

Attractor strength identifies classifier label geometry as construct- irrelevant variance in machine-scored developmental drawing assessment

Rafat Saleh Madani

rsmadani@uqu.edu.sa

Umm Al-Qura University

Article

Keywords: developmental assessment, children's drawings, stimulus sampling, generalizability theory, measurement validity, cross-classified random effects

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10787903/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Attractor strength identifies classifier label geometry as construct-irrelevant variance in machine-scored developmental drawing assessment

Rafat Saleh Madani

Product Design Department, College of Designs and Arts, Umm Al-Qura University, Makkah, Saudi Arabia

Corresponding author

Correspondence to: Rafat Saleh Madani (rsmadani@uqu.edu.sa)

Abstract

Children's drawings have been used to index cognitive development for a century, and machine-scored drawing tasks have revived the practice at scale. Such assessments rest on the assumption that a drawing score reflects the child who produced it rather than the object the child was asked to draw. This study tested that assumption in a public corpus of 37,770 drawings produced by 8,084 children aged 2 to 10 across 48 object categories, scored independently by two machine-vision pipelines. Exact restricted maximum likelihood estimation for cross-classified random effects showed that variance attributable to the object category exceeded variance attributable to the child under both pipelines (ratios of 1.65 and 1.60), and adjustment for visuomotor skill, drawing effort, and category familiarity strengthened rather than attenuated this asymmetry (ratios above 2.0). Expressed on a developmental metric introduced here, one standard deviation of object difficulty is equivalent to 4.4 to 7.3 years of chronological development. A generalizability-theory decision study quantified the cost: single-prompt assessment achieves a criterion-referenced dependability coefficient of 0.11 to 0.14, and 26 to 32 prompts are required to reach 0.80. Recovering the full classifier confusion structure explains why estimated category difficulty fails to replicate across pipelines. Each category's attractor strength, defined as the probability mass it draws from drawings of every other category, is a highly stable property of a scoring pipeline (split-half $r = 1.00$ and 0.97), accounts for 54% of category-effect variance under one pipeline, and does not transfer between pipelines ($r = 0.04$). Removing it unmask a referent-intrinsic component that the raw comparison had hidden ($r = 0.21$ to 0.40 , $p = 0.006$). Much of what presents as object difficulty is therefore construct-irrelevant variance introduced by the scoring instrument, which implies that difficulty norm tables are not transportable across pipelines and that classifier-scored drawing measures should report attractor-adjusted category effects alongside raw ones.

Keywords: developmental assessment; children's drawings; stimulus sampling; generalizability theory; measurement validity; cross-classified random effects

1. Introduction

The idea that a child's drawing reveals the child's mind is among the oldest working assumptions in developmental psychology. The survey of children's art by Ricci (1894) marks the beginning of systematic interest; the Draw-a-Man test (Goodenough 1926) and its revision (Harris 1963) converted the intuition into an instrument, and the family of figure-drawing measures that followed, Koppitz (1968), the Draw A Person system (Naglieri 1988), the emotional-disturbance screener built on it (Naglieri et al. 1991), and the more recent DAP:IQ (Reynolds and Hickman 2004), remains in circulation in schools and clinics because it is fast, language-independent, and requires almost no equipment. Those same properties have made drawing tasks attractive again in the machine-learning era, where classifier-scored recognizability can be computed automatically over corpora orders of magnitude larger than any hand-scored study (Long et al. 2018; Long et al. 2019; Long et al. 2024; Mukherjee et al. 2025).

Interest in drawing as a cognitive measure has, if anything, intensified recently: a 2025 special issue of *Memory & Cognition* collected twenty-five studies using drawing to probe cognition across the lifespan (Bainbridge et al. 2025), and a parallel developmental literature continues to debate what graphic change actually indexes (Thomas and Silk 1990; Morra 2005; Fabris et al. 2023).

The psychometric record for these instruments is, it should be said, already mixed. Inter-rater reliability is consistently high (Naglieri and Maxwell 1981; Abell et al. 2001) and score stability is adequate (Williams et al. 2006), but concurrent validity against established ability measures has been modest at best (Wisniewski and Naglieri 1989), and the largest evaluation to date (2,543 children) reported correlations with Raven's Matrices between 0.16 and 0.50, a sensitivity of 0.12, and a false-negative rate near 88%, concluding that the instrument fails as an assessment of intellectual functioning (Troncone et al. 2021). Reviews reaching similar conclusions have appeared periodically for over half a century (Goodenough and Harris 1950; Swenson 1968; Scott 1981), and a direct test concluded flatly that human figure drawings do not measure intellectual ability (Willcock et al. 2011). What that literature has not done is ask why the numbers come out this way. Our analysis suggests one structural answer that applies with equal force to hand-scored and machine-scored versions.

Both the classical and the computational literatures share a structural feature that has attracted little scrutiny. In each case, a child is asked to draw something (a person, a house, a tree, a cat), and the resulting score is interpreted as a property of the child. The object is treated as a neutral vehicle. Yet nothing guarantees neutrality. Objects differ in how many parts they have, how canonical their viewpoint is, how much they resemble other objects, how often children encounter them, and how forgiving they are of imprecise execution. The developmental literature on graphic representation has documented exactly these object-level dependencies for decades, in the canonical forms children settle on (Kellogg 1969; Freeman 1980), in

the rigidity and flexibility of drawing routines (Karmiloff-Smith 1990), in how depiction systems mature (Willats 2005; Golomb 2004; Cox 1992), and in the account given by Piaget (1929) of what children draw as opposed to what they see, without treating them as a measurement problem, and more recently in the semantic part structure of what children draw (Huey et al. 2022). If those differences translate into score differences of a magnitude comparable to developmental differences, then the score is not a measure of the child in any straightforward sense. Suggestive evidence has been available for some time. Asking children to draw a football player rather than a generic person changes figure size, posture, shading, and detail, and does so differently across countries (Baluch et al. 2017) a demonstration that the referent shapes the score, though it was not framed as a measurement problem.

The consequence is a threat to score interpretation of the kind Lord and Novick (1968) first formalized as systematic error, and it is a specific instance of a general problem that psychology has confronted before in other domains. Clark (1973) showed that treating linguistic materials as fixed rather than sampled inflates Type I error in psycholinguistics; Judd et al. (2012) made the parallel argument for social psychology, and Westfall et al. (2014) worked out its design and power consequences; Yarkoni (2022) generalized it into a claim about the field's inferential practices at large. Developmental drawing assessment has not received the same treatment, in part because the classical instruments use a single fixed prompt, which makes the stimulus factor invisible: with one object, there is no between-object variance to estimate, and the confound is silently absorbed into the norm table.

A parallel literature in machine learning has independently documented one half of what we report. Classifier predictions are known to be systematically imbalanced across labels even on balanced data; CLIP in particular inherits a skewed label prior from web-scale pretraining (Wang et al. 2022), and a family of corrections exists under the heading of logit adjustment (Menon et al. 2021). That work is concerned with classification accuracy. To our knowledge it has not been connected to measurement validity in the behavioural sciences, where the same pathology does something different and arguably worse.

Test theory has, in fact, two relevant vocabularies. The first concerns items: item-response models since Rasch (1960) and Masters (1982) treat item difficulty as an explicit parameter, and the machinery of differential item functioning (Holland and Thayer 1988; Holland and Wainer 1993; Swaminathan and Rogers 1990; Hagquist and Andrich 2017) exists precisely to detect items whose difficulty shifts between groups of equal ability. Developmental instruments are routinely audited this way: the PDMS-2 (Valentini and Zanella 2022), the TGMD-3 (Valentini et al. 2022), the Autism Behavior Checklist (Peng et al. 2024), and youth mental-health screeners (Cauffman and MacIntosh 2006) have all had their item difficulties and DIF characterized, with semi-automated tooling now available for the purpose (Wijayanto et al. 2023). Drawing-based measures have not received this treatment, and our results indicate why it matters.

The second vocabulary concerns scores: construct-irrelevant variance (CIV), systematic variation in scores produced by something other than the attribute the test intends to measure, and one of the two classical threats to validity (Messick 1989; Haladyna and Downing 2004; Kane 2013). Messick’s distinction between construct-irrelevant difficulty and construct-irrelevant easiness is exactly the right frame here: a category that a classifier readily assigns confers unearned easiness on every child asked to draw it, while a category the classifier disfavours confers unearned difficulty. The CIV literature has catalogued many sources (item flaws, testwiseness, rater severity, method effects; Palermo 2022), but all of them originate in the examinee, the task, or a human judge. An automated scorer whose label prior systematically advantages some prompts over others is a source of CIV with no precedent in that catalogue, and one that is invisible in the score it produces.

Our contribution is the bridge, not either endpoint. That classifiers have skewed label priors is established in machine learning; that CIV threatens validity is established in psychometrics. What has not been done, so far as we can determine, is to measure how much of a psychological instrument’s score variance the former consumes.

Large digital corpora make the problem tractable for the first time. When thousands of children each draw several objects from a common pool of dozens, the design becomes cross-classified (children crossed with categories), and the variance attributable to each can be separated. We exploit that structure here.

1.1 The present study

Machine-scored analyses of children’s drawings are also no longer confined to one corpus or one scoring method: deep-network feature representations have been applied to drawing-completion tasks (Philippson et al. 2022) as well as to free production, which makes the question of what such scores measure a general one rather than a property of a single dataset.

