Journal of Applied Data Sciences
Vol. 7, No. 3, September 2026, pp. 2335-2352
ISSN 2723-6471
2468

Data-Driven Evaluation of Task-Specific Performance in Refined Watercolor Brush Prototypes Using Repeated-Measures Analysis

Suraphan Chantanasut[1,*]

[1]Department of Art Education, Faculty of Fine and Applied Arts, Rajamangala University of Technology Thanyaburi, 39 Moo 1, Rangsit-Nakhon Nayok Road, Khlong Hok, Khlong Luang, Pathum Thani 12110, Thailand

(Received: February 18, 2026; Revised: May 1, 2026; Accepted: July 25, 2026; Available online: September 9, 2026)

Abstract

Watercolor brush performance emerges from the interaction between prototype configuration, bristle characteristics, material transfer, and the requirements of specific painting tasks. This study presents a data-driven evaluation of ten refined watercolor brush prototypes by integrating professional expert validation with controlled repeated-measures performance testing. The study used a sequential two-stage design. In Stage 1, 15 professional watercolor artists evaluated five round-handled and five flat-handled refined prototypes on material quality, grip comfort, movement control, water retention and color distribution, and professional-use potential. The overall expert evaluation mean was 4.65/5.00. In Stage 2, 24 experienced watercolor practitioners tested all ten prototypes across precision, graded-wash, and mixed technical tasks, producing 720 primary trials. Objective outcomes included task completion time, line-deviation error, stroke-width variability, wash-gradient error, wash smoothness, and paint consumption; task-specific usability was recorded after each trial. All 720 outputs were independently rated by three blinded watercolor experts, yielding 2,160 technical-quality ratings. Round configurations showed markedly lower precision error (0.442 vs. 1.243 mm) and shorter precision-task times (48.64 vs. 75.53 s), whereas flat configurations showed lower wash-gradient error (3.43 vs. 13.32 CIE L* units), higher wash smoothness (95.72% vs. 83.35%), and shorter wash-task times (39.40 vs. 67.42 s). Mixed-task differences were substantially smaller. A linear mixed-effects model confirmed a strong family-by-task interaction in completion time, while trial order was not significant ($p = 0.512$). Subjective control and satisfaction followed the objective task-specific pattern, whereas hand fatigue and wrist discomfort did not differ meaningfully between families. Blinded rating reliability was excellent (ICC(2,1)=0.969; ICC(2,3)=0.990). The findings indicate complementary task-specific specialization rather than universal superiority of one prototype family and demonstrate the value of repeated-measures, multimetric evaluation for data-driven artistic-tool design.

Keywords: Watercolor Brush, Data-Driven Design, Task-Specific Performance, Repeated-Measures Analysis, Mixed-Effects Modeling, Ergonomic Evaluation, Image-Based Measurement.

1. Introduction

Watercolor painting requires continuous coordination between the artist, the brush, water, pigment, and the painting surface. The brush therefore functions as a human–tool–material interface whose performance depends on both physical configuration and task demand. Human-factors research treats performance as an interaction among user capability, artifact characteristics, and the requirements of the activity [1]. Hand-tool studies further show that handle geometry, contact conditions, and perceived comfort can influence controllability and user experience, while objective and subjective responses do not necessarily change in parallel [2], [3], [4]. In professional watercolor practice, these interactions are especially important because artists shift rapidly between fine contour work, broad washes, glazing, and mixed detail–coverage operations within the same painting process.

Previous quantitative work by Chantanasut [5] evaluated a broad first-generation family of 24 innovative watercolor brushes against conventional brushes and reported advantages in task efficiency, paint use, satisfaction, and durability. That study established the feasibility of evaluating artistic tools through structured quantitative measurements rather than relying only on preference or anecdotal judgment. In parallel, contemporary data-driven design research increasingly emphasizes the acquisition and analysis of product- and user-related data to support evidence-based design decisions [6], [7]. However, the earlier watercolor-brush study treated the innovative designs primarily as an aggregate product family and did not resolve how a later refined set of prototypes behaves under different task requirements. Subsequent refinement therefore reduced the design space to ten prototypes—five round and five flat configurations—with clearer ergonomic, balance, and bristle-function distinctions. The refined prototypes first underwent professional

*Corresponding author: Suraphan Chantanasut (suraphan_c@rmutt.ac.th)
DOI: https://doi.org/10.47738/jads.v7i3.1574

expert evaluation. Fifteen professional watercolor artists assessed material quality, grip comfort, weight balance and movement control, water retention and color distribution, and potential for professional artistic creation. These evaluations were consistently high, with an overall mean of 4.65/5.00. Expert use also suggested a functional distinction: round configurations appeared well suited to precision and detailed control, while flat configurations appeared advantageous for broad application and graded washes. Nevertheless, descriptive expert ratings cannot by themselves establish whether these propositions are reflected in measurable performance. A data-driven validation requires controlled tasks, objective metrics, repeated observations, and statistical models that account for dependence when the same artists test multiple prototypes.

Accordingly, this study integrates the expert-validation stage with a controlled repeated-measures experiment using the same ten refined prototypes. Twenty-four experienced watercolor practitioners tested every prototype in three standardized contexts: a precision line-tracking task, a graded-wash task, and a mixed technical watercolor task. The resulting dataset combines 720 primary trials, prototype-level physical measurements, post-task usability responses, 720 uniquely coded physical outputs, and 2,160 blinded technical-quality ratings. The study evaluates family-level and prototype-level patterns while explicitly treating round-versus-flat differences as integrated configuration effects rather than isolated causal effects of handle geometry.

The study addresses four hypotheses. H1 proposes that round prototype configurations will show superior precision-task performance, reflected by lower line-deviation error, lower stroke-width variability, shorter task time, and higher perceived control. H2 proposes that flat configurations will show superior graded-wash performance, reflected by lower gradient error, higher smoothness, shorter task time, and higher wash-task usability. H3 proposes that the relative advantage of each family will depend on task type, producing a family-by-task interaction rather than a universal family hierarchy. H4 proposes that family differences will be smaller in the mixed technical task, where both fine and broad application capabilities are required. Through these hypotheses, the study aims to provide a reproducible, multimetric framework for applied data analysis in professional artistic-tool evaluation.

2. Literature Review

2.1. Watercolor Brush Design and Functional Interaction

Watercolor-brush function emerges from an integrated set of design characteristics that includes brush-head geometry, bristle composition and length, ferrule construction, handle dimensions, mass distribution, and liquid-holding capacity. The earlier 24-brush evaluation demonstrated that such artistic tools can be studied through measurable outcomes including task time, paint use, satisfaction, and durability [5]. More broadly, computational art research has shown that brushwork itself can be represented using quantitative geometric features such as direction, curvature, width, and spatial variability [8]. These developments support treating brush performance not only as a qualitative artistic phenomenon but also as a reproducible source of design and performance data. In the present prototype set, round and flat families differ simultaneously in handle width, overall length, mass, center of gravity, brush-head geometry, bristle blend, stiffness, and water-retention capacity. Consequently, a family comparison is an integrated configuration contrast. This distinction is important because task performance can reflect several interacting physical characteristics at once; the study therefore avoids attributing any observed family effect to handle cross-section alone. The practical question is instead whether the tested configurations show repeatable specialization when matched with different watercolor tasks.

