Supplementary Material 1

A Dimension definitions, boundaries and examples

Definitions and comparison criteria follow Section 4. Examples are drawn from the coded corpus; study IDs refer to S2.

Table 1: Dimension definitions and boundaries.

Dimension	Definition	Dimension boundary
Stimuli	The object presented to participants for task execution.	**Stimuli** are the materials that participants perceive or manipulate, rather than their responses (**Data**), the activity performed (**Task**), or the delivery system (**Interface**). Values shown in a chart are stimulus content; answers collected about that chart are evidence.
Task	The activity that participants are asked to execute using the stimuli.	**Task** concerns what the participant does. Instructions, timing and trial organisation belong to **Procedure**; the displayed object belongs to **Stimuli**. A changed activity remains in the replication core when it tests or qualifies the reference-linked question.
Procedure	The operational protocol through which the study is conducted.	**Procedure** covers practice, instructions, timing, ordering, allocation and study structure. It differs from the participant's activity (**Task**), input-output system (**Interface**), and surrounding setting (**Environment**). Prompting, training and repeated inference are procedural choices when models take the participant role.
Interface	The system through which participants access the study.	**Interface** covers the device, software and input-output mechanisms. The materials presented are **Stimuli**; the rules governing a session are **Procedure**. A machine participant's image/prompt input and generated output channel are its **Interface**, not the machine itself.
Environment	The surrounding setting where the study occurs.	**Environment** concerns the experimental setting and its surrounding conditions, not geographical location. It includes computational execution or deployment settings when these are where the experiment occurs. The actor is **Participant**, the access channel is **Interface**, and the operational protocol is **Procedure**.
Data	The evidence source on which the study claims are ultimately based, such as answers, logs, and response time.	**Data** are collected or reused observations, not the stimulus dataset or the methods that turn observations into claims (**Analysis**). Derived bias, precision or model-fit scores do not by themselves establish a new evidence form. Additional evidence for a separable question belongs to the addition fields.
Participant	The intended population and sampling approach, including recruitment and eligibility requirements.	**Participant** is the actor performing the focal study activity, rather than the research team. A model is **Participant** when it performs the reference participant's role. Population judgements use the reported broad target population and sampling characteristics; unreported demographics are not reconstructed.
Analysis	The methods used to transform the collected evidence into claims.	**Analysis** includes transformations, statistical models, qualitative interpretation and inferential comparisons. The observations are **Data**. **Analysis** remains in the replication core when it explains or qualifies the reference-linked finding, even if the method is new.

Table 2: Comparison criteria and corpus examples.

Dimension / level	Criterion	Corpus examples
Stimuli ***identical***	Exactly the same stimuli are used; only unavoidable rendering or implementation differences exist.	**Study 4:** The reference visual-metaphor charts are retained for the remote replication.
		Study 50.1: Published case-study figures are reused. Minor image-extraction differences are treated as implementation differences.
Stimuli ***similar***	The stimuli differ but serve a comparable perceptual, semantic, or experimental role.	**Study 3:** Standalone rectangles and a wider aspect-ratio range retain the area-judgement role.
		Study 44: Astronomy figures are reconstructed with matched materials for the same scholarly-figure comparison.
Stimuli ***different***	The stimuli differ substantially in representation, form, or material.	**Study 31:** Materials used with household smart artefacts are replaced by crafting and power-tool materials.
		Study 19: Two-dimensional graphical marks are replaced by physical bars and spheres.
Task ***identical***	Participants perform exactly the same task.	**Study 2.1:** Participants retain the proportional-judgement activity used in the reference experiment.
		Study 38: The same nine information-extraction tasks are used.

Continued on next page

Table 2 (continued)

