valid rationales while still achieving good inter-rater reliability.

Reliability among judges and challenges in decision-making

The scale separation reliability (SSR)—which essentially measures how all judges collectively produce a similar ranking of design works—was found to be good in this study and comparable to previous research (Bramley & Vitello, 2019; Jones & Alcock, 2014; Verhavert et al., 2019). It is important to note that the pairing of pupils' work in our study was random rather than adaptive, whereas studies that reported even higher SSR of above 0.90 were predominantly using algorithms that adaptively present pairs of works with similar number of wins for more efficient comparison (e.g., Bartholomew et al., 2020; Buckley et al., 2022).

The qualitative coding of judges' comments revealed several challenges during the evaluation process. Some expressed difficulty when the paired works were similar, as Judge 19 noted, "It is a bit hard to judge and compare the two concepts because the ideas are not very different; the concept of working on the packaging in different ways has more creativity." Conversely, when the paired works were markedly different, judges faced the challenge of balancing various attributes. As Judge 16 explained, "left has more practical creativity while right has more imaginative creativity… hard to judge…will go with left since some features like the 'dark color to keep heat' help enhance the experience." Future studies may want to explore how judges deal with potentially conflicting rationales, and whether targeted training, evaluation guidelines, or improved pairing algorithms could help reduce the cognitive load during judgment.

Another potential factor influencing SSR was the shared background of the judges. All judges were trained in the same industrial design department, which may have contributed to a high degree of consensus, particularly when evaluating design ideation**.** A frequently cited criterion was the quantity of ideas. For example, Judge 1 noted, "Right shows more ideas. Quantity will lead to quality during brainstorm, therefore it is more creative." Similarly, Judge 4 commented, "As you need a lot of ideas in order to come up with the real creative ideas, I see more potential for people who are able to generate a lot of ideas." While research often supports the claim that quantity breeds quality (Kudrowitz & Wallace, 2013; Paulus et al., 2011), other studies have challenged this notion. For example, Goldschmidt & Tatsa (2005) found no correlation between design quality and the sheer number of ideas generated, as many ideas do not meaningfully address the design problem and would likely be discarded after the ideation phase. Therefore, while consensus among judges could enhance consistency in decision-making, it also raises concerns about potential bias stemming from homogeneity in evaluative perspectives.

Limitations and future directions

One limitation of using comparative judgment as an assessment tool lies in its focus on relative, rather than absolute, quality among the compared works. Since judges were asked to select the more creative design from each pair, it is possible—as some judges noted in their comments—that neither design was particularly creative. Accordingly, it is important to clarify that our outcomes reflect the relative creativity observed within the studied sample of pupils' design work. While this is consistent with the foundational principle of comparative judgment, it may be viewed as a limitation, as the relativity of the judgments could restrict the generalizability of our results—particularly in comparison to results derived from standardized rubrics. However, this limitation pertains more to our study design than to the comparative judgment method itself. Recent research has proposed strategies to transform the produced relative rank order into grades, which could then be compared across assessments (cf. Egelandsdal et al., 2025). Moreover, methods have been developed to merge and compare independent comparative judgement rank orders (Benton, 2021; Buckley & Canty, 2022; Buckley et al., 2023; Verhavert et al., 2022), further extending the applicability of this method in broader assessment contexts.

A second limitation concerns the composition of the judging panel. All judges involved in this study were master's-level design students from a single university department, which may have shaped a shared conceptualization of creativity. Previous research suggested that judges' view of design quality can be influenced by their cultural backgrounds (Bartholomew et al., 2020) and levels of professional experience (Strimel et al., 2021). As this study focused on evaluating the creative design work of primary and secondary school pupils, the absence of school D&T teachers on the judging panel limited the educational relevance of our findings. Future research could address this by involving a more diverse panel of judges—including professional design practitioners and D&T educators—to gain a more comprehensive understanding of creativity in pupils' design. In addition, emerging research highlights the potential of involving learners themselves in comparative judgment as a means of providing formative peer feedback and enhancing learning (Bartholomew et al., 2019, 2022). Building on this, future studies could explore if engaging pupils in comparative judgment helps foster their understanding and development of creativity in design.