2.2. Ergonomics of Hand-Held Tools

Ergonomic hand-tool research consistently shows that handle dimensions, shape, hand size, force distribution, and surface interaction influence grip capability, comfort, torque production, and controllability [2], [3], [4], [9]–[18]. User-reported comfort in hand tools includes functionality, physical interaction, posture, irritation, and handle characteristics [2], while objective measures of pressure, muscle activity, productivity, and subjective comfort may show only partial correspondence [3], [4]. Experimental work has further demonstrated systematic effects of handle diameter and orientation on grip force and torque [9], [10], relationships between hand size and handle performance [11], [14], [18], effects of elliptic versus circular geometry [12], and changes in whole-hand force distribution across handle diameters [13]. These findings provide a strong rationale for recording both prototype dimensions and participant hand anthropometry rather than treating ergonomics as a single subjective score.

The ergonomics literature also cautions against equating controllability with reduced physical strain. Grip-force capacity can improve under a particular handle dimension without producing an equivalent change in perceived comfort, and a configuration that feels easier to control may not reduce short-duration fatigue [3], [4], [15], [16], [17]. For this reason, the present study separates grip stability, perceived control, ease of execution, hand fatigue, and wrist discomfort, allowing beneficial task-control effects to be distinguished from claims about biomechanical relief.

2.3. Data-Driven Product and Creative-Tool Evaluation

Data-driven design seeks to connect design decisions with structured evidence gathered from products, users, and operational contexts [6], [7]. In creative-tool evaluation, this is particularly valuable because performance is simultaneously technical and experiential. Completion time and material consumption quantify aspects of execution, but they do not establish whether the resulting output is technically acceptable. Conversely, usability and satisfaction ratings capture experience but can be influenced by familiarity, expectations, and task preference. Product-design and user-experience research therefore supports combining behavioral performance with subjective evaluation rather than treating either as sufficient on its own [19], [20], [21].

The previously published 24-brush study [5] used task time, paint use, satisfaction, and durability to compare an earlier innovative brush generation with conventional products. The present study advances that approach in four ways. First, it retains individual prototype identities instead of treating the new family only as one aggregated condition. Second, it introduces image-derived task-specific metrics for precision and wash execution. Third, it uses a fully crossed repeated-measures design in which every participant tests every prototype. Fourth, it adds blinded output-quality assessment so that faster performance or favorable user ratings are interpreted alongside the technical quality of the resulting watercolor output.

2.4. Image-Derived Measurement of Brushstroke and Color Performance

Digital image analysis provides a reproducible route for converting artistic outputs into quantitative variables. Recent computational work on painting has demonstrated that brushstroke structure can be modeled from image-derived geometric features and used to characterize changes in artistic execution [8]. ImageJ and related open image-analysis platforms provide established tools for spatial measurement, profile extraction, segmentation, and reproducible image processing [22], [23]. For wash evaluation, CIELAB-based color science provides standardized representations of lightness and color difference, with well-documented formulations and implementation considerations [24], [25]. These foundations motivate the use of line-deviation error, stroke-width variability, and graded-wash luminance error as task-specific outcome measures in the present study.

2.5. Repeated-Measures Modeling and Rating Reliability

Repeated-measures designs are advantageous when individual skill differences are large because every participant can serve as his or her own comparison. Traditional repeated-measures ANOVA can be restrictive when observations are nested, unbalanced, or correlated in more complex ways, whereas mixed-effects models explicitly represent repeated observations and participant-level heterogeneity [26]. Random-effects specification influences generalizability and inference [27], and modern mixed-model frameworks provide a principled way to estimate fixed task effects while retaining trial-level information [28]. These considerations are directly relevant to a 24-participant × 10-prototype × 3-task dataset. Blinded expert ratings create a second dependency structure because multiple judges evaluate the same physical outputs. Intraclass correlation coefficients (ICCs) provide a formal basis for quantifying inter-rater reliability [29], [30]. Reporting should identify the exact ICC formulation and whether reliability refers to a single rater or the mean of multiple raters [31]. Accordingly, the present study reports ICC(2,1) for a single blinded rater and ICC(2,3) for the mean of three raters.

3. Methodology

3.1. Study Design and Participants

A sequential integrated design was used. Stage 1 consisted of professional expert evaluation of ten refined watercolor brush prototypes. Stage 2 consisted of a controlled, fully crossed repeated-measures experiment using the same ten prototypes. The integration was designed to determine whether favorable expert judgments and proposed round-versus-

ISSN 2723-6471

flat functional distinctions were reflected in measurable task outcomes. Figure 1 summarizes the complete evidence pathway from prototype refinement to multimeric analysis.

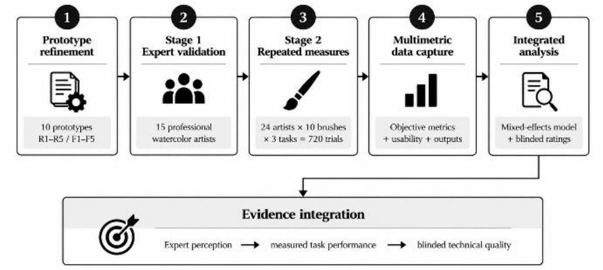

Figure 1. Integrated two-stage study design and evidence pathway from prototype refinement to repeated-measures and blinded-output analysis.

Stage 2 included 24 experienced watercolor practitioners. Thirteen participants (54.2%) were professional artists, six (25.0%) were advanced illustrators, and five (20.8%) were art educators. Mean watercolor experience was 12.88 ± 7.04 years (range 4–28 years). Twenty-one participants were right-handed and three were left-handed. Thirteen participants reported using both round and flat brushes as their primary preference, nine preferred round brushes, and two preferred flat brushes. Mean hand length was 178.40 ± 8.54 mm, mean palm width was 79.77 ± 5.91 mm, mean middle-finger length was 79.25 ± 4.21 mm, and mean preferred grip diameter was 11.92 ± 1.18 mm. Four participants had also participated in the earlier professional expert panel, while 20 formed an independent Stage 2 evaluation cohort. Recording hand anthropometry is consistent with evidence that hand size and handle dimensions jointly affect grip behavior and tool fit [11], [14], [18].

Stage 2 testing was conducted under ethics protocol IRB-RMUTT-2026-084, and informed consent was obtained from all participants before testing. Analytical datasets used participant codes rather than names. The Stage 1 component incorporated only previously collected aggregate, de-identified expert-evaluation results; no participant-level Stage 1 responses were reconstructed or reanalyzed in the present study.