We ask five questions.

Q1 (Magnitude). How much of the variance in machine-scored drawing recognizability is attributable to the object category, relative to the child? If the category component is small, stimulus dependence is a nuisance; if it rivals or exceeds the child component, it is a validity threat.

Q2 (Interpretation). Stimulus effects in log-odds units are hard to evaluate. We therefore introduce the Developmental Displacement Index (DDI), which rescales stimulus difficulty into the units the assessment claims to measure, years of chronological development.

Q3 (Explanation). Are category effects reducible to something already understood? We test whether they survive adjustment for visuomotor control (shape-tracing skill), drawing effort (stroke count, duration, ink coverage), and category familiarity (how often children spontaneously draw the category).

Q4 (Consequence and remedy). What does stimulus dependence cost an assessment in practice, and can it be corrected? We conduct a generalizability-theory decision study, simulate realistic short batteries, quantify age-ordering errors, and test whether category norming repairs the damage.

Q5 (Mechanism). If category difficulty is not a stable property of the referent, what is it a property of? We recover the full classifier confusion structure and test whether apparent object difficulty is generated by the scoring model’s label geometry rather than by the object.

Our central empirical claim is that the stimulus-linked component of a drawing score exceeds the child component. Our central mechanistic claim is that most of that component, at least under one widely used scoring pipeline, is not about the object at all: it is a property of where the classifier deposits probability mass. Our central practical claim follows from both: difficulty norm tables are not transportable, and correcting for classifier geometry is a prerequisite for comparing category effects across studies.

2. Methods

2.1 Corpus

We analysed the KidDraw corpus (Long et al. 2024), collected via a freestanding touchscreen kiosk at a children’s museum in San Jose, California. Children aged 2 to 10 were verbally prompted to draw object categories drawn from a pool of 48, spanning animals (cat, camel, octopus, whale), artefacts (chair, piano, scissors, train), and other familiar concepts (face, hand, house, tree). The kiosk also administered shape-tracing trials, providing a per-session index of visuomotor control.

Each drawing carries a recognizability score derived from a machine-vision classifier: the log-odds that the drawing is assigned to its intended category. Two independent scoring pipelines are distributed with the corpus, one based on VGG embeddings pretrained on ImageNet (Simonyan and Zisserman 2015; Deng et al. 2009) and one based on CLIP embeddings (Radford et al. 2021). Deep-network features of this kind are now the standard basis for quantifying sketches at scale (Yamins et al. 2014; Fan et al. 2018; Xu et al. 2022), and the reference sketch corpora against which such methods are calibrated, TU-Berlin (Eitz et al. 2012), Sketchy (Sangkloy et al. 2016), and Quick, Draw! (Ha and Eck 2017; Fernandez-Fernandez et al. 2019), are all adult or mixed-age, which is one reason a child-specific validity analysis is needed. We treat these as parallel measurement instruments and report every analysis under both. The CLIP pipeline covers the full corpus ($n = 37,770$ drawings, 8,084 children); the VGG pipeline covers a category-balanced subset

($n = 22,272$ drawings, 7,032 children). Sample characteristics appear in Table 1. The corpus is openly available, which makes every result below independently reproducible.

2.2 Statistical model

Let $y(i,j,k)$ denote the recognizability score of the drawing produced by child i for category k . We fit the cross-classified linear mixed model

$$y(i,j,k) = \mathbf{x}(i,j,k)^T \boldsymbol{\beta} + u(i) + v(k) + \varepsilon(i,j,k)$$

with

$$u(i) \sim N(0, \sigma^2 C), \quad v(k) \sim N(0, \sigma^2 K), \quad \varepsilon(i,j,k) \sim N(0, \sigma^2 e),$$

all mutually independent. Here $u(i)$ is the child effect, $v(k)$ the object-category effect, and $\mathbf{x}(i,j,k)$ the fixed-effect design vector. Children are crossed with categories rather than nested within them, which is what permits $\sigma^2 C$ and $\sigma^2 K$ to be separately identified. This is the crossed-random-effects structure standard in psycholinguistics and increasingly elsewhere (Baayen et al. 2008; Barr et al. 2013), normally fitted with lme4 (Bates et al. 2015); variance-partition coefficients follow the conventions of Nakagawa and Schielzeth (2013). In matrix form, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_C \mathbf{u} + \mathbf{Z}_K \mathbf{v} + \boldsymbol{\varepsilon}$, so that

$$\mathbf{V} = \text{Var}(\mathbf{y}) = \sigma^2 C \mathbf{Z}_C \mathbf{Z}_C^T + \sigma^2 K \mathbf{Z}_K \mathbf{Z}_K^T + \sigma^2 e \mathbf{I}.$$

We report three variance-partition coefficients,

$$\text{ICC}_{\text{child}} = \sigma^2 C / (\sigma^2 C + \sigma^2 K + \sigma^2 e), \quad \text{ICC}_{\text{cat}} = \sigma^2 K / (\sigma^2 C + \sigma^2 K + \sigma^2 e),$$

together with the stimulus-to-child variance ratio $R = \sigma^2 K / \sigma^2 C$, which is the quantity of primary interest. $R > 1$ means the object contributes more to the score than the child.

2.3 Estimation

Standard mixed-model software for crossed random effects was unavailable in our computing environment, so we implemented exact restricted maximum likelihood (REML; Patterson and Thompson 1971) using a factorization tailored to this design. The approach exploits two structural facts: each drawing belongs to exactly one child, and the number of categories is small.

Because children are a partition of the observations, the matrix $\mathbf{A} = \sigma^2 C \mathbf{Z}_C \mathbf{Z}_C^T + \sigma^2 e \mathbf{I}$ is block diagonal, with one compound-symmetric block per child. Each block inverts in closed form:

$$\begin{aligned} \mathbf{A}(i)^{-1} &= (1/\sigma^2 e) \left[\mathbf{I}(n_i) - (\sigma^2 C / (\sigma^2 e + n_i \sigma^2 C)) \mathbf{J}(n_i) \right], \\ \log|\mathbf{A}(i)| &= (n_i - 1) \log \sigma^2 e + \log(\sigma^2 e + n_i \sigma^2 C), \end{aligned}$$

where n_i is the number of drawings by child i and J is the all-ones matrix. The category effects are then folded in with the Woodbury identity, which requires inverting only a $q \times q$ matrix ($q = 48$, or $q = 96$ when random slopes are included):

$$V^{-1} = A^{-1} - A^{-1} Z_K (G_K^{-1} + Z_K^T A^{-1} Z_K)^{-1} Z_K^T A^{-1},$$

$$\log|V| = \log|A| + \log|G_K| + \log|G_K^{-1} + Z_K^T A^{-1} Z_K|.$$

Variance components are obtained by minimizing the REML criterion

$$-2 \log L_R = \log|V| + \log|X^T V^{-1} X| + y^T P y,$$

$$P = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1},$$

over the log-transformed variance parameters, initialized at method-of-moments estimates. Generalized least squares gives $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ with $\text{Var}(\hat{\beta}) = (X^T V^{-1} X)^{-1}$. Category effects are recovered as best linear unbiased predictors (Henderson 1975). This is exact REML, not an approximation. All computation used NumPy (Harris et al. 2020) and SciPy (Virtanen et al. 2020); cross-validated prediction used scikit-learn (Pedregosa et al. 2011) and figures were produced with Matplotlib (Hunter 2007).

We validated the implementation against simulated data with known parameters (Figure S1): across six independent datasets the estimator recovered $\sigma^2 C$ (mean 0.519 against a true value of 0.50), $\sigma^2 K$ (0.840 against 0.80), $\sigma^2 e$ (1.189 against 1.20), and the age slope (0.699 against 0.70).

2.4 The Developmental Displacement Index

Variance components in classifier log-odds have no intrinsic interpretation. Because the same model estimates the fixed effect of age, β_{age} , in the same units, stimulus difficulty can be rescaled into developmental time. We define

$$\text{DDI} = \sigma K / \beta_{\text{age}},$$

the number of years of chronological development equivalent to one standard deviation of object-category difficulty. A range-based variant uses the spread of the category BLUPs, $(\max \hat{v}(k) - \min \hat{v}(k)) / \beta_{\text{age}}$. The DDI answers the question a practitioner actually cares about: if I change the object, how many years of development have I apparently added or removed?

2.5 Models fitted

M1 (unconditional): intercept only, child and category random effects.

M2 (+ age): adds the fixed linear effect of age, centred at 6 years.

M3 (+ random slopes): adds a category-specific random slope of age, testing whether developmental trajectories differ by object.

M4 (+ covariates): adds shape-tracing score, stroke count, drawing duration, mean ink intensity, and category drawing frequency, each standardized.

2.6 Decision study

We treat the assessment as a person-by-item design with items sampled from a universe of admissible object prompts (Cronbach et al. 1972; Brennan 2001). For a battery of n_k prompts, the relative (norm-referenced) coefficient is

$$E\rho^2(n_k) = \sigma^2C / (\sigma^2C + \sigma^2e / n_k),$$

and the absolute (criterion-referenced) coefficient is

$$\Phi(n_k) = \sigma^2C / (\sigma^2C + (\sigma^2K + \sigma^2e) / n_k).$$

Generalizability theory has been applied to a drawing instrument once before, to our knowledge: Hyland and Rae (2001) analysed Koppitz’s human-figure-drawing scoring system, treating scoring items and raters as facets. Our design differs in the facet that matters here. Their items are scoring criteria applied to one drawing; ours are the objects children are asked to draw. Theirs asks how dependably a rubric can be applied; ours asks how dependably a child can be characterized when the referent varies.