DIMENSION / level	**Criterion**	**Corpus examples**
TASK ***similar***	Wording, difficulty, or operational details differ, but the task targets a comparable activity.	**Study 16.1:** Selecting the most-connected node becomes estimating its number of neighbours, retaining node-degree assessment.
		Study 21.1: Graphical comparison is reframed as a percentage-of-whole estimate.
TASK ***different***	The tasks are not meaningfully comparable.	**Study 35:** Visible ratio estimation becomes reproduction of values from memory.
		Study 26.2: Target identification under perceptual load becomes a same/different judgement of shape sets.
PROCEDURE ***identical***	The same study protocol and experimental process are followed.	**Study 4:** The paired visual-metaphor protocol is retained despite moving the study online.
		Study 5: The control replication retains the reference graphical-perception protocol.
PROCEDURE ***similar***	Some protocol elements change, but the core experimental structure remains comparable.	**Study 3:** The MTurk proportion-judgement structure is retained, with revised factorial trials and administration.
		Study 44: The core evaluation protocol is retained, with changes to study organisation and timing.
PROCEDURE ***different***	The experimental processes are substantially different.	**Study 28.2:** Human experimental trials are replaced by supervised network training and testing.
		Study 50.1: Interactive human exploration becomes structured, multi-model conversational prompting.
INTERFACE ***identical***	The same device type, software, and input-output mechanisms are used.	**Study 4:** The same Java-based visual-metaphor interface is retained.
		Study 30: The side-by-side scatterplot display and binary-selection interaction are retained.
INTERFACE ***similar***	The interface changes, but the interaction logic remains comparable.	**Study 34:** A custom web application retains the view-and-allocate interaction.
		Study 40: Printed charts and written answers become browser displays and typed answers, preserving the response logic.
INTERFACE ***different***	The interface changes the nature of interaction.	**Study 43.1:** Desktop keyboard/mouse access becomes embodied interaction in virtual reality.
		Study 50.1: An interactive visualisation system is replaced by static images in LLM chat interfaces.
ENVIRONMENT ***identical***	The setting and constraints are controlled to be as similar as possible.	**Study 44:** Both studies use an AAS conference exhibit hall with a nearly identical physical set-up. The change of convention centre does not itself change the code.
		Study 19: Both studies use controlled laboratory settings.
ENVIRONMENT ***similar***	The same setting type is retained, but constraints differ.	**Study 9.1:** The computer-based study retains its broad setting but uses participants' own computers rather than fixed experimental equipment.
		Study 31: Broad setting types are retained, while the crafting and power-tool situations change the surrounding constraints.
ENVIRONMENT ***different***	The setting is of a different type.	**Study 4:** A supervised student-laboratory study becomes remote MTurk participation.
		Study 50.1: A human case-study setting becomes local or cloud-based model inference.
DATA ***identical***	**DATA** form and values are the same.	**Study 1:** The Springboard logs are reused; one unavailable participant log is treated as incidental, non-material loss. This is the corpus's only **DATA**-identical case.
DATA ***similar***	**DATA** form is the same, but the values differ.	**Study 2.1:** New proportion judgements have the same evidence form as the reference responses.
		Study 17: Shared observations are reused, but previously excluded observations and conditions are deliberately restored.
DATA ***different***	The collected evidence is in a different form.	**Study 35:** Ratio estimates are replaced by records of values reconstructed from memory; the distinction concerns recorded responses, not derived bias or precision.
		Study 51.2: Qualitative interviews and body/landscape maps replace the reference quantitative body and gait evidence.

Continued on next page

Table 2 (continued)

DIMENSION / level	Criterion	Corpus examples
PARTICIPANT *identical*	The population and demographics are the same.	**Study 5:** Both studies target the same broad MTurk population.
		Study 30: The same intended MTurk population and broadly aligned reported demographics are retained.
PARTICIPANT *similar*	The population is the same, but demographics differ.	**Study 44:** The astronomy community remains the target population, with different realised demographics.
		Study 46.1: Online adults remain the target population, but recruitment moves from MTurk to Prolific with different filters.
PARTICIPANT *different*	The intention is to use a different population.	**Study 37:** Children are a focal target population rather than only the adults studied previously.
		Study 49.1: Human graph readers are replaced by multimodal LLMs.
ANALYSIS *identical*	The same analysis approach is used.	**Study 2.1:** The core graphical-perception error analysis is retained.
		Study 45.1: Ward clustering and the reference comparison metrics are retained.
ANALYSIS *similar*	The data are treated differently using a comparable approach.	**Study 3:** Factorial ANOVA and post-hoc comparisons extend the proportion-error analysis while retaining comparable factor-effect inference.
		Study 26.2: Factors change, but response-time and error analyses still use comparable factor-effect inference.
ANALYSIS *different*	The analysis approach is substantively different.	**Study 17:** Condition-mean Weber models become individual-observation, censored Bayesian models and uncertainty-aware partial rankings.
		Study 40: Aggregate graphical-perception rankings are reassessed through substantive modelling of individual differences.