3.2. Prototype Characterization and Stage 1 Expert Evaluation

The research object was fixed at ten refined prototypes: five round configurations (R1–R5) and five flat configurations (F1–F5). Each prototype was individually characterized for overall length, maximum handle diameter or width, minimum grip dimension, total mass, center of gravity, handle material and geometry, bristle blend and cut, bristle stiffness, and water-retention capacity. These variables were selected because hand-tool studies show that diameter, shape, hand-size compatibility, orientation, and force distribution can affect handling performance [9]–[18]. Water retention was measured in five bench replicates per brush after oven drying, immersion in deionized water to the full bristle length up to the ferrule junction for 5.0 s, 3.0 s vertical drainage, and re-weighing on an Ohaus Pioneer PX224 analytical balance. Water-retention capacity was calculated according to Eq. (1).

$$R_w = m_{wet} - m_{dry} \quad (1)$$

Rw is retained water mass, mwet is the brush mass immediately after standardized immersion and drainage, and mdry is the dry brush mass. Bristle stiffness was represented as transverse force per unit deflection (N/mm) under a standardized 3.0-mm point-deflection procedure. Representative flat- and round-handled prototypes from the refined study set are illustrated in figure 2, while figure 3 summarizes the measured multidimensional physical profiles of the ten prototypes on a standardized scale.

Figure 2. Representative refined watercolor brush prototypes illustrating the flat-handled and round-handled configuration types used in the ten-prototype study set.

Figure 3. Standardized physical profiles of R1–R5 and F1–F5. Values are z-score standardized across the ten prototypes for visualization; raw measurements are reported in Table 1.

The physical specifications and standardized bench-test characteristics of the ten refined watercolor brush prototypes are presented in table 1. The prototypes varied in length, maximum width, minimum grip diameter, mass, center of gravity, stiffness, and water-retention capacity, providing an objective physical basis for interpreting the subsequent professional evaluations. Figure 3 shows the corresponding standardized multivariable physical profiles. Because the round and flat brush families differ across multiple physical characteristics, subsequent family effects are interpreted as configuration-level effects rather than as effects attributable to a single physical parameter.

Table 1. Physical specifications and bench-test characteristics of the ten refined watercolor brush prototypes.

Brush	Family	Length (mm)	Max width (mm)	Min grip (mm)	Mass (g)	CG (mm)	Stiffness (N/mm)	Water retention (g, mean±SD)
R1	Round	248.0	12.5	9.0	21.8	108.5	0.42	1.543±0.075
R2	Round	252.0	13.0	9.4	23.4	105.2	0.48	2.052±0.049
R3	Round	256.5	14.2	9.8	25.1	102.0	0.55	2.857±0.090
R4	Round	260.0	15.0	10.2	27.6	98.4	0.62	3.590±0.118
R5	Round	264.0	16.2	10.8	31.2	95.0	0.38	4.929±0.151
F1	Flat	198.0	18.2	12.0	26.5	92.0	0.72	2.438±0.091
F2	Flat	204.0	20.5	13.2	29.8	88.5	0.68	3.818±0.140
F3	Flat	208.5	23.0	14.5	33.2	85.0	0.58	5.477±0.170
F4	Flat	212.0	25.4	15.6	37.0	81.2	0.52	7.295±0.129
F5	Flat	216.0	28.0	16.8	42.5	78.0	0.44	10.400±0.309

*CG = center of gravity measured from the brush tip. Water-retention values summarize five standardized bench replicates per prototype. Figure 3 shows the corresponding standardized multivariable physical profiles. Because round and flat families differ on multiple physical characteristics, subsequent family effects are interpreted as configuration-level effects.

In Stage 1, 15 professional watercolor artists evaluated the refined prototypes using a five-point scale. The evaluation dimensions comprised: (1) quality of materials and craftsmanship; (2) comfort and ease of grip during use; (3) balance of weight and control of brush movement; (4) water retention capacity and color distribution on paper; and (5) potential to support professional-level artistic creation. The resulting item-level means and standard deviations are summarized in table 2 and were retained as descriptive baseline evidence. Because the original participant-level Stage 1 questionnaire dataset is not currently available, Stage 1 and Stage 2 are integrated at the evidence-pattern level rather than through unsupported participant-level correlation. Accordingly, the expert evaluations in table 2 are interpreted alongside the physical and bench-test characteristics reported in table 1, without implying participant-level statistical associations that cannot be independently verified.

Table 2. Stage 1 professional expert evaluation of the refined prototypes (n=15).

Evaluation dimension	Mean	SD
Quality of materials and craftsmanship of the brush	4.73	0.46
Comfort and ease of grip during use	4.67	0.49
Balance of weight and control of brush movement	4.60	0.51
Water retention capacity and color distribution on paper	4.67	0.40

Evaluation dimension	Mean	SD
Potential of the brush to support professional-level artistic creation	4.60	0.25
Overall satisfaction	4.65	0.42*

*The overall mean is retained from the Stage 1 report. The reported overall SD of 0.42 has not been independently recomputed from participant-level responses and should therefore be treated as a reported aggregate value pending recovery of the original questionnaires.

3.3. Task Design and Variable Definition

Three standardized tasks were selected to test complementary functional requirements. The precision task required participants to trace a 200-mm sinusoidal contour with a 1.5-mm target line and complete 10-mm-spaced hatching on A5 Arches 300-gsm cold-press paper using a standardized French Ultramarine preparation. Timing started at initial brush-to-paper contact and ended at the final stroke lift. Finished outputs were scanned at 600 dpi and registered against the reference path. Mean line-deviation error was calculated according to Eq. (2),

$$E_{line} = \left(\frac{1}{N}\right) \Sigma_{i=1}{}^{N} \|p_i - r_i\| \tag{2}$$

p_i is a sampled point on the extracted stroke centerline, r_i is the corresponding nearest point on the registered reference path, and N is the number of sampled centerline points. Stroke-width variability was calculated according to Eq. (3) from 50 equidistant orthogonal cross-sections.

$$SD_w = \sqrt{\left[\left(\frac{1}{K-1}\right) \Sigma_{j=1}{}^{K} (w_j - \bar{w})^2\right]} \tag{3}$$

w_j is the stroke width at the j-th cross-section, $\bar{w}$ is the mean stroke width, and K = 50. Image-derived measurements were implemented using procedures consistent with established ImageJ/Fiji-based quantitative image analysis [22], [23]. The wash task required a 100 × 150-mm graded wash on the same paper mounted at a 15° incline. The primary quality metric was wash-gradient root-mean-square error (RMSE), calculated using Eq. (4) from the difference between observed CIE L* values and the target linear luminance profile.

$$RMSE_{wash} = \sqrt{\left[\left(\frac{1}{K}\right) \Sigma_{k=1}{}^{K} (L*_k - L*target,_k)^2\right]} \tag{4}$$

L^*_k is the observed CIE L* value at sampling position k, $L^*target,_k$ is the corresponding target luminance, and K is the number of sampled positions. Lower RMSE represented better gradient fidelity. A derived smoothness index was calculated using Eq. (5), where 80 CIE L* units represented the predefined dynamic luminance range. Because smoothness is a deterministic transformation of RMSE, RMSE was treated as the primary wash-quality measure. The use of CIE L* is grounded in established CIELAB color-difference methodology and implementation guidance [24], [25].

$$S_{wash} = 100 \times \left(1 - \frac{RMSE_{wash}}{80}\right) \tag{5}$$

The mixed technical task was a standardized 90–105-s watercolor exercise combining contour work, a small wash region, and glazing. This task required both fine and broad manipulation without representing a full finished artwork. Across tasks, completion time was recorded in seconds, and paint consumption was calculated from the tared palette measurements using Eq. (6). Post-task responses assessed grip stability, perceived control, ease of task execution, hand fatigue, wrist discomfort, and overall satisfaction on five-point scales. These constructs were kept separate because perceived usability, comfort, and discomfort are related but non-equivalent aspects of product experience [2], [3], [4], [21]. All 720 outputs were assigned unique blind codes and subsequently scored for overall technical quality on a 1.0–10.0 scale by three independent master watercolor raters. The three standardized task protocols are summarized in figure 4.

$$M_{used} = M_{pre} - M_{post} \tag{6}$$

Mpre and Mpost are the tared palette masses immediately before and after the task, respectively.