The difference between them is precisely the reliability cost of stimulus dependence. The distinction matters because developmental screening is a criterion-referenced activity: a child is compared against an age norm, not against the other children who happened to be tested that day.

2.7 Applied simulations

Short batteries. For $k \in \{1, 2, 3, 4, 6, 8\}$ we resampled k prompts per child (200 replications), fit a regression predicting chronological age from the mean score on a random half of children, and evaluated mean absolute error on the held-out half. We compared raw scoring against category-standardized scoring.

Age-order violations. We paired children aged 7–10 with children aged 2–4, gave each a single randomly assigned prompt, and recorded how often the older child received the lower score. Chance is 0.500; a perfect instrument would approach 0.

Cross-pipeline convergence. To test whether category difficulty is a property of the referent or of the scorer, we refit both models on the 22,272 drawings scored by both pipelines, so that any difference is attributable to the scoring method rather than to sampling.

2.8 Classifier label geometry

Both scoring pipelines distribute a full probability vector over all 48 categories for every drawing, not merely a score for the intended category. Recovering these vectors allows the confusion structure of each pipeline to be characterized directly. Let $P(j | d)$ be the probability that a pipeline assigns label j to drawing d , and let $L(d)$ be the category the child was asked to draw. We form the mean confusion matrix

$$C(i, j) = (1 / |\{d : L(d) = i\}|) \cdot \sum \text{over } d \text{ with } L(d) = i \text{ of } P(j | d),$$

and define the attractor strength of category k as

$$A(k) = (1 / (K - 1)) \cdot \sum \text{over } i \neq k \text{ of } C(i, k),$$

the mean probability that drawings of every other category assign to k . The exclusion $i \neq k$ is essential: no drawing of category k contributes to $A(k)$, so $A(k)$ contains no information whatsoever about how children render k . It is a pure description of the classifier’s label geometry, how much probability mass a category absorbs when the model is looking at something else.

$A(k)$ is not a new statistic so much as a targeted variant of one already known in machine learning. The label-bias literature estimates a classifier’s implicit label prior by averaging its predicted probabilities across a dataset (Wang et al. 2022; Allingham et al. 2023), and corrects it by logit adjustment (Menon et al. 2021). Our contribution to the statistic itself is only the exclusion $i \neq k$, which is unnecessary for the ML use case but essential here: without it, the estimator is contaminated by the very scores whose validity we are trying to assess. What is new in this paper is not the quantity but the demonstration that it governs the behaviour of a psychological measurement instrument.

We deliberately do not use neighbourhood entropy or any index computed from a category’s own drawings. Such indices are mechanically related to the recognizability score itself, and we found that including them inflates the apparent explanatory power of geometry while simultaneously absorbing genuine referent variance (see Section 3.9). $A(k)$ is the only geometry index used in the analyses reported below. To confirm that $A(k)$ is a stable property rather than sampling noise, we estimated it independently on two disjoint random halves of the non- k drawings and correlated the two estimates.

2.9 Use of large language models

Analysis code, figures, and manuscript drafts were prepared with the assistance of a large language model (Anthropic Claude). All analyses were specified, executed, and verified against the released code and result files by the author, who takes full responsibility for the content, the accuracy of all reported values, and the interpretation of the findings. The model is not an author and does not meet authorship criteria.

3. Results

3.1 The object contributes more than the child (Q1)

Figure 1 and Table 2 give the variance decomposition. Under the CLIP pipeline with age in the model (M2), the object category accounts for 22.6% of score variance while the child accounts for 13.7%, a stimulus-to-child ratio of $R = 1.65$. The VGG pipeline gives 17.8% versus 11.1%, $R = 1.60$. The two pipelines differ in absolute variance magnitudes but agree closely on the ratio, which is the quantity of interest.

The ordering is not an artefact of including age. In the unconditional model (M1), $R = 1.25$ (CLIP) and 1.14 (VGG); adding age removes between-child variance, as it should, since age is a child-level property, and thereby sharpens the asymmetry. In every specification and under both scorers, the object category outweighs the child.

3.2 Stimulus effects are large in developmental units (Q2)

The estimated age slope is 0.2146 log-odds per year ($SE = 0.0060$) under CLIP and 0.1750 ($SE = 0.0058$) under VGG, both highly reliable and consistent with the developmental gains reported for this corpus by Long et al. (2024). Dividing the category standard deviation by these slopes yields DDI = 5.00 years (CLIP) and 4.41 years (VGG).

One standard deviation of object difficulty is therefore worth roughly four and a half to five years of development in a sample spanning eight years. The range-based index is more extreme still: the gap between the most and least favourable category corresponds to 24.4 years (CLIP) and 16.7 years (VGG) of apparent development, two to three times the entire age range studied. Figure 2 displays the full category effect distribution with a secondary axis in developmental years.

3.3 Stimulus effects are not reducible to motor skill, effort, or familiarity (Q3)

The obvious deflationary explanations fail. Supplementary Table S1 reports the covariate-adjusted model. Shape-tracing skill predicts recognizability robustly (CLIP $b = 0.262$, $SE = 0.015$; VGG $b = 0.279$, $SE = 0.015$, both $p < 0.001$), confirming that visuomotor control matters. Drawing duration contributes as well. But category drawing frequency how often children spontaneously draw a category, the most intuitive candidate explanation for category difficulty, is not reliable under either pipeline (CLIP $p = .568$; VGG $p = .836$).

Critically, adjustment does not shrink the category component. It grows: R rises from 1.65 to 2.06 (CLIP) and from 1.60 to 2.08 (VGG). Because the covariates absorb child-level variance without absorbing category-level variance, the stimulus effect becomes proportionally larger in the better-specified model. Correspondingly, the DDI rises to 7.31 and 7.20 years. A researcher who carefully controls for motor ability and effort ends up with a score that is more, not less, dominated by the choice of object.

3.4 Developmental trajectories differ by object

Adding category-specific random slopes of age improves fit decisively (CLIP: $\chi^2(1) = 423.7$, $p < 0.001$; VGG: $\chi^2(1) = 184.5$, $p < 0.001$). The slope standard deviations are 0.104 (CLIP) and 0.080 (VGG) log-odds per year, against mean slopes of 0.215 and 0.175. A category one standard deviation above the mean in developmental sensitivity therefore shows roughly 1.5 times the age-related gain of the average category, and a category one standard deviation below shows roughly half. Stimulus dependence is not a constant offset that shifts all children equally; it interacts with development itself.

3.5 The measurement cost (Q4)

Figure 3 and Table 3 presents the decision study. A single-prompt assessment, which is the administration format of every classical figure-drawing instrument, yields an absolute dependability coefficient of $\Phi = 0.137$ (CLIP) and 0.111 (VGG). Reaching the conventional threshold of 0.80 for individual-level decisions requires 26 prompts (CLIP) or 32 prompts (VGG); reaching 0.90 requires 57 and 72. The gap between the relative and absolute coefficients peaks in the range where real batteries live, around 0.07 at eight prompts, and this gap is exactly the stimulus-dependence penalty.

Figure 5 translates this into error rates. When a child aged 7–10 and a child aged 2–4 are each given one randomly chosen prompt, the older child receives the lower score 34.7% of the time (CLIP) or 34.3% (VGG); rates and intervals for both scoring rules are given in Table 5. An instrument intended to discriminate across an eight-year developmental gap misorders more than a third of maximally separated pairs, purely as a function of which object each child was asked to draw.

3.6 Norming does not fix it

Category standardization helps, but only marginally. Across battery lengths, the reduction in age-recovery error is on the order of 0.010 to 0.022 years, statistically reliable (bootstrap 95% CIs exclude zero at most lengths; Supplementary Table S2) but practically negligible against a baseline error near 1.8 years. Age-order violations fall only from 0.347 to 0.331. Section 3.7 explains why, and this is our most consequential finding.

3.7 A dissociation: magnitude replicates, ordering does not

Restricting to the 22,272 drawings scored by both pipelines, so that sampling is held constant, the two scorers agree moderately at the level of individual drawings ($r = 0.426$). They agree almost exactly on the magnitude of stimulus dependence: $R = 1.61$ (CLIP) versus $R = 1.60$ (VGG). And they agree substantially on which categories are developmentally sensitive: the category-specific age slopes correlate at $r = 0.734$ ($p < 0.001$).

But they do not agree on which categories are hard. The correlation between the two pipelines' estimates of category baseline difficulty is $r = 0.205$ ($p = .166$) across the 47 categories common to both pipelines, statistically indistinguishable from zero (Figure 4, Table 4). Under CLIP, snail and hand are the most favourable categories and tiger and piano the least; under VGG, clock and apple are most favourable and bird and dog least.

The implication is direct. Stimulus dependence in drawing assessment is large, reproducible in magnitude, and reproducible in its interaction with age; but its direction is a joint property of the referent and the scoring instrument, not of the referent alone. That is why norming fails, and why it will keep failing: a difficulty table derived from one scoring pipeline does not transfer to another. There is no fixed set of object difficulty constants waiting to be estimated and published.