Conclusion

By analyzing a total of 1200 comments provided by judges, this study unpacked the key rationales guiding their comparative judgment of pupils' creative design work. Our results revealed a consistent use of evaluative criteria across the judging process, with judges applying similar criteria from their initial to final comparisons. These criteria, identified through qualitative coding, aligned closely with established frameworks for assessing product creativity. Our findings support the growing body of research attesting to the validity and reliability of comparative judgment as an assessment method. Beyond this, comparative judgment appeared especially useful in capturing nuanced insights when assessing complex constructs such as creativity and design quality. Judges naturally integrated multiple evaluative criteria to form holistic decisions. These findings suggest that comparative judgment is not only a feasible alternative to rubric-based assessment but also offers rich insights for

evaluating creative, open-ended tasks. Future research could further examine how judges balance trade-offs among criteria, and how optimized pairing algorithms might help reduce cognitive load while maintaining judgment quality. Additionally, involving a broader range of evaluators—including educators and learners themselves—may yield a more comprehensive understanding of creativity in the context of design education.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10798-025-10027-w.

Acknowledgements We thank all the pupils and design students/graduates who participated in this study for their valuable contributions. We also thank Prof. Kay Stables and Ir. Eveline Holla for their feedback during the development of the design task used in this research. This research was funded by the Marie Skłodowska-Curie Innovative Training Network (grant no. 956124).

Declarations

Ethics Ethical clearance was obtained from the Human Research Ethics Committee, TU Delft. During the preparation of this work, the authors used ChatGPT by OpenAI to assist with improving writing and readability. The authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Informed consent Informed consent for participation was obtained from the parents or guardians of all pupils.

Conflict of interest We have no conflicts of interest including no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H., & Reeve, E. M. (2018). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educational Assessment, 23*(2), 85–101. https://doi.org/10.1080/10627197.2018.1444986

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education, 29*, 363–385. https://doi.org/10.1007/s10798-018-9442-7

Bartholomew, S. R., Ruesch, E. Y., Hartell, E., & Strimel, G. J. (2020). Identifying design values across countries through adaptive comparative judgment. *International Journal of Technology and Design Education, 30*(2), 321–347. https://doi.org/10.1007/s10798-019-09506-8

Bartholomew, S. R., Mentzer, N., Jones, M., Sherman, D., & Baniya, S. (2022). Bartholomew, S. R., Mentzer, N., Jones, M., Sherman, D., & Baniya, S. (2022). Learning by evaluating (LbE) through adaptive comparative judgment. *International Journal of Technology and Design Education, 32*(2), 1191–1205. https://doi.org/10.1007/s10798-020-09639-1

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues And Practice,31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

BenMahmoud-Jouini, S., & Midler, C. (2020). Unpacking the notion of prototype archetypes in the early phase of an innovation process. *Creativity and Innovation Management,29*(1), 49–71. https://doi.org/10.1111/caim.12358

Benson, C., & Lunt, J. (2011). We're creative on a Friday afternoon: Investigating children's perceptions of their experience of design & technology in relation to creativity. *Journal of Science Education and Technology,20*, 679–687. https://doi.org/10.1007/s10956-011-9304-5

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education, 6*, 775203. https://doi.org/10.3389/feduc.2021.775203

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice,26*(1), 43–58. https://doi.org/10.1080/0969594X.2017.1418734

Buckley, J. (2024a). An open-source adaptive comparative judgement app for technology education research and practice: Alpha version. *Journal of Technology Education, 36*(1), 58–82. https://doi.org/10.21061/jte.v36i1.a.4

Buckley, J. (2024c). Conducting power analyses to determine sample sizes in quantitative research: A primer for technology education researchers using common statistical tests. *Journal of Technology Education, 35*(2), 81–109. https://doi.org/10.21061/jte.v35i2.a.5

Buckley, J., Canty, D., & Seery, N. (2022). An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educational Studies, 41*(2), 313–331. https://doi.org/10.1080/03323315.2020.1814838

Buckley, J., & Canty, D. (2022). Assessing performance: Addressing the technical challenge of comparing novel portfolios to the "ACJ-Steady State". In*PATT39 conference: PATT on the Edge - Technology, Innovation and Education*, 523–537. St. John's, Newfoundland and Labrador, Canada.

Buckley, J., Seery, N., & Kimbell, R. (2023). Modelling approaches to combining and comparing independent adaptive comparative judgement ranks. In S. Davies, M. McLain, A. Hardy, & D. Morrison-Love (Eds.), PATT40: The 40th International Pupils' Attitudes Towards Technology Conference Proceedings 2023 (Vol. 1, October). Liverpool John Moores University.