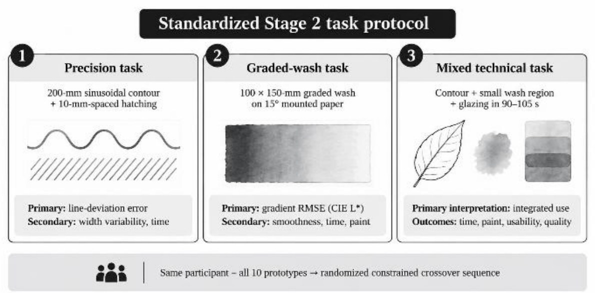

Figure 4. Standardized Stage 2 task protocol for precision line tracking, graded-wash execution, and the mixed technical watercolor exercise.

3.4. Experiment Execution and Data Collection

Stage 2 involved 24 participants whose demographic, professional, and hand-anthropometric characteristics are summarized in table 3. The sample included professional artists, advanced illustrators, and art educators, with varying levels of watercolor experience, dominant hand, painting frequency, preferred brush family, and hand dimensions. Four participants (16.7%) had also participated in the Stage 1 expert panel, while the remaining 20 participants (83.3%) were new to the study.

Every Stage 2 participant tested every prototype in every task, yielding a fully crossed design of 24 participants × 10 prototypes × 3 tasks = 720 primary trials. Trial order followed a randomized constrained crossover sequence rather than a fixed R1–R5/F1–F5 order. The same paper, pigment formulation, workspace geometry, timing rules, and measurement procedures were used across participants. Testing was performed at 23.0 ± 1.0 °C and 55 ± 5% relative humidity under D50 5000-K illumination calibrated to 1000 lux. Participants worked at a 74-cm drafting desk with a 15° easel mount and adjustable chair. Each physical output was assigned a unique output identifier and blind code. Three independent raters with 28, 22, and 19 years of watercolor experience evaluated the 720 outputs under calibrated D50 viewing conditions, producing 2,160 technical-quality ratings. The raters were not Stage 2 participants. The workbook preserved participant, brush, task, trial-order, output, and rating identifiers to support complete traceability between performance observations and blinded evaluation.

Table 3. Characteristics of Stage 2 participants (n=24).

Characteristic	Value
Professional status	13 professional artists (54.2%); 6 advanced illustrators (25.0%); 5 art educators (20.8%)
Watercolor experience	12.88±7.04 years; range 4–28
Dominant hand	21 right-handed (87.5%); 3 left-handed (12.5%)
Painting frequency	15 daily (62.5%); 9 three–four times/week (37.5%)
Preferred brush family	9 round (37.5%); 2 flat (8.3%); 13 both (54.2%)
Hand length	178.40±8.54 mm; range 165–192
Palm width	79.77±5.91 mm; range 71.5–89.0
Middle-finger length	79.25±4.21 mm; range 73–86
Preferred grip diameter	11.92±1.18 mm; range 10–14
Prior Stage 1 panel participation	4 yes (16.7%); 20 no (83.3%)

3.5. Data Analysis

Descriptive statistics were calculated for prototype properties, participant characteristics, Stage 1 expert ratings, task outcomes, and post-task usability responses. For task-specific family contrasts, each participant's five round-prototype observations and five flat-prototype observations were first averaged within task. The resulting 24 paired family values were compared using paired-samples t-tests, 95% confidence intervals based on the Student t distribution, and standardized paired effect size dz. Trial-level means and standard deviations are reported descriptively to show the full distribution of observations. The primary repeated-measures analysis modeled completion time using a linear mixed-effects model with fixed effects for brush family, task, family-by-task interaction, and trial order, together with a participant-specific random intercept, as expressed in Eq. (7).

$$Y_{ijkt} = \beta_0 + \beta_1 F_i + \beta_2 T_j + \beta_3 (F_i \times T_j) + \beta_4 O_{ijkt} + u_k + \varepsilon_{ijkt} \quad (7)$$

Y_{ijkt} is task completion time, F_i denotes prototype family, T_j denotes task, O_{ijkt} denotes trial order, u_k is the participant-specific random intercept, and ε_{ijkt} is the residual error. Mixed-effects modeling was selected because repeated observations from the same artist are not statistically independent and because the design retains trial-level information [26], [27], [28]. Precision line error and stroke-width variability were analyzed within the precision task, while wash-gradient RMSE was analyzed within the wash task. This task-specific strategy avoided treating metrics as interchangeable across contexts in which they did not have the same functional meaning. Prototype-level means were retained to identify within-family variation, but multivariable physical-feature modeling was considered exploratory because only ten unique prototype configurations were available.

For blinded output evaluation, inter-rater reliability was calculated using a two-way random-effects absolute-agreement model [29], [30], [31]. ICC(2,1) represents single-rater reliability, whereas ICC(2,3) represents the reliability of the mean of the three raters. Statistical significance was evaluated at $\alpha = 0.05$, but interpretation emphasized effect direction, confidence intervals, practical magnitude, and consistency across objective, subjective, and blinded evidence rather than p-values alone.

4. Results and Discussion

4.1. Prototype Characteristics and Stage 1 Expert Evaluation

The ten refined prototypes covered a deliberately broad functional range. Round configurations were longer and narrower, with total lengths from 248.0 to 264.0 mm, masses from 21.8 to 31.2 g, and minimum grip dimensions from 9.0 to 10.8 mm. Flat configurations were shorter but wider and heavier, with lengths from 198.0 to 216.0 mm, masses from 26.5 to 42.5 g, and minimum grip widths from 12.0 to 16.8 mm. Water-retention capacity ranged from 1.543 g for R1 to 10.400 g for F5, illustrating the substantially larger reservoir capacity of the broader flat configurations. These physical differences confirm that family comparisons represent integrated configuration contrasts rather than isolated handle-shape effects.

Stage 1 professional evaluation was favorable across all five dimensions. Material quality and craftsmanship received the highest mean score (4.73±0.46), while comfort and grip and water retention/color distribution each averaged 4.67. Balance/control and professional-use potential each averaged 4.60. The overall reported mean was 4.65/5.00. These results established strong professional acceptance but did not determine which configurations were most effective for specific tasks, motivating the controlled Stage 2 analysis.