3.8 What category difficulty is actually a property of (Q5)

Section 3.7 establishes that apparent object difficulty does not transfer between scoring pipelines. It does not explain why. The full classifier confusion structure supplies the answer, summarised in Figure 6 and Table 6.

Attractor strength is an extraordinarily stable property of a scoring pipeline. Estimated on two disjoint random halves of the non-target drawings, the two estimates correlate at $r = 1.000$ (CLIP) and $r = 0.966$ (VGG). Whatever $A(k)$ measures, it is not noise.

It predicts apparent category difficulty. Under CLIP, $A(k)$ correlates with the estimated category effect at $r = 0.737$ ($p < 0.001$), accounting for 54.3% of the variance in category effects on its own. Under VGG the relationship is weaker but in the same direction, $r = 0.288$ ($p = 0.049$), accounting for 8.3%. Recall that $A(k)$ is computed entirely from drawings of other categories. A variable containing no information about how children draw snails predicts more than half the variance in how easy snails appear to be.

The concrete case is instructive. Under CLIP, the single most favourable category in Figure 2 is snail. Snail is also CLIP's largest probability sink by a wide margin ($A = 0.0775$, against a median across categories of 0.0148). Children's snail drawings are not unusually good. CLIP simply assigns probability to snail readily, including when it is looking at drawings of something else, and this inflates the score of every child asked to draw one.

Crucially, attractor structure is itself pipeline-specific: $A(k)$ under CLIP and under VGG correlate at $r = 0.040$ ($p = 0.79$). Two embedding spaces impose two unrelated label geometries. This is the mechanism behind the Section 3.7 dissociation: a large component of apparent difficulty is an instrument property, and instrument properties do not transfer between instruments. We stress that the magnitude of this component differs sharply between the two pipelines, dominant under CLIP, minor under VGG, and Section 4.4 shows

that the difference is predictable from how each scorer was built. The general claim we defend is not that label geometry always dominates, but that it can, and that nothing in current reporting practice would reveal it if it did.

3.9 Unmasking the referent-intrinsic component

If apparent difficulty is a compound of an instrument artefact and a genuine object effect, then removing the artefact should reveal the object effect. This is a falsifiable prediction, and it is confirmed.

Residualizing each pipeline's category effects on its own attractor strength and re-correlating the residuals raises cross-pipeline agreement from $r = 0.205$ ($p = 0.166$, n.s.) to $r = 0.396$ ($p = 0.006$). The increase is $\Delta r = 0.190$, 95% bootstrap CI [0.013, 0.344], excluding zero. There is a real, scorer-independent component of object difficulty. It was invisible in the raw comparison because a larger, pipeline-specific artefact was sitting on top of it.

Two cautions bound this result. First, $r = 0.396$ sits close to the smallest correlation detectable at 80% power with 47 categories ($r = 0.399$), so the referent-intrinsic component is estimated with limited precision, and we would not defend a point estimate of its size. Second, the adjustment is a first-order correction using a single geometry index; a richer characterization of label geometry would likely recover more.

We note a methodological trap encountered here, because it may catch others. Our first attempt at this analysis also included neighbourhood entropy, the Shannon entropy of a category's own mean probability vector, as a geometry predictor. It appeared to perform superbly, raising joint explanatory power to 74% and 81%. But entropy is computed from the category's own drawings, and its principal component is the diagonal of the confusion matrix, which is the recognizability score itself. The apparent explanatory power was partly an identity. Worse, because entropy absorbs genuine referent variance along with instrument variance, including it destroyed the unmasking effect ($\Delta r = +0.014$, 95% CI [-0.315, 0.285]). Only the strictly non-circular index recovers the referent signal. Analysts characterizing classifier geometry should verify that their indices exclude the target category's own data.

4. Discussion

4.1 What the score measures

The results support a specific and uncomfortable conclusion. In a drawing-based developmental assessment scored by contemporary machine-vision methods, more of the variance in the score is attributable to the object the child was asked to draw than to the child. This holds under two independent scoring pipelines, survives adjustment for visuomotor skill and effort, and is not explained by how familiar the object is.

Rescaled into the units the instrument claims to measure, the effect of the object is on the order of five to seven years of development.

This does not mean drawing tasks measure nothing about children. The age effect is robust and replicates prior work on this corpus (Long et al. 2024); tracing skill predicts performance; developmental gains are real. The claim is narrower and more specific: a drawing score is a joint function of child, referent, and scoring instrument, in which the child's contribution is the smallest of the three, and single-prompt administration provides no way to separate them.

Sections 3.8 and 3.9 sharpen this further, and in a direction we did not anticipate. Most of what presents as object difficulty is not about the object. Attractor strength, a quantity computed without using a single drawing of the category in question, accounts for 54% of category-effect variance under CLIP. A genuine referent-intrinsic component does exist, and becomes visible once the artefact is removed, but it is the smaller part. The honest summary of this paper is therefore not that the object outweighs the child, which was our starting hypothesis, but something more specific: the second-largest identifiable source of variance in a machine-scored drawing assessment is the label geometry of the scoring model.

Framed in test-theoretic terms, the label geometry of an automated scorer is a source of construct-irrelevant variance, and specifically of Messick's construct-irrelevant easiness: a child asked to draw a category the classifier favours receives a higher score without any corresponding difference in the attribute being measured. The classical remedies for CIV (item review, rater training, balancing forms) are all aimed at sources that reside in the task or the judge, and none of them would detect or correct this one.

This matters because that source is entirely invisible in the score. A practitioner who sees a child score poorly on tiger and well on snail has no way to know that the difference is substantially about where CLIP deposits probability mass rather than about the child or the animals.

4.2 Implications for practice

Single-prompt instruments are not defensible for individual decisions. A dependability coefficient near 0.12 is far below any professional standard for individual-level interpretation (American Educational Research Association et al. 2014). Classical figure-drawing tests administer one prompt and interpret the result against age norms, which is an inherently criterion-referenced use and therefore the use most damaged by stimulus variance. Our estimates suggest such instruments would need on the order of thirty prompts to support the decisions routinely made from them.

Convergence with the clinical record. It is worth noting how closely our estimate converges with independent clinical evidence. Troncone et al. (2021) found a sensitivity of 0.12 for the DAP:QSS in 2,543 children, strikingly close to the single-prompt dependability coefficient of 0.11 to 0.14 we estimate here by

an entirely different route. Two literatures that do not cite each other arrive at the same number, and our variance decomposition supplies a mechanism: with one prompt, the score is dominated by which object was requested rather than by who drew it.

Norm tables are not the remedy. The dissociation in Section 3.7 shows that object difficulty does not have a scorer-independent value. Any correction calibrated on one scoring pipeline is not transportable to another, which means the field cannot resolve stimulus dependence by publishing difficulty constants.

Report attractor-adjusted category effects. Any study that scores drawings with a classifier and reports category-level results should also report those results adjusted for attractor strength. The computation requires only the full probability matrix the classifier already produces, costs nothing additional to collect, and is the difference between a category effect that transfers across studies and one that does not. We provide the code.

Prefer in-domain balanced scorers, or debias zero-shot ones. The pipeline whose label prior was near-uniform showed an instrument artefact roughly six times smaller. Researchers scoring drawings with a zero-shot vision-language model are importing a web-scale label prior into a developmental measurement, and should either fit a classifier on balanced in-domain data or apply an off-the-shelf logit-adjustment correction (Menon et al. 2021; Wang et al. 2022) before interpreting category-level results.

Prompt sampling should be reported. At minimum, studies using drawing tasks should state how prompts were selected, report results across prompts rather than pooled, and treat the object as a random factor in the analysis, which is the discipline psycholinguistics adopted after Clark (1973) and social psychology after Judd et al. (2012).

4.3 Implications for cross-cultural and equity research

Object familiarity varies systematically across cultural and socioeconomic contexts. Our finding that category frequency does not predict category difficulty is reassuring on one narrow point, but the deeper concern has a long empirical trail in the figure-drawing literature specifically: scoring systems transplanted across national samples have repeatedly produced norm shifts, in Greek children (Bardos et al. 1989) and Canadian ones (Saklofske and Braun 1992), even where the systems agreed internally. The concern is old: the collection assembled by Paget (1932) of more than sixty thousand drawings from outside Europe already made clear that what children draw and how they draw it varies markedly by cultural context, and later work documented systematic differences between ethnic groups (Gonzalez 1982), between Japanese and American children in what they depict (Steward et al. 1982), and in the developmental items that appear at a given age (Groves and Fried 1991). The deeper concern also stands: because the choice of prompt can displace a score by several developmental years, any prompt set carries assumptions about which objects

are equally available to all children being compared, and children’s engagement with depicted content itself varies across contexts (Zhu et al. 2025). Long et al. (2023) found broadly similar developmental trajectories in U.S. and Chinese samples, which is encouraging; our results suggest that such comparisons should nonetheless be conducted and reported at the level of individual prompts, since a difference localized to a few categories could be masked or manufactured by pooling.

4.4 Limitations

Several constraints bound these conclusions.

The corpus is a convenience sample recruited at a single children’s museum in California, and museum visitors are not representative of children generally in socioeconomic background, exposure to touchscreens, or prior drawing experience. The age variable is recorded in whole years, which coarsens the age slope and therefore inflates the DDI to an unknown degree; a finer age measure would likely reduce the index somewhat, though it would not change the variance ratio, which does not depend on age at all.