Buckley, J. (2024b). Adaptive comparative judgement app. Accessed on January 15, 2025, from https://jeffbuckley1992.shinyapps.io/comparative_judgement/

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474. https://doi.org/10.1037/0022-0663.79.4.474

Casakin, H., & Kreitler, S. (2005). The nature of creativity in design. In J. S. Gero & N. Bonnardel (Eds.), *Studying designers*, *5* (pp. 87–100). Key centre of design computing and cognition.

Casakin, H., Davidovitch, N., & Milgram, R. M. (2010). Creative thinking as a predictor of creative problem solving in architectural design students. *Psychology of Aesthetics, Creativity, and the Arts, 4*(1), 31–35. https://doi.org/10.1037/a0016965

Christiaans, H. H. (2002). Creativity as a design criterion. *Communication Research Journal,14*(1), 41–54. https://doi.org/10.1207/S15326934CRJ1401_4

Cropley, D., & Cropley, A. (2010). Recognizing and fostering creativity in technological design education. *International Journal of Technology and Design Education, 20*, 345–358. https://doi.org/10.1007/s10798-009-9089-5

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: Non-expert raters and the creative solution diagnosis scale. *The Journal of Creative Behavior,46*(2), 119–137. https://doi.org/10.1002/jocb.9

Cropley, A. J. (2005). *Creativity and problem-solving: Implications for classroom assessment*. British psychological society.

Department for Education. (2013). National curriculum in England: Design and technology programmes of study. https://www.gov.uk/government/publications/national-curriculum-in-england-design-and-technology-programmes-of-study/national-curriculum-in-england-design-and-technology-programmes-of-study. Accessed 3 Apr 2025.

Dorst, K., & Cross, N. (2001). Creativity in the design process: Co-evolution of problem–solution. *Design Studies, 22*(5), 425–437. https://doi.org/10.1016/S0142-694X(01)00009-6

Egelandsdal, K., Hartell, E., & Færstad, J. O. (2025). Exploring the practical feasibility of adaptive comparative judgment as a summative assessment method. *Assessment & Evaluation in Higher Education*, 1–16. https://doi.org/10.1080/02602938.2025.2511787

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—Principles and practices. *Health Services Research,48*(6 Pt 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. *Thinking Skills And Creativity,23*, 129–139. https://doi.org/10.1016/j.tsc.2016.12.005

Goldschmidt, G., & Tatsa, D. (2005). How good are good ideas: Correlates of design creativity. *Design Studies,26*, 593–611. https://doi.org/10.1016/j.destud.2005.02.004

Hartell, E., & Buckley, J. (2021). Comparative judgment: An overview. *Handbook for Online Learning Contexts: Digital, Mobile and Open: Policy and Practice*, 289–307.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,6*(2), 65–70. https://www.jstor.org/stable/4615733

Horn, D., & Salvendy, G. (2006). Product creativity: Conceptual model, measurement and characteristics. *Theoretical Issues in Ergonomics Science, 7*(4), 395–412. https://doi.org/10.1080/14639220500078195

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden state examination. *Journal of Educational Measurement,38*(2), 121–145. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

Houde, S., & Hill, C. (1997). What do prototypes prototype?. In Handbook of human-computer interaction (pp. 367–381). North-Holland. https://doi.org/10.1016/B978-044481862-1.50082-0

IDEO.org. (2015). The field guide to human-centered design. https://www.designkit.org/resources/1.html . Accessed 10 Jul 2023

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education, 22*, 135–155. https://doi.org/10.1007/s10798-011-9190-4

Kimbell, R., & Stables, K. (2007). *Researching design learning: Issues and findings from two decades of research and development.* Springer Dordrecht. https://doi.org/10.1007/978-1-4020-5115-9

Klapwijk, R., Gielen, M., Schut, A., van Mechelen, M., & Stables, K. (2021). Your turn for the teacher: Guidebook to develop real-life design lessons for use with 8–14-year-old pupils. Accessed on June 21, 2023, from https://studiolab.ide.tudelft.nl/studiolab/codesignwithkids/files/Your-Turn-for-the-teacher-Guidebook.pdf

Kudrowitz, B. M., & Wallace, D. (2013). Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design, 24*(2), 120–139. https://doi.org/10.1080/09544828.2012.676633

Lawson, B. (2005). *How designers think: The design process demystified (4th ed.).* Routledge.