4.2. Task-Specific Performance Results

Table 4 summarizes the principal family-level performance contrasts. In the precision task, the round family produced substantially lower mean line-deviation error than the flat family (0.442 vs. 1.243 mm; mean paired difference −0.801 mm, 95% CI −0.815 to −0.787; $p<0.0001$). Stroke-width variability was also lower for round prototypes (0.127 vs. 0.407 mm), and precision-task completion time was shorter (48.64 vs. 75.53 s). Overall usability during precision work was 4.76 for round prototypes versus 3.11 for flat prototypes. These results support H1 and are directionally consistent with the professional observations that motivated refinement of the round family.

The pattern reversed in the graded-wash task. Flat prototypes showed much lower gradient error than round prototypes (3.43 vs. 13.32 CIE L* units; mean paired difference 9.89, 95% CI 9.72–10.06; $p<0.0001$) and consequently higher derived smoothness (95.72% vs. 83.35%). Flat prototypes also completed the wash task more quickly (39.40 vs. 67.42 s) and used less paint (1.96 vs. 2.81 g). Wash-task usability averaged 4.87 for flat prototypes and 3.55 for round prototypes. Thus, H2 was supported under the standardized graded-wash condition.

Family differences were smaller in the mixed technical task. Round prototypes were faster by approximately 2.82 s (95.02 vs. 97.84 s; $p<0.0001$), used approximately 0.21 g less paint, and received higher overall usability scores (4.79 vs. 4.50). Although these differences remained statistically detectable, their scale was much smaller than the cross-

 ISSN 2723-6471

family separation observed in the specialized precision and wash tasks. This pattern supports H4 and suggests that the strongest family advantages emerge when task requirements align closely with the functional specialization of the prototype configuration.

Table 4. Task-specific performance summary by brush family. Descriptive mean±SD values use all 120 trials per family/task. Paired tests, confidence intervals, and dz are calculated from 24 participant-level family averages.

Task	Outcome	Round mean±SD	Flat mean±SD	Round−Flat	95% CI	t(23)	p	dz
Precision	Line error (mm)	0.442±0.053	1.243±0.117	−0.801	[−0.815, −0.787]	−117.90	<0.001	−24.07
Precision	Stroke-width variability (mm)	0.127±0.026	0.407±0.051	−0.280	[−0.287, −0.273]	−87.08	<0.001	−17.78
Precision	Time (s)	48.64±2.69	75.53±3.74	−26.89	[−27.41, −26.37]	−107.21	<0.001	−21.88
Precision	Paint used (g)	0.984±0.127	1.663±0.194	−0.679	[−0.700, −0.658]	−65.97	<0.001	−13.47
Precision	Overall usability	4.76±0.43	3.11±0.34	+1.65	[1.55, 1.75]	35.10	<0.001	7.16
Wash	Gradient RMSE (L*)	13.32±0.90	3.43±0.37	+9.89	[9.72, 10.06]	119.20	<0.001	24.33
Wash	Smoothness (%)	83.35±1.12	95.72±0.46	−12.37	[−12.58, −12.15]	−119.84	<0.001	−24.46
Wash	Time (s)	67.42±3.03	39.40±2.41	+28.02	[27.62, 28.41]	146.00	<0.001	29.80
Wash	Paint used (g)	2.805±0.191	1.964±0.170	+0.841	[0.817, 0.866]	71.70	<0.001	14.64
Wash	Overall usability	3.55±0.50	4.87±0.34	−1.32	[−1.46, −1.18]	−19.36	<0.001	−3.95
Mixed	Time (s)	95.02±3.72	97.84±4.09	−2.82	[−3.69, −1.95]	−6.72	<0.001	−1.37
Mixed	Paint used (g)	3.406±0.205	3.611±0.219	−0.205	[−0.241, −0.169]	−11.82	<0.001	−2.41
Mixed	Overall usability	4.79±0.41	4.50±0.50	+0.292	[0.178, 0.406]	5.29	<0.001	1.08

4.3. Prototype-Level Performance

Prototype-level performance results are presented in table 5, showing that the family-level findings were not driven by a single extreme design. Precision time and line error generally increased from R1 toward R5 as brush size and reservoir capacity increased, while the flat family showed a corresponding progression from F1 to F5. In contrast, all flat prototypes substantially outperformed all round prototypes in wash smoothness. R1 produced the fastest precision time (45.86 s) and the lowest precision error (0.381 mm), whereas F1 produced the fastest wash time (36.93 s), and F1–F2 achieved the highest wash-smoothness values. Differences among prototypes were comparatively modest in the mixed task. Figure 5 provides a visual summary of these prototype-level performance patterns after rescaling each metric to a common higher-is-better direction.

Table 5. Prototype-level performance matrix.

Brush	Family	Precision time (s)	Precision error (mm)	Wash time (s)	Wash smoothness (%)	Mixed time (s)	Mixed paint (g)	Mixed usability
R1	Round	45.86	0.381	64.96	82.40	92.54	3.186	4.79
R2	Round	48.10	0.409	65.73	82.93	93.34	3.300	4.92
R3	Round	48.60	0.445	67.38	83.49	94.29	3.419	4.71
R4	Round	49.69	0.474	68.65	83.73	97.43	3.534	4.79
R5	Round	50.93	0.498	70.38	84.22	97.48	3.592	4.75
F1	Flat	71.95	1.109	36.93	96.13	95.17	3.364	4.50
F2	Flat	73.15	1.172	37.92	96.03	96.53	3.466	4.46
F3	Flat	75.78	1.252	39.50	95.78	96.78	3.608	4.46
F4	Flat	77.47	1.327	40.54	95.48	99.43	3.745	4.54
F5	Flat	79.30	1.355	42.12	95.19	101.29	3.873	4.54

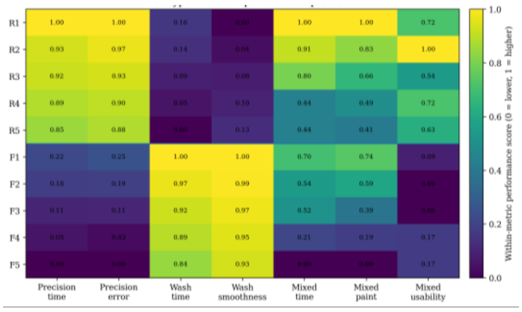

Figure 5. Prototype-level task-performance profile. Values are rescaled within each metric so that 0 indicates the lower and 1 the higher task-specific performance score; this normalization is for visualization only.

4.4. Subjective Ergonomic and Usability Results

The post-task ratings closely followed the objective performance pattern for control-related outcomes (table 6). During precision work, round prototypes received higher ratings for grip stability (4.80 vs. 3.38), perceived precision control (4.88 vs. 3.11), ease of execution (4.76 vs. 3.11), and overall satisfaction (4.93 vs. 3.08). During the wash task, the direction reversed: flat prototypes received higher grip stability (4.93 vs. 3.92), perceived control (4.84 vs. 3.47), ease (4.87 vs. 3.55), and satisfaction (4.98 vs. 3.78). In the mixed task, the round family retained smaller advantages in control, ease, and satisfaction.