Our outcome is machine-scored recognizability, not a clinical developmental score. This is the appropriate outcome for the computational literature we address, and the two-pipeline design guards against idiosyncrasies of any one scorer, but we cannot demonstrate directly that the same asymmetry holds for hand-scored instruments such as the DAP:IQ, whose own construct status remains contested (Rehrig and Stromswold 2018). That the effect appears under two quite different embeddings, and that its magnitude is nearly identical under both, makes a purely algorithmic explanation less plausible, but the extension to human scoring is an inference, not a finding.

The design is observational: children were not randomly assigned to categories in a fully balanced fashion, and children who drew more categories may differ from those who drew fewer. The cross-classified model accommodates unbalanced data, but residual selection effects cannot be excluded.

The geometry analysis has its own bounds. With 47 categories, the unmasked referent-intrinsic correlation ($r = 0.40$) is estimated imprecisely, and although the bootstrap CI on the improvement excludes zero, the analysis cannot support a confident estimate of how large the true object effect is, only that it is nonzero and smaller than the artefact. The asymmetry between pipelines, by contrast, does have an explanation, and a useful one. The two pipelines differ sharply in how skewed their label priors are. Under CLIP the attractor distribution spans a 30-fold range across categories (Gini = 0.406, normalized entropy = 0.928); under VGG it spans a 1.4-fold range and is nearly uniform (Gini = 0.045, normalized entropy = 0.999). This is what one would predict from how the two scorers were built. The CLIP pipeline scores zero-shot and therefore imports a label prior inherited from web-scale pretraining, precisely the imbalance documented by Wang et al. (2022). The VGG pipeline uses a classifier fit to the balanced drawing corpus itself, so its label prior

is close to uniform by construction. The size of the instrument artefact is thus predictable from the scoring architecture rather than mysterious, though confirming the generalization would require additional pipelines.

Finally, our remedy analysis is deliberately conservative. We tested the simplest corrections, category centering and standardization. More sophisticated approaches may do better, and the most promising is to treat prompts explicitly as items within an item-response framework, estimating a difficulty and a discrimination parameter for each object and testing directly for differential functioning across age, culture, and scoring pipeline (Rasch 1960; Masters 1982; Holland and Wainer 1993). Our attractor adjustment is best understood as a first-order approximation to what a properly specified IRT model would deliver. Such a model would also address a limitation of the variance-component approach used here: person reliability and item reliability are not the same quantity, and conflating them obscures which end of the instrument is failing (Linacre 1997). We regard this as the natural next step, and the code released here is intended to make it easy.

4.5 Conclusion

For a century, drawing-based assessment has proceeded as though the object were a window through which the child could be observed. These results indicate that the window is doing more of the work than what lies behind it, and that a large part of what we had taken for the window is in fact the frame. The stimulus-linked variance in a machine-scored drawing assessment exceeds the child-linked variance, and the majority of that stimulus-linked variance, under at least one widely used pipeline, is a property of the scoring model rather than of the object or the child.

The finding is not that children's drawings are uninformative. They plainly are informative, and the developmental signal in this corpus is robust. The finding is that three things are superimposed in a single number child, referent, and instrument, and that current practice reports the sum as though it were the first. Separating them is not optional bookkeeping. It is the difference between a measurement that transfers across studies and one that does not, and the arithmetic required is already available in output that every classifier-based pipeline produces and then discards.

References

- Abell SC, Wood W, Liebman SJ (2001) Children's human figure drawings as measures of intelligence: the comparative validity of three scoring systems. *Journal of Psychoeducational Assessment* 19(3):204-215
- Allingham JU, Ren J, Dusenberry MW et al (2023) A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. *Proceedings of the 40th International Conference on Machine Learning*, 547-568
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) Standards for educational and psychological testing. American Educational Research Association
- Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4):390-412
- Bainbridge WA, Chamberlain R, Wammes J et al (2025) Drawing as a means to characterize memory and cognition. *Memory & Cognition* 53:1-5
- Baluch B, Duffy LJ, Badami R et al (2017) A cross-continental study on children's drawings of football players: implications for understanding key issues and controversies in human figure drawings. *Europe's Journal of Psychology* 13(3):452-465
- Bardos AN, Softas BC, Petrogiannis K (1989) Comparison of the Goodenough-Harris and Naglieri's Draw-a-Person scoring systems for Greek children. *School Psychology International* 10(3):205-209
- Barr DJ, Levy R, Scheepers C et al (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* 68(3):255-278
- Bates DMM, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1):1-48
- Brennan RL (2001) Generalizability theory. Springer
- Cauffman E, MacIntosh R (2006) A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument. *Educational and Psychological Measurement* 66(3):502-521
- Clark HH (1973) The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12(4):335-359
- Cox MV (1992) Children's drawings. Penguin
- Cronbach LJ, Gleser GC, Nanda H et al (1972) The dependability of behavioral measurements: theory of generalizability for scores and profiles. Wiley
- Deng J, Dong W, Socher R et al (2009) ImageNet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248-255
- Eitz M, Hays J, Alexa M (2012) How do humans sketch objects? *ACM Transactions on Graphics*, 31(4), 44
- Fabris MA, Lange-Kuttner C, Shiakou M et al (2023) Editorial: children's drawings: evidence-based research and practice. *Frontiers in Psychology* 14:1250556
- Fan JE, Yamins DLK, Turk-Browne NB (2018) Common object representations for visual production and recognition. *Cognitive Science* 42(8):2670-2698

- Fernandez-Fernandez R, Victores JG, Estevez D et al (2019) Quick, stat! A statistical analysis of the Quick, Draw! dataset. arXiv:1907.06417
- Freeman NH (1980) Strategies of representation in young children. Academic Press
- Golomb C (2004) The child's creation of a pictorial world (2nd ed.). Erlbaum
- Gonzalez E (1982) A cross-cultural comparison of the developmental items of five ethnic groups in the southwest. *Journal of Personality Assessment* 46(1):26-31
- Goodenough FL (1926) Measurement of intelligence by drawings. World Book Company
- Goodenough FL, Harris DB (1950) Studies in the psychology of children's drawings: II. 1928-1949. *Psychological Bulletin* 47(5):369-433
- Groves JR, Fried PA (1991) Developmental items on children's human figure drawings: a replication and extension of Koppitz to younger children. *Journal of Clinical Psychology* 47(1):140-148
- Ha D, Eck D (2017) A neural representation of sketch drawings. arXiv:1704.03477
- Hagquist C, Andrich D (2017) Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes* 15:181
- Haladyna TM, Downing SM (2004) Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice* 23(1):17-27
- Harris CR, Millman KJ, DWSJ, Gommers R et al (2020) Array programming with NumPy. *Nature* 585:357–362
- Harris DB (1963) Children's drawings as measures of intellectual maturity. Harcourt, Brace & World
- Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31(2):423–447
- Holland PW, Thayer DT (1988) Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Erlbaum
- Holland PW, Wainer H (eds) (1993) *Differential item functioning*. Erlbaum
- Huey H, Long B, Yang J et al (2022) Developmental changes in the semantic part structure of drawn objects. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*
- Hunter JD (2007) Matplotlib: a 2D graphics environment. *Computing in Science & Engineering* 9(3):90–95
- Hyland P, Rae G (2001) Generalisability and classical test theory analyses of Koppitz's Scoring System for Human Figure Drawings. *British Journal of Educational Psychology* 71(3):369-382
- Judd CM, Westfall J, Kenny DA (2012) Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* 103(1):54–69
- Kane MT (2013) Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50(1):1-73
- Karmiloff-Smith A (1990) Constraints on representational change: evidence from children's drawing. *Cognition* 34(1):57–83
- Kellogg R (1969) *Analyzing children's art*. National Press Books

- Koppitz EM (1968) Psychological evaluation of children's human figure drawings. Grune & Stratton
- Linacre JM (1997) KR-20 / Cronbach alpha or Rasch person reliability: which tells the 'truth'? *Rasch Measurement Transactions*, 11(3), 580-581
- Long B, Fan JE, Chai Z et al (2019) Developmental changes in the ability to draw distinctive features of object categories. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*
- Long B, Fan JE, Frank MC (2018) Drawings as a window into developmental changes in object representations. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*
- Long B, Fan JE, Huey H et al (2024) Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications* 15(1):1191. <https://doi.org/10.1038/s41467-023-44529-9>
- Long B, Wang Y, Christie S et al (2023) Developmental changes in drawing production under different memory demands in a U.S. and Chinese sample. *Developmental Psychology* 59(10):1784–1798
- Lord FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley
- Masters GN (1982) A Rasch model for partial credit scoring. *Psychometrika* 47(2):149-174
- Menon AK, Jayasumana S, Rawat AS et al (2021) Long-tail learning via logit adjustment. *International Conference on Learning Representations*
- Messick S (1989) Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed. pp. 13–103). American Council on Education / Macmillan
- Morra S (2005) Cognitive aspects of change in drawings: a neo-Piagetian theoretical account. *British Journal of Developmental Psychology* 23(3):317-341
- Mukherjee K, Huey H, Lu X et al (2025) Drawings of THINGS: a large-scale drawing dataset of 1,854 object concepts. *Behavior Research Methods* 58:57
- Naglieri JA (1988) Draw A Person: a quantitative scoring system. Psychological Corporation
- Naglieri JA, Maxwell S (1981) Inter-rater reliability and concurrent validity of the Goodenough-Harris and McCarthy Draw-a-Child scoring systems. *Perceptual and Motor Skills* 53(2):343-348
- Naglieri JA, McNeish TJ, Bardos AN (1991) Draw A Person: screening Procedure for Emotional Disturbance. PRO-ED., Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2):133–142
- Paget GW (1932) Some drawings of men and women made by children of certain non-European races. *Journal of the Royal Anthropological Institute* 62:127-144
- Palermo C (2022) Rater characteristics, response content, and scoring contexts: decomposing the determinates of scoring accuracy. *Frontiers in Psychology* 13:937097
- Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3):545–554
- Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
- Peng K, Chen M, Zhou L et al (2024) Differential item functioning in the autism behavior checklist in children with autism spectrum disorder based on a machine learning approach. *Frontiers in Psychiatry* 15:1447080