B OPEN-CODE CODEBOOKS

Motivation, design changes and design additions are unified separately. Counts give the number of replication studies assigned to each phrase; each phrase is counted at most once per study. A study may have multiple change or addition phrases. The **EXPERIMENT** group combines **STIMULI**, **TASK**, **PROCEDURE**, **INTERFACE** and **ENVIRONMENT** in the open-code summaries.

B.1 Motivation

Table 3: Motivation codebook.

Unified phrase	Count	Explanation	Corpus example
Explain effects or conflicting findings	21	Investigate mechanisms, moderators, strategies, or alternative explanations for an established effect or disagreement.	**Study 14.1:** The study isolates separation and distractor contributions to adjacent-versus-separated bar error.
Test generalisability across designs or settings	20	Test whether findings or model predictions extend to other materials, tasks, representations, conditions, or settings.	**Study 19:** The study tests transfer from two-dimensional graphical marks to physical bars and spheres.
Reassess the robustness of a reported finding	12	Recheck an established effect, ranking, or model prediction, including a closer, better-powered, or deliberately varied replication.	**Study 39.2:** A close synchronous remote replication reassesses the chart-description and recall findings.
Evaluate a research method or tool	10	Assess a platform, recruitment method, or tool used to conduct or analyse studies.	**Study 46.1:** The adaptive-staircase replication tests reVISit's dynamic-sequencing capability.
Assess computational observers against established findings	10	Assess whether algorithms or models acting as observers reproduce established human or computational findings.	**Study 49.1:** Multimodal LLMs are tested against human graph-aesthetic readability effects.
Examine generalisability across human populations	7	Examine whether a finding applies to a different human population, including age, disability, or recruitment-defined populations.	**Study 37:** Adult graphical-perception findings are examined with children aged 8–12.
Evaluate an intervention using a replicated baseline	4	Establish or use a reference-linked baseline to evaluate a visualisation, interaction, accessibility, or behavioural intervention.	**Study 23:** A replicated CO2-exploration baseline supports evaluation of interaction-history encoding.
Develop or refine a model of performance	2	Use the replication relationship to develop a more general or individualised model of perception or task performance.	**Study 9.1:** The study develops and validates a more general model of slope-ratio judgement error.

B.2 Design changes

Table 4: Design changes codebook.

Unified phrase	Count	Explanation	Corpus example
EXPERIMENT			
Change the study interface	72	Change the presentation/response system, interaction channel, software implementation, input controls, or access medium.	**Study 38:** Conventional screen-reader access is replaced by VoxLens voice, summary, sonification, and keyboard modes.
Adjust the study administration protocol	70	Adjust trial counts and ordering, practice, instructions, timing, prompting, response support, checks, or computational run details within a study structure.	**Study 40:** Repetitions, scheduled breaks, and attention checks are grouped as protocol adjustments, not three separate design-choice categories.
Reconstruct or adapt stimulus materials	47	Recreate or adapt stimulus instances, source datasets, values, layout details, sizes, wording, or content.	**Study 44:** Matched 3D astronomy figures are rebuilt from real astronomy data.
Change the study setting	46	Change the surrounding physical, remote, computational, or deployment setting, or consequential contextual constraints within a setting type.	**Study 51.1:** A quiet laboratory corridor is replaced by an outdoor university-campus path.
Restrict stimulus or condition coverage	33	Select a subset of reference materials, tasks within a battery, parameter ranges, or conditions.	**Study 46.1:** The replication retains scatterplots and parallel coordinates but omits seven other chart forms.
Change the task or response requirement	33	Change the activity, requested judgement, response form, or constituent questions participants must complete.	**Study 35:** Ratio estimation is replaced by immediate memory reproduction of every displayed value.
Change the experimental structure or paradigm	32	Change how conditions, phases, or comparison groups are organised, such as within- versus between-participant allocation, concurrent versus historical baselines, or human sessions versus computational experiments.	**Study 38:** The no-VoxLens comparison uses the published sample as a historical baseline rather than recruiting a concurrent control.
Expand stimulus or condition coverage	26	Add stimulus forms, parameter ranges, factor levels, or experimental contrasts within the replication relationship.	**Study 3:** The rectangle experiment expands aspect-ratio coverage and adds orientation.
Change stimulus representation or modality	19	Replace the encoding family, material form, or sensory modality through which the task object is represented.	**Study 47:** Visual graphical marks are translated into raised swell-paper tactile graphics.
DATA			
Collect new observations	84	Collect or generate new human or model observations for the reference-linked comparison.	**Study 30:** The study collects 19,000 new correlation judgements instead of reusing the reference observations.
Revise the evidence measures collected	38	Add, remove, replace, or redefine particular recorded measures while retaining a broadly comparable kind of primary evidence.	**Study 49.1:** Correctness remains the primary evidence, while output-token count replaces human response time as a secondary analogue.
Reuse existing evidence	5	Reuse reference observations, labels, or benchmark evidence, wholly or after selecting or restoring observations, alone or alongside new evidence.	**Study 17:** Shared raw observations are reused with previously excluded outliers and conditions restored; these details remain visible in the original **DATA** field.
Replace the primary evidence type	3	Replace the main kind of evidence used for the comparison, such as quantitative observations with qualitative accounts, ratio estimates with reconstructed-value records, or illustrative evidence with measured model outputs.	**Study 51.2:** Interview transcripts, route annotations, and body/landscape maps replace the reference quantitative body, gait, and physiological observations.
PARTICIPANT			
Adjust sampling and recruitment	45	Adjust recruitment route, incentives, eligibility or exclusion filters, sample size, or realised composition within a broad target population.	**Study 46.2:** Expert recruitment changes from network email and design-community Slack to social media; other eligible instances include altered filters or sample size.