Importantly, the family differences did not extend to short-duration fatigue and wrist discomfort. Precision-task hand fatigue did not differ significantly between families (1.84 vs. 1.78, $p=0.236$), nor did wash-task fatigue (1.85 vs. 1.84, $p=0.862$). Wrist discomfort was likewise non-significant in precision ($p=0.607$), wash ($p=0.117$), and mixed conditions. This divergence is informative: the configurations altered task control and user satisfaction much more strongly than they altered short-term physical discomfort. Consequently, the present data support claims of task-specific controllability and usability, but not a broad claim that one family reduces fatigue.

Table 6. Post-task subjective evaluation by prototype family.

Dimension	Task	Round	Flat	t(23)	p
Grip stability	Precision	4.80	3.38	27.66	<0.001
Grip stability	Wash	3.92	4.93	−24.72	<0.001
Grip stability	Mixed	4.72	4.59	2.61	0.016
Precision control	Precision	4.88	3.11	34.45	<0.001
Precision control	Wash	3.47	4.84	−32.55	<0.001
Precision control	Mixed	4.76	4.40	6.21	<0.001
Ease of task	Precision	4.76	3.11	35.10	<0.001
Ease of task	Wash	3.55	4.87	−19.36	<0.001
Ease of task	Mixed	4.79	4.50	5.29	<0.001
Hand fatigue	Precision	1.84	1.78	1.22	0.236
Hand fatigue	Wash	1.85	1.84	0.18	0.862
Hand fatigue	Mixed	1.87	1.81	1.00	0.328
Wrist discomfort	Precision	1.51	1.47	0.52	0.607
Wrist discomfort	Wash	1.55	1.45	1.63	0.117
Wrist discomfort	Mixed	1.52	1.49	0.34	0.734
Overall satisfaction	Precision	4.93	3.08	53.50	<0.001
Overall satisfaction	Wash	3.78	4.98	−31.01	<0.001
Overall satisfaction	Mixed	4.88	4.53	5.90	<0.001

4.5. Blinded Technical Output Quality

The blinded expert evaluation results are presented in table 7 and independently support the task-specific family pattern observed in the performance metrics. Precision outputs produced with round prototypes received a pooled technical-quality score of 7.80 ± 0.26, compared with 4.94 ± 0.48 for flat prototypes. In contrast, wash outputs produced with flat prototypes received a pooled score of 8.70 ± 0.21, compared with 4.94 ± 0.39 for round prototypes.

Table 7. Blinded technical output quality and inter-rater reliability.

Task	Family	Rater 1	Rater 2	Rater 3	Pooled mean±SD
Precision	Round	7.81	7.79	7.82	7.80±0.26
Precision	Flat	4.97	4.91	4.95	4.94±0.48
Wash	Round	4.92	4.93	4.97	4.94±0.39
Wash	Flat	8.72	8.68	8.70	8.70±0.21
Mixed	Round	8.77	8.79	8.79	8.78±0.19
Mixed	Flat	8.49	8.49	8.52	8.50±0.20
Overall	All outputs	7.28	7.26	7.29	7.28±1.71

Mixed-task quality was high for both families, with a smaller difference favoring round prototypes (8.78 ± 0.19 vs. 8.50 ± 0.20). Inter-rater reliability was excellent, with ICC(2,1) = 0.969 for a single rater and ICC(2,3) = 0.990 for the mean of three raters. Thus, objective task metrics, user evaluations, and blinded technical judgments converged on the same broad conclusion: the performance advantage of each brush family depends on task context. Figure 6 visualizes the blinded technical-quality ratings across the three tasks and prototype families.

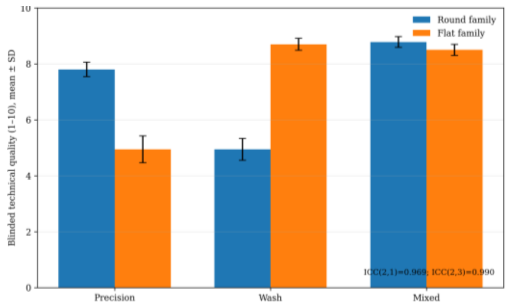

Figure 6. Blinded technical-quality ratings by task and prototype family (mean ± SD). Inter-rater reliability across the 720-output matrix was ICC(2,1)=0.969 and ICC(2,3)=0.990.

4.6. Repeated-Measures Model and Hypothesis Evaluation

The linear mixed-effects model results are presented in table 8. The model included all 720 completion-time observations with participant-specific random intercepts and used Round/Precision as the reference condition. Relative to this reference, flat configurations were 26.90 s slower in the precision task (β = 26.898, 95% CI [26.065, 27.732], $p < 0.001$). For round configurations, completion time was 18.78 s longer in the wash task and 46.39 s longer in the mixed task than in the precision task. The interaction terms were strongly negative for Flat × Wash (β = −54.901) and Flat × Mixed (β = −24.084), indicating that the relative effect of prototype family changed substantially across tasks. Trial order was not significant (β = 0.009 s per trial, 95% CI [−0.019, 0.037], p = 0.512), providing no evidence of a systematic completion-time trend across the 30-trial sequence.

Table 8. Linear mixed-effects model for task completion time (n=720 trials).

Model parameter	Estimate	SE	95% CI	z	p
Intercept: Round, Precision	48.488	0.393	[47.719, 49.258]	123.49	<0.001
Family: Flat	26.898	0.425	[26.065, 27.732]	63.25	<0.001
Task: Wash	18.781	0.425	[17.947, 19.614]	44.17	<0.001
Task: Mixed	46.385	0.425	[45.552, 47.219]	109.07	<0.001
Flat × Wash	−54.901	0.601	[−56.080, −53.722]	−91.29	<0.001
Flat × Mixed	−24.084	0.602	[−25.263, −22.904]	−40.03	<0.001
Trial order	0.009	0.014	[−0.019, 0.037]	0.66	0.512

The participant random-intercept variance was 0.304, compared with a residual variance of 10.846, indicating relatively limited between-participant variation within the standardized task structure. Figure 7 provides the corresponding descriptive family-by-task interaction pattern, with the crossover between precision and wash performance visually reflecting the family × task interaction identified by the mixed-effects model.

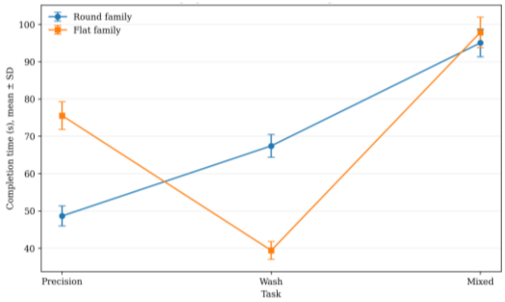

Figure 7. Observed family-by-task interaction in completion time (mean ± SD). The crossover between Precision and Wash visually reflects the family×task interaction confirmed by the mixed-effects model.