- Philippsen A, Tsuji S, Nagai Y (2022) Quantifying developmental and individual differences in spontaneous drawing completion among children. *Frontiers in Psychology* 13:783446
- Piaget J (1929) *The child's conception of the world*. Routledge & Kegan Paul
- Radford A, Kim JW, Hallacy C et al (2021) Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763
- Rasch G (1960) *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research
- Rehrig G, Stromswold K (2018) What does the DAP:IQ measure? Drawing comparisons between drawing performance and developmental assessments. *The Journal of Genetic Psychology* 179(1):9–18
- Reynolds CR, Hickman JA (2004) *Draw-A-Person Intellectual Ability Test for Children, Adolescents, and Adults*. PRO-ED., Ricci, C. (1894). *L'arte dei bambini*. *Pedagogical Seminary* 3:302-307
- Saklofske DH, Braun SM (1992) A psychometric study of the Draw-A-Person: a Quantitative Scoring System. *Canadian Journal of School Psychology* 8(1):106-113
- Sangkloy P, Burnell N, Ham C et al (2016) The Sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics* 35(4):119
- Scott LH (1981) Measuring intelligence with the Goodenough-Harris Drawing Test. *Psychological Bulletin* 89(3):483-505
- Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*
- Steward M, Furuya T, Steward O et al (1982) Japanese and American children's drawings of the outside and inside of their bodies. *Journal of Cross-Cultural Psychology* 13(1):87-104
- Swaminathan H, Rogers HJ (1990) Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement* 27(4):361-370
- Swenson CH (1968) Empirical evaluations of human figure drawings: 1957-1966. *Psychological Bulletin* 70(1):20-44
- Thomas GV, Silk AMJ (1990) *An introduction to the psychology of children's drawings*. Harvester Wheatsheaf
- Troncone A, Chianese ADLA, Grasso M et al (2021) Validity of the Draw a Person: a Quantitative Scoring System (DAP:QSS) for clinically evaluating intelligence. *Child Psychiatry & Human Development* 52:728-738
- Valentini NC, Duarte MG, Zanella LW et al (2022) Test of Gross Motor Development-3: item difficulty and item differential functioning by gender and age with Rasch analysis. *International Journal of Environmental Research and Public Health* 19(14):8667
- Valentini NC, Zanella LW (2022) Peabody Developmental Motor Scales-2: the use of Rasch analysis to examine the model unidimensionality, motor function, and item difficulty. *Frontiers in Pediatrics* 10:852732
- Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 17:261–272
- Wang X, Wu Z, Lian L et al (2022) Debiased learning from naturally imbalanced pseudo-labels. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14647-14657

- Westfall J, Kenny DA, Judd CM (2014) Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General* 143(5):2020–2045
- Wijayanto F, Bucur IG, Mul K et al (2023) Semi-automated Rasch analysis with differential item functioning. *Behavior Research Methods* 55:3129-3148
- Willats J (2005) *Making sense of children's drawings*. Erlbaum
- Willcock E, Imuta K, Hayne H (2011) Children's human figure drawings do not measure intellectual ability. *Journal of Experimental Child Psychology* 110(3):444-452
- Williams TOFAM Jr, Eaves RC, Woods-Groves S (2006) The reliability of scores for the Draw-A-Person intellectual ability test for children, adolescents, and adults. *Journal of Psychoeducational Assessment* 24(2):137-144
- Wisniewski JJ, Naglieri JA (1989) Validity of the Draw a Person: a Quantitative Scoring System with the WISC-R. *Journal of Psychoeducational Assessment* 7(4):346-351
- Xu P, Hospedales TM, Yin Q et al (2022) Deep learning for free-hand sketch: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(1):285–312
- Yamins DLK, Hong H, Cadieu CF et al (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111(23):8619–8624
- Yarkoni T (2022) The generalizability crisis. *Behavioral and Brain Sciences* 45:e1
- Zhu R, Kilonzo TN, Zhu LZ et al (2025) Cross-contextual variability in children's early understanding of visual media. *Topics in Cognitive Science*

Figure legends

Fig. 1 Variance decomposition of drawing recognizability

Fig. 2 Estimated object-category effects

Fig. 3 Decision study: reliability as a function of battery length

Fig. 4 Cross-pipeline convergence of category-level estimates

Fig. 5 Short-battery age recovery and single-prompt ordering errors

Fig. 6 Classifier label geometry and apparent object difficulty

Tables

Table 1. Corpus and analytic sample

Property	CLIP pipeline	VGG pipeline
Drawings analyzed	37,770	22,272
Children	8,084	7,032
Object categories	48	48
Age range (years)	2–10	2–10
Mean drawings per child	4.67	3.17
Drawings in covariate model (M4)	35,807	20,980
Drawings scored by both pipelines	22,272	22,272

Table 2. Variance components and the Developmental Displacement Index

Scorer	Model	σ^2C	σ^2K	σ^2e	ICC_child	ICC_cat	R	β_{age} (SE)	DDI (yr)
CLIP	M1	0.926	1.157	3.249	0.174	0.217	1.25	n/a	n/a
CLIP	M2	0.696	1.150	3.245	0.137	0.226	1.65	0.2146 (0.0060)	5.00
CLIP	M4	0.601	1.235	3.197	0.119	0.245	2.06	0.1521 (0.0069)	7.31
VGG	M1	0.522	0.597	2.389	0.149	0.170	1.14	n/a	n/a
VGG	M2	0.372	0.595	2.382	0.111	0.178	1.60	0.1750 (0.0058)	4.41
VGG	M4	0.304	0.634	2.361	0.092	0.192	2.08	0.1106 (0.0066)	7.20

Note. $R = \sigma^2K / \sigma^2C$. $DDI = \sigma K / \beta_{age}$, the number of years of chronological development equivalent to one standard deviation of object-category difficulty. M1 = unconditional; M2 = + age; M4 = + covariates.

Table 3. Decision study

Prompts (n_k)	CLIP $E\rho^2$	CLIP Φ	CLIP gap	VGG $E\rho^2$	VGG Φ	VGG gap
1	0.177	0.137	0.040	0.135	0.111	0.024
2	0.300	0.241	0.060	0.238	0.200	0.038
4	0.462	0.388	0.074	0.385	0.334	0.051
8	0.632	0.559	0.073	0.556	0.500	0.056
12	0.720	0.655	0.065	0.652	0.600	0.052
16	0.774	0.717	0.057	0.714	0.667	0.048
24	0.837	0.792	0.046	0.790	0.750	0.039
32	0.873	0.835	0.038	0.833	0.800	0.033
48	0.911	0.884	0.028	0.882	0.857	0.025

Note. $E\rho^2$ is the relative (norm-referenced) coefficient; Φ is the absolute (criterion-referenced) coefficient. The gap between them is the reliability cost attributable to stimulus dependence. Prompts required for $\Phi \geq 0.80$: 26 (CLIP), 32 (VGG); for $\Phi \geq 0.90$: 57 and 72.

Table 4. Cross-pipeline convergence

Quantity	Pearson r	p	Spearman ρ
Drawing-level recognizability score	0.426	< .001	0.425

Quantity	Pearson r	p	Spearman ρ
Category baseline effect	0.205	.166	n/a
Category baseline effect, attractor-adjusted	0.396	.006	n/a
Category-specific age slope	0.734	< .001	n/a

Table 5. Single-prompt age-order violation rates

Scorer	Scoring rule	Violation rate	95% CI
CLIP	Raw score	0.347	[0.325, 0.369]
CLIP	Category-centered score	0.331	[0.311, 0.352]
VGG	Raw score	0.343	[0.324, 0.362]
VGG	Category-centered score	0.329	[0.307, 0.349]

Note. The proportion of randomly paired children aged 7–10 versus 2–4 for whom the older child received the lower score on a single randomly assigned prompt. Chance = 0.500.

Table 6. Classifier label geometry

Quantity	CLIP	VGG
Split-half reliability of attractor strength $A(k)$	1.000	0.966
$A(k)$ to category effect, r	0.737	0.288
$A(k)$ to category effect, p	< .001	0.049
Category-effect variance explained by $A(k)$	54.3%	8.3%
Largest probability sink	snail (0.0775)	octopus (0.0205)
Cross-pipeline agreement of $A(k)$, r	0.040	0.040
Cross-pipeline agreement of category effect, raw	0.205	0.205
Cross-pipeline agreement, attractor-adjusted	0.396	0.396

Note. $A(k)$ is computed exclusively from drawings of categories other than k . Improvement in cross-pipeline agreement after adjustment: $\Delta r = 0.190$, 95% bootstrap CI [0.013, 0.344]. Smallest correlation detectable at 80% power with 47 categories: $r = 0.399$.

