Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Two lenses on one score: geometric and psychometric evidence that classifier label structure contaminates machine-graded children's drawings

Rafat Saleh Madani

rsmadani@uqu.edu.sa

Umm Al-Qura University

Article

Keywords: educational measurement, construct-irrelevant variance, automated scoring, children's drawings, item response theory, representational geometry, algorithmic fairness

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10842930/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Two lenses on one score: geometric and psychometric evidence that classifier label structure contaminates machine-graded children's drawings

Rafat Saleh Madani

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

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

ORCID: 0000-0001-7945-4459

Abstract

Machine scoring is increasingly used to assess children's drawings at scale, and the resulting scores are read as measurements of developmental ability. Much of the variance in such scores is shown to belong to the scoring model's label space rather than to the child. Sinkhorn matrix balancing is repurposed here as a validity diagnostic. The column scaling of a scorer's aggregate posterior mass matrix, termed **attractor strength**, is an additive logit offset, and therefore enters an item-response model directly as an item parameter. Across 37,770 drawings by 8,084 children aged 2–10, scored by two independent classifiers, the analysis finds: (i) attractor strength far exceeds an accuracy-preserving label-exchangeability null (z = 24.5 supervised, z = 428.3 zero-shot; $p < .001$); (ii) between-category variance exceeds between-child variance roughly sixfold, and attractor strength alone explains 78% of it, making 67% of systematic score variance construct-irrelevant; (iii) attractor structure does not replicate across scorers (r = −.16), with 25 of 48 categories reversing role; (iv) benchmarked against 4,372 child recognisers judging the same drawings under an identical four-alternative choice, the supervised scorer's label geometry is human-comparable while the zero-shot scorer's exceeds it 2.4–12.6×; and (v) correction steepens the estimated developmental trajectory by 27% and improves category-level agreement with human recognition, though it cannot improve individual-drawing measurement. Machine and human scorers weight production effort equivalently, localising the contamination to the label space. Automating a rater does not inherit that rater's biases; it substitutes new ones.

Keywords: educational measurement; construct-irrelevant variance; automated scoring; children's drawings; item response theory; representational geometry; algorithmic fairness

1. Introduction

Children draw prolifically, and adults have been reading those drawings as evidence of mind for a century. Goodenough's (1926) Draw-A-Man test proposed that the number and organization of features in a child's human figure drawing indexes conceptual maturity; Harris (1963) revised and restandardised it; Koppitz (1968) and Naglieri (1988) extended the approach to emotional and intellectual screening. The appeal has always been the same: drawing is nearly universal, requires little language, takes minutes, and appears to offer a culture-fair window onto cognitive development (Goodenough & Harris, 1950; Scott, 1981; Cox, 1993; Jolley, 2010).

That appeal has never been matched by uncontested validity evidence. Correlations between human figure drawing scores and standardised intelligence measures range from substantial to negligible depending on sample and scoring system (Abell et al., 2001; Ter Laak et al., 2005; Van Meerbeke et al., 2011), and careful comparisons against Wechsler scales have concluded that these instruments should not be used as proxies for intellectual ability at all (Willcock et al., 2011; Imuta et al., 2013). Reliability across raters and scoring systems is likewise variable (Naglieri & Maxwell, 1981; Strommen et al., 1987; Baraheni et al., 2018). The field's core difficulty has been that scoring drawings by hand is slow, subjective, and hard to standardize.

Machine scoring changes the arithmetic. Long et al. (2024) collected more than 37,000 drawings from 8,084 children at a museum kiosk and scored every one automatically by passing deep network features to a linear classifier over 48 object categories, recovering a clean developmental trajectory in recognizability. This is a genuine methodological advance: it makes drawing assessment scalable, reproducible, and free of rater drift. It also silently changes what the score *is*. A hand score is a judgment against a rubric. A machine score is the output of a classifier's decision over a label space, and that label space has structure of its own.

1.1 Machine scores are measurements, and measurements have validity requirements

Once a machine output is interpreted as an index of a child's ability, it enters the domain of measurement, where accuracy is necessary but nowhere near sufficient. Messick's (1989, 1995) unified framework identifies two principal threats to construct validity: construct underrepresentation, where the assessment is too narrow, and **construct-irrelevant variance** (CIV), where scores carry systematic variance unrelated to the intended construct. CIV is not noise. It is reliable, reproducible, and precisely for that reason dangerous: it inflates or deflates scores for reasons external to the attribute being measured (Downing, 2002; Kane, 2013; AERA, APA, & NCME, 2014).