4.7. Discussion

The principal finding is not that one brush family is universally superior, but that the refined configurations exhibit complementary specialization. Round prototypes showed a pronounced advantage in precision tracking, stroke-width stability, speed, perceived control, and blinded technical quality. Flat prototypes showed an equally pronounced advantage in graded-wash fidelity, broad-task speed, material consumption, perceived wash control, and blinded wash quality. When task requirements were mixed, the separation narrowed considerably. This pattern directly supports a task-specific interpretation of professional artistic-tool performance and extends the earlier aggregate innovative-versus-conventional comparison [5].

The precision advantage of the round family is plausible in light of the integrated configuration of those prototypes. The round brushes were narrower, generally lighter, and equipped with pointed or tapered bristle profiles. Hand-tool research shows that handle diameter, hand-size compatibility, shape, and force distribution can materially alter grip mechanics and control [9]–[18], so the observed precision advantage is consistent with a broader human–tool interaction framework. However, the experiment was not designed to isolate the causal contribution of any single characteristic. The round family also differed from the flat family in overall length, center of gravity, head geometry, bristle composition, stiffness, and reservoir capacity. Therefore, the correct inference is that the tested round configurations were more effective for the standardized precision task, not that round handle cross-section alone caused superior precision.

The flat-family wash advantage is similarly consistent with its broader heads and greater reservoir capacity. Water retention increased markedly across the wider flat prototypes, reaching 10.4 g in F5. The flat configurations also reduced graded-wash error and completion time while using less paint under the standardized wash protocol. These results suggest that broad contact area, reservoir capacity, bristle arrangement, and handle stability act together to support controlled fluid transfer. The finding advances the earlier first-generation comparison [5] by showing that performance is not merely an aggregate "innovative brush" effect: within the later refined generation, different configurations are optimized for different task demands.

The subjective results provide an important qualification to the ergonomic interpretation. Participants strongly preferred the task-appropriate family for grip, control, ease, and satisfaction, yet hand fatigue and wrist discomfort did not differ significantly. This divergence is consistent with hand-tool and product-experience research showing that functionality, comfort, discomfort, and objective performance are related but distinct constructs [2], [3], [4], [19], [20], [21]. The tasks in this study were short and highly standardized; they were appropriate for detecting control and execution differences but may have been too brief to generate large cumulative fatigue effects. Accordingly, the study supports task-specific usability advantages but does not establish long-duration fatigue reduction.

A further strength is convergence across three evidence layers. Stage 1 professional artists evaluated the refined prototypes favorably. Stage 2 objective measurements demonstrated strong task-specific family differences. Independent blinded experts then rated the resulting outputs in the same direction, with excellent agreement. This convergence reduces reliance on any single measurement type and aligns with the broader movement toward data-driven design, where heterogeneous product and user data are integrated to support design decisions [6], [7]. It also demonstrates how quantitative image analysis and color measurement can contribute to creative-tool research: stroke geometry and wash luminance become analyzable variables rather than purely descriptive impressions [8], [22], [23], [24], [25].

The very large standardized paired effects observed in the specialized tasks should nevertheless be interpreted in the context of the experimental design. Family values were averaged over five prototypes within each participant before paired contrasts were computed, and the tasks were deliberately constructed to emphasize precision versus graded-wash requirements. The magnitude of these standardized effects therefore describes separation under the tested standardized conditions; it should not be generalized directly to all watercolor practice. The mixed-task results provide a more conservative view of practical overlap between the two families.

Several limitations should be acknowledged. First, the study includes only ten prototype configurations, so broader claims about all round and flat watercolor brushes are not warranted. Second, physical properties covary across family; a factorial prototype experiment would be required to isolate handle geometry from bristle, mass, head size, and center-of-gravity effects. Third, participants were experienced artists and art professionals, limiting direct generalization to novice users. Fourth, the standardized tasks improve measurement precision but do not reproduce the duration and complexity of complete watercolor artworks. Fifth, the Stage 1 results are available as aggregate descriptive statistics, preventing participant-level longitudinal analysis between the two stages. Finally, durability and long-term biomechanical outcomes were outside the scope of the current refined-generation evaluation.

Future work should therefore test the refined designs in longer painting sessions, include novice and intermediate artists, evaluate durability and manufacturing repeatability, and develop factorial prototypes that vary one design parameter at a time. Larger prototype sets would also permit more reliable modeling of relationships between physical variables—such as reservoir capacity, mass, center of gravity, and stiffness—and task outcomes. Such work could extend the present data-driven framework from comparative evaluation toward predictive design support.

5. Conclusion

This study integrated professional expert validation with a controlled repeated-measures experiment to evaluate ten refined watercolor brush prototypes. Stage 1 showed high professional acceptance of the refined set, with an overall expert evaluation mean of 4.65/5.00. Stage 2 demonstrated that performance was strongly task-dependent: round prototype configurations produced lower precision error, more stable stroke widths, faster execution, and higher precision-task usability, whereas flat configurations produced more accurate and smoother graded washes, faster wash execution, lower material consumption, and higher wash-task usability. Differences were substantially smaller in the mixed technical task. Subjective control and satisfaction aligned with objective and blinded-output results, while fatigue and wrist discomfort did not show meaningful family differences. The primary repeated-measures model confirmed a strong family-by-task interaction, supporting a complementary-specialization interpretation rather than a universal hierarchy between brush families. These findings demonstrate how multimetric data collection, repeated-measures analysis, and independent output assessment can provide an evidence-based foundation for the design and refinement of professional artistic tools.

6. Declarations

6.1. Author Contributions

Conceptualization: S.C.; Methodology: S.C.; Validation: S.C.; Formal Analysis: S.C.; Investigation: S.C.; Resources: S.C.; Data Curation: S.C.; Writing—Original Draft Preparation: S.C.; Writing—Review and Editing: S.C.; Visualization: S.C. The author has read and agreed to the submitted version of the manuscript.

6.2. Data Availability Statement

The data from this study are available upon request from the corresponding author.

6.3. Funding

Prototype development and refinement were supported by the National Science, Research and Innovation Fund (Thailand Science Research and Innovation: TSRI) through Rajamangala University of Technology Thanyaburi (Grant Nos. FRB68E0801 and FRB680045/0168).

6.4. Institutional Review Board Statement

The Stage 2 human-participant protocol was conducted under institutional ethics protocol IRB-RMUTT-2026-084. The Stage 1 component used previously collected aggregate, de-identified expert-evaluation results and did not involve participant-level reanalysis in the present manuscript.

6.5. Informed Consent Statement

Informed consent was obtained from all Stage 2 participants before data collection.

6.6. Declaration of Competing Interest

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.

References

[1] W. Karwowski, "Ergonomics and human factors: The paradigms for science, engineering, design, technology and management of human-compatible systems," *Ergonomics*, vol. 48, no. 5, pp. 436–463, May 2005, doi: 10.1080/00140130400029167.