Data availability

The datasets analysed during the current study are publicly available in the KidDraw repository, <https://github.com/brialorelle/kiddraw>. The specific files used are data/preprocessed_data/merged_clip_class_and_meta.csv, data/preprocessed_data/merged_vgg_class_and_meta.csv, data/clip_outputs/fullset/clip_output.csv, and data/compiled_classifications/C_0.1_T_0.1batchtotal_232.csv. No new data were generated for this study.

Code availability

All custom code required to reproduce the analyses is provided as Supplementary Software: the restricted maximum likelihood implementation (reml.py), the analysis scripts (analysis_01_models.py through analysis_06_geometry_clean.py), and the figure and table generation scripts. Analyses were carried out in Python using NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), scikit-learn (Pedregosa et al. 2011), and Matplotlib (Hunter 2007). On acceptance the code will be deposited in a repository with a persistent identifier.

Author contributions

R.S.M. conceived the study, designed and conducted the analyses, interpreted the results, and wrote the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

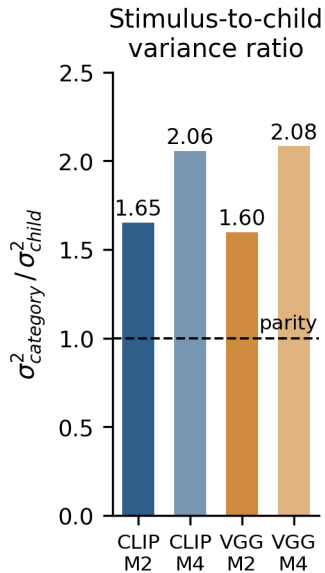
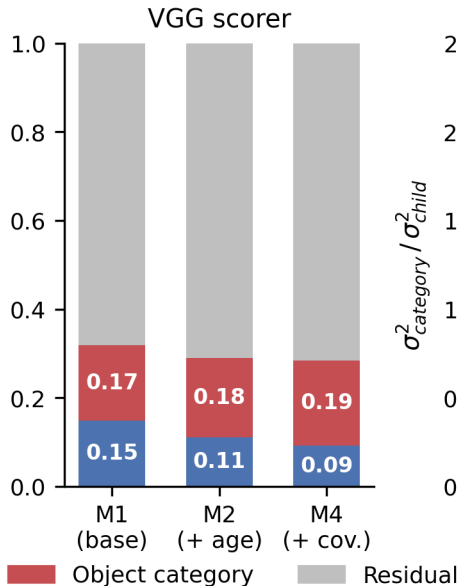
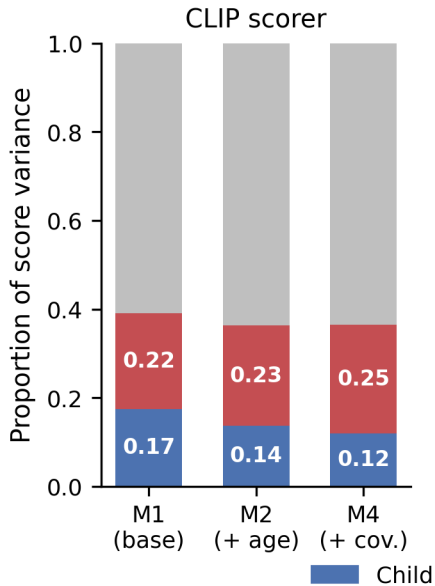
The author declares no competing interests.

Ethics declarations

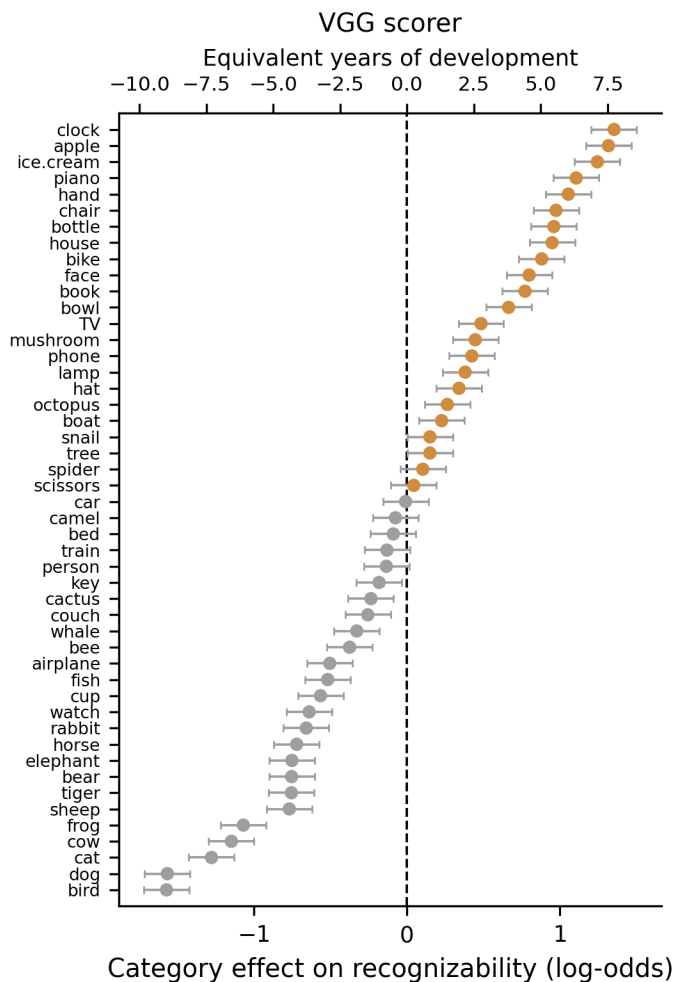
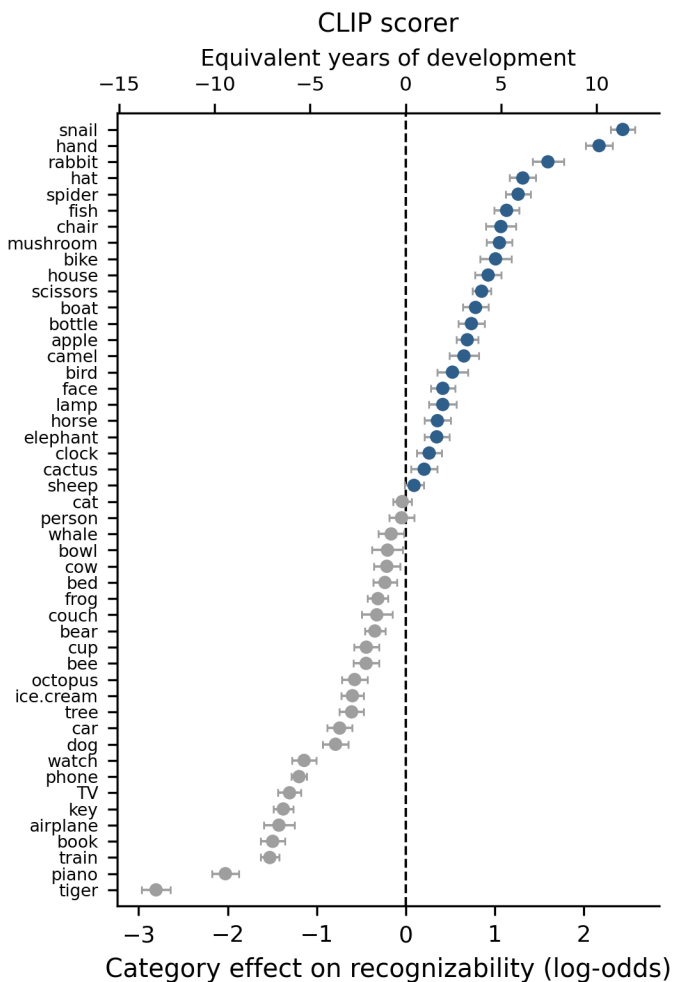
Ethical approval was not required for this study. The research consists of secondary analysis of an existing, fully de-identified, publicly available dataset. No human participants were recruited by the author, no participants were contacted, and no identifying information was accessed; the archived data comprise drawings, an age in whole years, and an anonymous session identifier. The original data collection was approved by the Institutional Review Board at Stanford University under Protocol 43992, Development of Children's Drawing Abilities, and a waiver of consent was completed by a parent or guardian before each session (Long et al. 2024).