The automated scoring literature has taken this seriously, though almost entirely for text. Work on automated essay scoring has developed fairness frameworks covering overall score accuracy, overall score differences, and conditional score differences across groups (Loukina et al., 2019; Johnson & McCaffrey, 2023), documented differential functioning across learner groups (Litman et al., 2021; Schaller et al., 2024; Chen et al., 2026), and asked what validity arguments look like when scores come from generative models

(Rodriguez et al., 2019; Lottridge et al., 2023; Kumar & Boulanger, 2021; Casabianca et al., 2025). Recent work extends this to large language models as scorers, finding that zero-shot systems can appear "fair" simply because they rarely assign extreme scores (Chen et al., 2026), and that construct validity for model-based measurement requires more than benchmark agreement (Freiesleben, 2026; Lin, 2025; Bean et al., 2025).

What this literature shares is a *person-side* conception of bias. The question asked is whether scores differ unfairly across groups of test takers defined by demographic or linguistic attributes. That framing is essential and incomplete. It cannot detect a bias that lives in the *item* side of the design: specifically, in the geometry of the label space against which every response is scored.

1.2 Label spaces have geometry

A multi-way classifier does not evaluate a drawing in isolation; it distributes probability mass across a fixed set of labels. Those labels are not exchangeable. Some occupy large, centrally located regions of representational space and absorb ambiguous inputs; others are narrow and absorb almost nothing. In the vision–language setting this is well documented under other names: CLIP inherits a skewed label prior from web-scale pretraining and benefits measurably from post-hoc logit adjustment (Radford et al., 2021; Agarwal et al., 2021; Alabdulmohsin et al., 2024; Shin et al., 2024; Wang et al., 2024). Related work shows severe class collapse and extreme prompt sensitivity when such models are applied zero-shot to educational observation (Goyal et al., 2026), and that debiasing at the representation or prompt level materially changes predictions (Chuang et al., 2023; Dehdashtian et al., 2024; Seth et al., 2023; Kim et al., 2025).

Separately, a large literature in computational neuroscience has developed tools for characterising representational geometry (the arrangement of stimulus representations in a system's internal space) and for comparing that geometry across systems (Kriegeskorte et al., 2008a, 2008b; Kriegeskorte & Kievit, 2013; Kriegeskorte & Wei, 2021; Nili et al., 2014; Kornblith et al., 2019). A recurring finding is that geometry is model-specific in ways that performance metrics conceal: networks differing only in random seed learn measurably different representational geometries despite indistinguishable accuracy (Mehrer et al., 2020), diverse architectures predict cortex about equally well after fitting (Storrs et al., 2021), and correspondence between networks and brains breaks down at higher levels (Xu & Vaziri-Pashkam, 2021).

These two literatures have not been brought to bear on educational measurement. Doing so yields a specific, testable hypothesis: **if a machine scorer's label space contains attractors, then the apparent difficulty of an assessment item is partly a property of the scorer rather than of the child or the task**, precisely Messick's construct-irrelevant variance, arriving from the item side rather than the person side.

1.3 The present study

We formalize this as **attractor strength**, obtained by Sinkhorn-balancing the aggregate posterior mass matrix of a scorer (Sinkhorn, 1964; Sinkhorn & Knopp, 1967; Knight, 2008). The construction yields a label-level parameter in the units of a logit, which lets us carry a

quantity computed from classifier geometry directly into an explanatory item response model (De Boeck & Wilson, 2004; De Boeck, 2008; Rijmen et al., 2003). That is the paper's two lenses: a geometric estimator and a psychometric decomposition, joined by an exact algebraic identity rather than an analogy.

We address four questions.

- **RQ1.** Does the off-target posterior mass of a machine scorer distribute across labels more unevenly than label-exchangeability predicts?
- **RQ2.** How much of the apparent difficulty of a drawing category is attributable to label geometry rather than to the concept or the child?
- **RQ3.** Does attractor structure replicate across scoring architectures, and how does it compare to the label geometry of human recognisers performing the identical task?
- **RQ4.** Does correcting for label geometry change substantive developmental conclusions?

---

2. Data

We use the *kiddraw* corpus (Long et al., 2024), collected at a free-standing touchscreen kiosk in a children's museum and released under CC BY 4.0. Children aged 2–10 were verbally prompted to draw each of 48 object categories, spanning commonly drawn categories (*face*, *cat*) and rarely drawn ones (*octopus*, *piano*).

Two machine scorers distributed with the corpus are analysed.

- **VGG-19 linear probe**: a multinomial logistic classifier ($C = 0.1$) over VGG-19 features; supervised and in-domain. Applied to the age × category balanced subset ($N = 22{,}272$).
- **CLIP zero-shot**: a vision–language classifier scoring each drawing against 48 textual prompts; no in-domain training (Radford et al., 2021). Available for the full corpus ($N = 37{,}770$). We analyse the full set and a balanced subsample (22 drawings per age × category cell; $N = 9{,}504$).