[2] L. F. M. Kuijt-Evers, L. Groenesteijn, M. P. de Looze, and P. Vink, "Identifying factors of comfort in using hand tools," *Appl. Ergon.*, vol. 35, no. 5, pp. 453–458, Sep. 2004, doi: 10.1016/j.apergo.2004.04.001.

[3] L. F. M. Kuijt-Evers, T. Bosch, M. A. Huysmans, M. P. de Looze, and P. Vink, "Association between objective and subjective measurements of comfort and discomfort in hand tools," *Appl. Ergon.*, vol. 38, no. 5, pp. 643–654, Sep. 2007, doi: 10.1016/j.apergo.2006.05.004.

[4] P. Vink and S. Hallbeck, "Comfort and discomfort studies demonstrate the need for a new model," *Appl. Ergon.*, vol. 43, no. 2, pp. 271–276, Mar. 2012, doi: 10.1016/j.apergo.2011.06.001.

[5] S. Chantanasut, "Quantitative evaluation of watercolor brush performance: A comparative study of user satisfaction and task efficiency using 24 innovative brush designs," *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 2131–2142, 2024, doi: 10.47738/jads.v5i4.462.

[6] A. Liu, S. Lu, F. Tao, and N. Anwer, "Integration of data science with product design towards data-driven design," *CIRP Ann.*, vol. 73, no. 2, pp. 509–532, 2024, doi: 10.1016/j.cirp.2024.06.003.

[7] J. C. Quiñones-Gómez, E. Mor, and J. Chacón, "Data-driven design in the design process: A systematic literature review on challenges and opportunities," *Int. J. Hum.-Comput. Interact.*, vol. 41, no. 4, pp. 2227–2252, 2025, doi: 10.1080/10447318.2024.2318060.

[8] J. Li, C. Chun, K. Brown, and J. Z. Wang, "Computational investigation of abstraction in Claude Monet's *Water Lilies* through brushstroke analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 6, pp. 7129–7146, 2026, doi: 10.1109/TPAMI.2026.3664028.

[9] Y. K. Kong and B. D. Lowe, "Optimal cylindrical handle diameter for grip force tasks," *Int. J. Ind. Ergon.*, vol. 35, no. 6, pp. 495–507, Jun. 2005, doi: 10.1016/j.ergon.2004.11.003.

[10] Y. K. Kong and B. D. Lowe, "Evaluation of handle diameters and orientations in a maximum torque task," *Int. J. Ind. Ergon.*, vol. 35, no. 12, pp. 1073–1084, Dec. 2005, doi: 10.1016/j.ergon.2005.04.009.

[11] N. J. Seo and T. J. Armstrong, "Investigation of grip force, normal force, contact area, hand size, and handle size for cylindrical handles," *Hum. Factors*, vol. 50, no. 5, pp. 734–744, Oct. 2008, doi: 10.1518/001872008X354192.

[12] N. J. Seo and T. J. Armstrong, "Effect of elliptic handle shape on grasping strategies, grip force distribution, and twisting ability," *Ergonomics*, vol. 54, no. 10, pp. 961–970, Oct. 2011, doi: 10.1080/00140139.2011.606923.

[13] J. Rossi, E. Berton, L. Grélot, C. Barla, and L. Vigouroux, "Characterisation of forces exerted by the entire hand during the power grip: Effect of the handle diameter," *Ergonomics*, vol. 55, no. 6, pp. 682–692, Jun. 2012, doi: 10.1080/00140139.2011.652195.

[14] C. Y. Wang and D. C. Cai, "Hand tool handle size and shape determination based on hand measurements using a contour gauge," *Hum. Factors Ergon. Manuf. Serv. Ind.*, vol. 30, no. 5, pp. 349–364, Sep. 2020, doi: 10.1002/hfm.20846.

[15] K. A. Grant, D. J. Habes, and L. L. Steward, "An analysis of handle designs for reducing manual effort: The influence of grip diameter," *Int. J. Ind. Ergon.*, vol. 10, no. 3, pp. 199–206, 1992, doi: 10.1016/0169-8141(92)90033-V.

[16] J. L. Sancho-Bru, D. J. Giurintano, A. Pérez-González, and M. Vergara, "Optimum tool handle diameter for a cylinder grip," *J. Hand Ther.*, vol. 16, no. 4, pp. 337–342, Dec. 2003, doi: 10.1197/S0894-1130(03)00160-1.

[17] A. Maleki-Ghahfarokhi, M.-R. Azghani, and I. Dianat, "Effects of handle characteristics of manual hand tools on maximal torque exertions: A literature review," *Int. J. Occup. Saf. Ergon.*, vol. 28, no. 3, pp. 1387–1402, 2022, doi: 10.1080/10803548.2021.1896139.

[18] R. G. García-Cáceres, S. Felknor, J. E. Córdoba, J. P. Caballero, and L. H. Barrero, "Hand anthropometry of the Colombian floriculture workers of the Bogota plateau," *Int. J. Ind. Ergon.*, vol. 42, no. 2, pp. 183–198, Mar. 2012, doi: 10.1016/j.ergon.2011.12.002.

[19] N. Crilly, J. Moultrie, and P. J. Clarkson, "Seeing things: Consumer response to the visual domain in product design," *Des. Stud.*, vol. 25, no. 6, pp. 547–577, Nov. 2004, doi: 10.1016/j.destud.2004.03.001.

[20] M. Hassenzahl and N. Tractinsky, "User experience—A research agenda," *Behav. Inf. Technol.*, vol. 25, no. 2, pp. 91–97, Mar. 2006, doi: 10.1080/01449290500330331.

[21] J. R. Lewis, "The System Usability Scale: Past, present, and future," *Int. J. Hum.-Comput. Interact.*, vol. 34, no. 7, pp. 577–590, 2018, doi: 10.1080/10447318.2018.1455307.

[22] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nat. Methods*, vol. 9, no. 7, pp. 671–675, Jul. 2012, doi: 10.1038/nmeth.2089.

[23] J. Schindelin *et al.*, "Fiji: An open-source platform for biological-image analysis," *Nat. Methods*, vol. 9, no. 7, pp. 676–682, Jul. 2012, doi: 10.1038/nmeth.2019.

[24] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Res. Appl.*, vol. 26, no. 5, pp. 340–350, Oct. 2001, doi: 10.1002/col.1049.

[25] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, Feb. 2005, doi: 10.1002/col.20070.

[26] R. Gueorguieva and J. H. Krystal, "Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the *Archives of General Psychiatry*," *Arch. Gen. Psychiatry*, vol. 61, no. 3, pp. 310–317, Mar. 2004, doi: 10.1001/archpsyc.61.3.310.

[27] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang.*, vol. 68, no. 3, pp. 255–278, Apr. 2013, doi: 10.1016/j.jml.2012.11.001.

[28] D. Bates, M. Mächler, M. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[29] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, Mar. 1979, doi: 10.1037/0033-2909.86.2.420.

[30] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychol. Methods*, vol. 1, no. 1, pp. 30–46, Mar. 1996, doi: 10.1037/1082-989X.1.1.30.

[31] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2016.02.012.