Lesterhuis, M., Bouwer, R., Van Daal, T., Donche, V., & De Maeyer, S. (2022). Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education,7*, Article 823895. https://doi.org/10.3389/feduc.2022.823895

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In *Innovative practices for higher education assessment and measurement* (pp. 119–138). IGI Global. https://doi.org/10.4018/978-1-5225-0531-0.ch007

Lewis, T. (2005). Creativity: A framework for the design/problem solving discourse in technology education. *Journal of Technology Education, 17*(1), 36–53. https://doi.org/10.21061/jte.v17i1.a.3

Lewis, T. (2009). Creativity in technology education: Providing children with glimpses of their inventive potential. *International Journal of Technology and Design Education, 19*, 255–268. https://doi.org/10.1007/s10798-008-9051-y

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

No More Marking. (n.d.). https://www.nomoremarking.com/. Accessed 1 May 2024.

O'Quin, K., & Besemer, S. P. (1989). The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal, 2*(4), 267–278. https://doi.org/10.1080/10400418909534323

Paulus, P. B., Kohn, N. W., & Arditti, L. E. (2011). Effects of quantity and quality instructions on brainstorming. *The Journal of Creative Behavior, 45*(1), 38–46. https://doi.org/10.1002/j.2162-6057.2011.tb01083.x

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Robitzsch, A. (2021). sirt: Supplementary item response theory models (Version R package version 3.10–118) [R]. https://cran.r-project.org/web/packages/sirt/index.html. Accessed 4 Oct 2024.

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences, 15*(5), 537–546. https://doi.org/10.1016/0191-8869(93)90337-3

Sadler, D. R. (2008). Transforming holistic assessment and grading into a vehicle for complex learning. *Assessment, learning and judgement in higher education* (pp. 1–19). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8905-3_4

Sarkar, P., & Chakrabarti, A. (2011). Assessing design creativity. *Design Studies, 32*(4), 348–383. https://doi.org/10.1016/j.destud.2011.01.002

Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2019). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *International Journal of Technology and Design Education, 29*(4), 701–715. https://doi.org/10.1007/s10798-018-9468-x

Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies, 24*(2), 111–134. https://doi.org/10.1016/S0142-694X(02)00034-0

Sluis-Thiescheffer, W., Bekker, T., Eggen, B., Vermeeren, A., & De Ridder, H. (2016). Measuring and comparing novelty for design solutions generated by young children through different design methods. *Design Studies, 43*, 48–73. https://doi.org/10.1016/j.destud.2016.01.001

Stables, K., & Kimbell, R. (2000). The unpickled portfolio: Pioneering performance assessment in design and technology. In R Kimbell (Ed.), *Design and technology international millennium conference*, 195--202. London: The design and technology association.

Starkey, E., Toh, C. A., & Miller, S. R. (2016). Abandoning creativity: The evolution of creative ideas in engineering design course projects. *Design Studies, 47*, 47–72. https://doi.org/10.1016/j.destud.2016.08.003

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education, 29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Strimel, G. J., Bartholomew, S. R., Purzer, S., Zhang, L., & Ruesch, E. Y. (2021). Informing engineering design through adaptive comparative judgment. *European Journal of Engineering Education, 46*(2), 227–246. https://doi.org/10.1080/03043797.2020.1718614

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice,26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Verhavert, S., Furlong, A., & Bouwer, R. (2022). The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. *Frontiers in Education, 6*, 785919. https://doi.org/10.3389/feduc.2021.785919

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies, 44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the adaptive comparative judgement method*. Manchester: AQA Centre for Education Research and Policy.

Xu, M., Williams, P. J., Gu, J., & Zhang, H. (2020). Hotspots and trends of technology education in the International journal of technology and design education: 2000–2018. *International Journal of Technology and Design Education,30*(2), 207–224. https://doi.org/10.1007/s10798-019-09508-6

Zhu, C. & Klapwijk, R. (2024). Assessing pupils' creative design ideas and prototypes using comparative judgment. In the 41st International Pupils' Attitudes Towards Technology (PATT) Research Conference in Nanjing, China (pp. 290–296).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.