Both emit a full row-stochastic posterior over all 48 labels for every drawing.

For the human benchmark (§4.4) we use the corpus's recognition games, in which 4,372 child recognisers made 46,688 four-alternative forced-choice judgments about 2,160 of these drawings. The corpus and scoring configuration are summarised in Table 1.

Table 1. Corpus and scoring configuration.

Property	VGG-19 probe	CLIP (balanced)	CLIP (full)
Drawings (N)	22,272	9,504	37,770
Children (N)	7,032	5,067	8,084
Categories (K)	48	48	48

Property	VGG-19 probe	CLIP (balanced)	CLIP (full)
Age range (years)	2–10	2–10	2–10
Age × category balanced	yes	yes (22/cell)	no
Chance accuracy	0.0208	0.0208	0.0208

Table 1 makes the design constraint visible: the VGG-19 set is balanced by age and category by construction, and the CLIP results are reproduced on a subsample balanced the same way, so no comparison reported below is confounded by unequal cell sizes. Both scorers face the same 48-way decision, and chance performance is 0.0208 throughout.

Data integrity. Recomputing the scoring decision from the released posteriors reproduces the archived accuracy exactly (0.3353), confirming that posterior matrix and scoring labels are mutually consistent. One label-name mismatch (*ice cream* / *ice.cream*) was identified and corrected before analysis.

---

3. Method

3.0 The idea in plain terms

Before the formalism, the intuition, because everything that follows is one simple observation made precise.

Suppose a scorer must decide, for each drawing, which of four things a child drew: a *bottle*, a *cup*, a *hat*, or a *lamp*. A child draws a cup badly. The scorer spreads its confidence across the four options and picks the highest. So far, nothing is wrong.

Now consider what happens across *all* the drawings. Every drawing emits exactly one unit of confidence, spread over the four labels. So four hundred drawings emit four hundred units in total, and if the label set were even-handed, each label would absorb roughly a quarter of them. That is the whole idea of neutrality: **a fair label absorbs about as much confidence as the drawings of its own category emit.**

Real scorers are not even-handed. In our data one label under CLIP absorbs **four times** its fair share. A label like this behaves as a sink: ambiguous drawings of *any* category drift into it. And the consequence lands on the children, not on the label. A child who drew a cup competently may be scored wrong because a neighboring label was busy vacuuming up uncertain drawings: a failure that has nothing to do with that child, that drawing, or the concept of cups.

Attractor strength is simply the number that says how much each label over- or under-absorbs. Computing it means asking: what correction would each label need so that every label absorbs its fair share? That question has a classical answer: Sinkhorn balancing (§3.2). The correction it returns is exactly the attractor strength.

Two consequences make this worth doing, and they are the paper's two lenses.

It is a measurement quantity, not just a diagnostic. The correction turns out to be an additive adjustment on the log-odds scale (§3.3), which is the same scale on which test theory measures item difficulty. So a number computed from a classifier's internal geometry can be placed directly into an item-response model beside a child's ability, and the two can be compared on equal terms. That comparison is what §4.2 reports.

It belongs to the scorer, not to the world. If attractor strength measured something real about drawing cups, two different scorers would agree about it. They do not (§4.3). And if it measured something about children, it would shift as we look at different age groups. It does not (§4.6). It is a property of the label space, which is precisely why variance it explains is construct-irrelevant in Messick's sense.

The remainder of this section makes each step exact.

3.1 The aggregate posterior mass matrix

Let $P \in [0,1]^{N \times K}$ be the posterior matrix of a machine scorer, rows summing to one, and $y_i \in \{1, \dots, K\}$ the target category of drawing i. Define the **aggregate mass matrix** $M \in \mathbb{R}^{K \times K}$:

$$M_{cj} \;=\; \frac{1}{N_c} \sum_{i\,:\,y_i = c} P_{ij}\,, \qquad N_c = |\{i : y_i = c\}|$$

Each row is the mean posterior emitted by drawings of true category c and sums to one by construction. Column sums are unconstrained:

$$\kappa_j \;=\; \sum_{c=1}^{K} M_{cj}$$

is the aggregate mass **absorbed** by label j.

3.2 Label neutrality and attractor strength

> **Definition 1 (Label neutrality).** A scorer is *label-neutral* on a corpus if M is doubly stochastic, i.e. $\kappa_j = 1$ for all j.

Under neutrality every label absorbs exactly as much mass as it emits. Departures mean some labels are net absorbers and others net donors: the label space contains sinks.