Continued on next page

Table 4 (continued)

Unified phrase	Count	Explanation	Corpus example
Change the target human population	34	Replace or extend the intended human population, including lab-to-crowd population shifts, age groups, or disability groups.	**Study 33:** The focal population includes people with intellectual disabilities or autism alongside a non-disabled control.
Replace the participant agent	10	Replace human observers with computational models, or replace one focal computational observer family with another.	**Study 45.1:** Pretrained CNN similarity metrics replace the focal tuned MS-SSIM computational agent.
ANALYSIS			
Change statistical tests or models	57	Replace or expand statistical tests, regressions, psychophysical fits, or inferential model families.	**Study 35:** Aggregate ranking analyses are replaced by a Bayesian multilevel censored mixture model.
Revise uncertainty or robustness assessment	48	Change interval estimation, uncertainty propagation, equivalence/evidence checks, multiplicity control, sensitivity checks, or diagnostics.	**Study 24.1:** Significance tests are replaced by bootstrap and score intervals, with response-bias checks.
Change the scope of comparisons	42	Expand, narrow, or reorganise which reference-linked outcomes, conditions, models, or cross-study estimates are compared.	**Study 45.2:** Cost and sample-sensitivity analyses are omitted while network selection and MSE comparisons are added.
Change outcome scoring or aggregation	36	Change error definitions, transformations, scoring, normalisation, aggregation, or derived outcome summaries.	**Study 14.1:** Log-error ranking emphasis is replaced by absolute-error effect-size decomposition.
Analyse individual differences or moderators	24	Model or compare participant-level variation, strategy groups, population subgroups, or factors that moderate the replication result.	**Study 40:** A multilevel model estimates participant and chart variance, correlations, and ranking probabilities.
Change qualitative or visual analysis	6	Replace, introduce, or revise qualitative coding, affinity grouping, thematic interpretation, or visual analysis of the replication evidence.	**Study 1:** Visual interaction-pattern clustering replaces inferential statistics and aggregate KLM analysis.

B.3 Design additions

Table 5: Design additions codebook.