Additional information

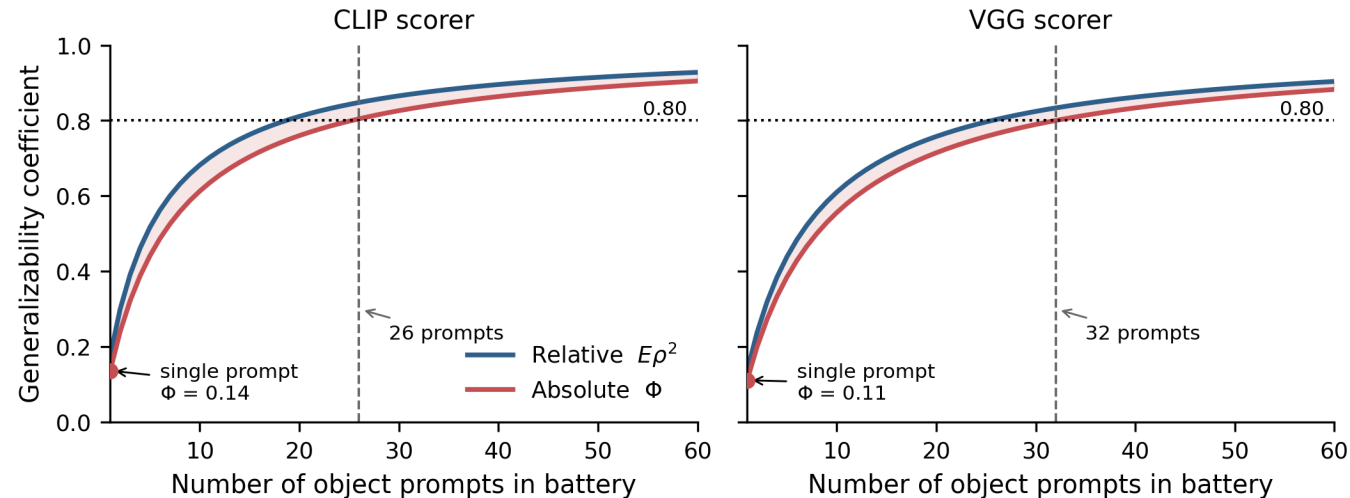
Supplementary Information accompanies this paper. Supplementary Fig. S1 reports validation of the restricted maximum likelihood implementation against simulated data with known parameters. Supplementary Tables S1 and S2 report the covariate-adjusted fixed effects and the short-battery age-recovery results.



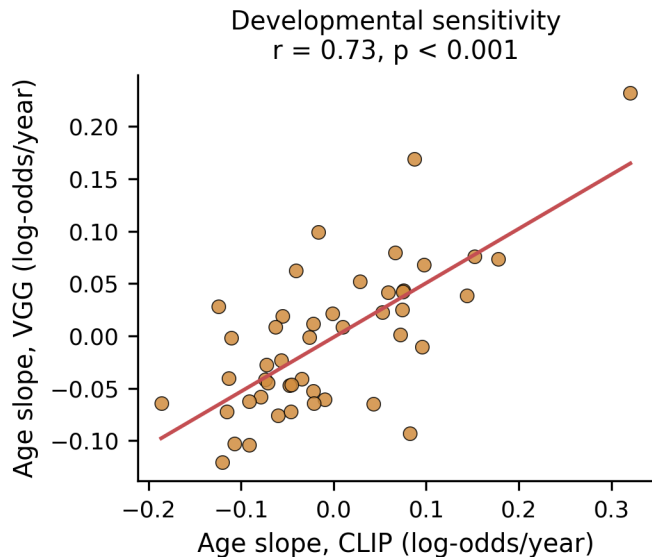
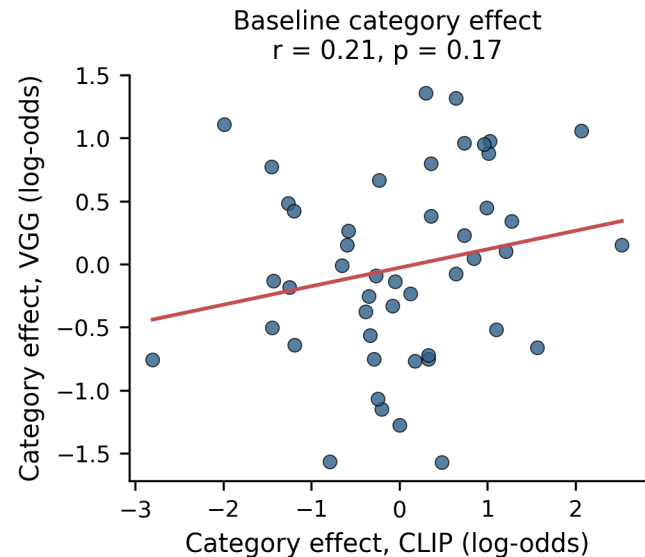
Estimated object-category effects (positive = more recognizable)

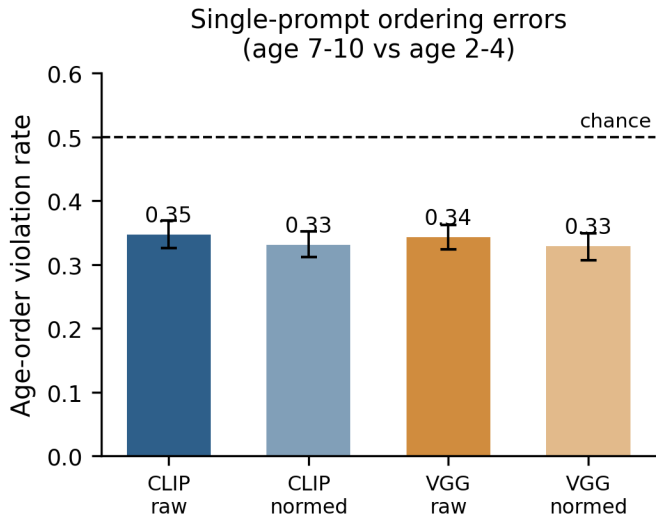
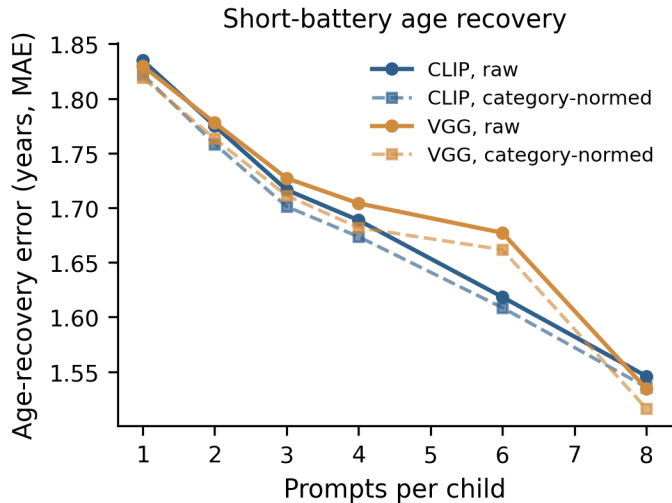


Decision study: reliability as a function of battery length

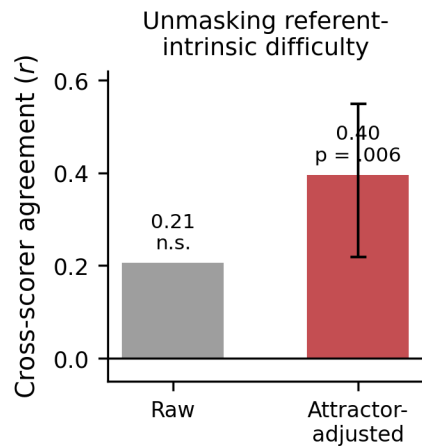
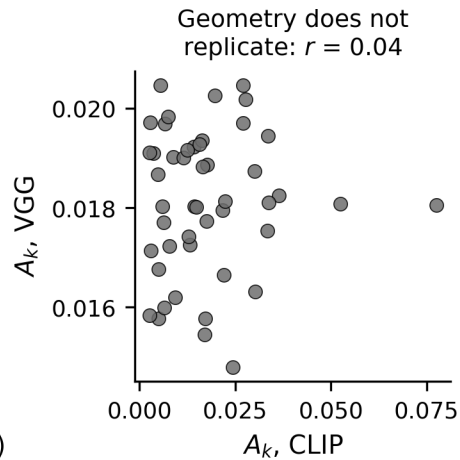
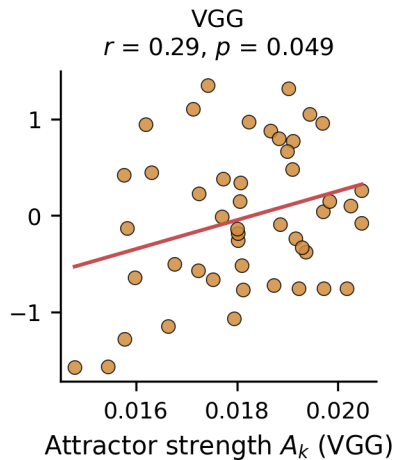
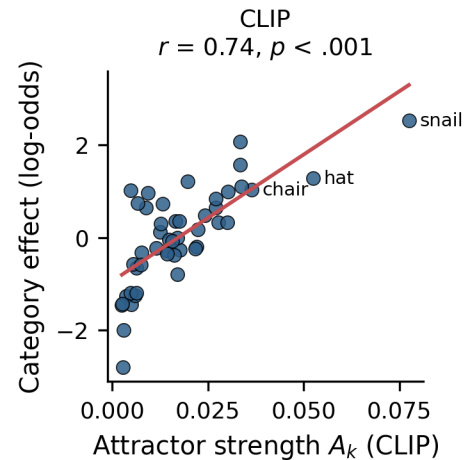


Difficulty is scorer-specific; age sensitivity is not





Classifier label geometry, not the object, drives most apparent category difficulty



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [MadaniSupplementaryInformation.docx](#)
- [SupplementarySoftware.zip](#)