Unified phrase	Count	Explanation	Corpus example
EXPERIMENT			
Add a condition or benchmark comparison	14	Add a separable comparison of representations, manipulations, settings, systems, or human benchmark conditions.	**Study 28.4:** A newly collected human comparison for bars and framed rectangles provides an additional benchmark condition.
Add a method or model evaluation	10	Add a separable evaluation of a research tool, workflow, computational model, or learning method, including feasibility, ablations, training capacity, and sensitivity experiments.	**Study 28.3:** Additional multi-task and enlarged-training-set experiments assess network training capacity beyond the reference chart comparison.
Add a new task or question	8	Introduce a separable task or research question beyond the primary replication relationship.	**Study 29:** An outlier-detection task is added beyond the four replicated binary tasks.
Add a formative or pilot study	4	Conduct preparatory, formative, feasibility-pilot, or design-development work recorded as an addition.	**Study 38:** Two formative Wizard-of-Oz studies inform the VoxLens modes.
Add a qualitative inquiry component	3	Add a separable interview, group interview, concurrent think-aloud, or similar qualitative inquiry component.	**Study 47:** Think-aloud, post-task interviews, and a follow-up group interview explore tactile measurement strategies.
Add a follow-up assessment or questionnaire	3	Add a separable ability test, participant assessment, targeted questionnaire, or follow-up survey.	**Study 44:** A Mini-VLAT assessment is added to the AR replication.
DATA			

Continued on next page

Table 5 (continued)

Unified phrase	Count	Explanation	Corpus example
Collect additional performance evidence	29	Collect additional responses, choices, errors, completion times, or model predictions describing task or benchmark performance.	**Study 26.2:** Response times and error rates from three added visualisation tasks form one performance-evidence category.
Collect participant-background or ability measures	24	Collect additional demographics, familiarity, experience, numeracy, literacy, or ability measures.	**Study 44:** Mini-VLAT scores provide an additional visualisation-literacy measure.
Collect qualitative accounts or recordings	23	Collect additional interviews, explanations, reflections, comments, verbalised strategies, transcripts, or contextual qualitative records.	**Study 36.1:** Semi-structured interviews record memorable, confusing, and preferred chart features.
Record additional operational or contextual metadata	15	Record checks, participation or completion outcomes, incentives, configuration costs, reliability, or device/browser/context information.	**Study 16.1:** Study-configuration times of 40 and 35 minutes provide evidence about research-tool setup cost.
Collect subjective ratings or preferences	13	Collect additional confidence, difficulty, workload, preference, attitude, or other subjective ratings.	**Study 38:** NASA-TLX workload ratings supplement accuracy and interaction-time evidence.
Record additional process or interaction evidence	8	Record how task execution, interaction, or learning unfolds, using action histories, phase-specific timing, provenance, or training trajectories.	**Study 46.2:** Interaction provenance, replay trajectories, and event counts record the authoring process.
Derive additional features or evidence representations	4	Construct additional visual features, rank patterns, similarity matrices, cluster labels, or other evidence representations for extension analyses.	**Study 45.1:** Distance matrices and cluster labels are generated for additional random-initialisation, feature-layer, and training-condition sensitivity tests.
PARTICIPANT			
Recruit a separate extension sample	13	Recruit a separate human sample for an additional experiment, pilot, benchmark, or tool study.	**Study 16.1:** Two visualisation graduate students take part in the configuration-time study.
Recontact or select existing participants	3	Invite returning participants or select a subsample for a separable follow-up component.	**Study 7.1:** Fifty-three **EXPERIMENT** 1 participants return for a follow-up questionnaire.
ANALYSIS			
Analyse additional outcomes or comparisons	23	Summarise or compare outcomes for added tasks, conditions, groups, ratings, or behavioural measures.	**Study 34:** Round 2 tests visualisation-treatment effects and evaluation-period interactions.
Analyse background or contextual associations	16	Examine additional associations with demographics, ability, prior experience, device characteristics, confidence, or contextual factors.	**Study 44:** OLS regressions relate Mini-VLAT scores to workload.
Analyse additional qualitative accounts	15	Code, group, interpret, or synthesise additional interviews, reflections, strategies, or verbal records.	**Study 36.1:** Interview responses receive thematic analysis.
Develop or evaluate additional models	8	Develop, compare, or diagnose explanatory, predictive, perceptual, or computational models beyond the primary replication comparison.	**Study 37:** AIC compares perceptual-bias models for the separately framed proportional-reasoning hypothesis, beyond the primary encoding-ranking comparison.
Synthesise findings across studies	7	Combine or compare findings across multiple studies in a separable meta-analysis or cross-study synthesis.	**Study 24.2:** The replication contributes to the four-experiment standardised-effect meta-analysis.
Assess operational feasibility or study process	6	Analyse tool configuration, cost, platform reliability, participation, workflow, or replay-supported study operation.	**Study 16.1:** Configuration times of 40 and 35 minutes describe GraphUnit setup feasibility.