By the Sinkhorn–Knopp theorem (Sinkhorn & Knopp, 1967; see also Sinkhorn, 1964; Menon, 1967; Brualdi et al., 1966), for any nonnegative matrix with total support there exist positive diagonal matrices $\mathrm{diag}(r)$, $\mathrm{diag}(s)$ with

$$B \;=\; \mathrm{diag}(r)\, M\, \mathrm{diag}(s)$$

doubly stochastic, r and s unique up to a reciprocal positive scalar. Convergence is linear for fully indecomposable matrices (Soules, 1991; Knight, 2008), and the iteration is equivalent to iterative proportional fitting (Deming & Stephan, 1940; Stephan, 1942; Rüschendorf, 1995).

Definition 2 (Attractor strength). For label j,

$$\alpha_j \;=\; -\log s_j \;-\; \frac{1}{K}\sum_{k=1}^{K}(-\log s_k)$$

in nats, centred so $\sum_j \alpha_j = 0$.

$\alpha_j > 0$ marks an **attractor**; $\alpha_j < 0$ a **repeller**.

3.3 The bridge between the two lenses

The correction implied by s acts multiplicatively on the posterior and therefore *additively on the logit*:

$$\tilde{P}_{ij} \;=\; \frac{P_{ij}\, e^{-\alpha_j}}{\sum_k P_{ik}\; e^{-\alpha_k}} \qquad \Leftrightarrow \qquad \text{logit}\tilde{P}_{ij} \;=\; \text{logit}P_{ij} - \alpha_j + c_i$$

Attractor strength is thus already in the units of an item-difficulty parameter. This is the formal bridge that permits a quantity computed from classifier geometry to enter an item response model as an item covariate, and it connects the construction to logit-adjustment methods used to correct label priors in zero-shot vision (Shin et al., 2024; Wang et al., 2024; Alexandari et al., 2020).

We call $\tilde{P}$ the **geometry-corrected** posterior and $\tilde{S}_i = \mathbb{1}\left[\text{argmax}_j \tilde{P}_{ij} = y_i\right]$ the geometry-corrected score. A scalar summary of distortion is

$$\text{TV} \;=\; \frac{1}{2K}\sum_{j=1}^{K}\left|\kappa_j - 1\right|$$

TV is the mean absolute departure of the column masses from neutrality, halved so that it lies in [0,1). It is zero for a perfectly label-neutral scorer and grows as labels diverge in how much mass they absorb, giving a single number by which scorers can be compared.

3.4 An accuracy-preserving null

The scientific question is not whether a scorer has label structure (trivially it does) but whether, *given its discriminability and per-drawing confidence*, off-target mass is distributed more unevenly than label-exchangeability implies.

> **Null model (rank-preserving label permutation).** For each drawing i independently: hold P_{i,y_i} fixed and randomly permute the remaining $K-1$ posterior values across the remaining $K-1$ labels.

This **exactly preserves**, for every drawing, the target-label probability, the full multiset of posterior values (hence entropy, margin, and maximum), and whether the drawing is scored correct. Overall accuracy and the entire developmental accuracy curve are invariant under the null; only the identity of which labels absorb off-target mass is destroyed. Any excess in α is therefore attributable to label geometry alone. We use 1,000 permutations.

Confidence intervals on α_j come from a child-level cluster bootstrap (500 resamples), respecting the nesting of drawings within children.

3.5 The psychometric lens

We fit a crossed random-effects (explanatory Rasch) model to the binary machine score S_{ic} of child i on category c (Rasch, 1960/1980; De Boeck & Wilson, 2004; De Boeck, 2008; Van den Noortgate et al., 2003):

$$\text{logitPr}(S_{ic} = 1) \;=\; \mu + \theta_i + \delta_c, \qquad \theta_i \sim \mathcal{N}\left(0, \sigma_\theta^2\right), \quad \delta_c \sim \mathcal{N}\left(0, \sigma_\delta^2\right)$$

Here μ is the grand mean of the logit scale, θ_i is the ability of child i, and δ_c is the easiness of category c. The two variance components are the quantities of interest: σ_θ^2 is the score variance attributable to differences between children, which is the construct the assessment intends to measure. σ_δ^2 is the variance attributable to differences between categories, which it does not. Their ratio is reported in §4.2.

Estimation is by penalized alternating Newton updates with curvature-corrected variance components, converging in 460 iterations and recovering marginal category log-odds at r = 0.967 and child mean scores at ρ = 0.834 (appropriately shrunk given a median of 3 drawings per child).

The structural stage regresses category easiness on attractor strength:

$$\delta_c \;=\; \gamma_0 + \gamma_1 \alpha_c + \varepsilon_c$$

where γ_1 measures how strongly a category's apparent easiness tracks the attractor strength of its label, and ε_c is the residual easiness not accounted for by label geometry. Because α_c is a property of the scorer's label space rather than of the child or the concept, the share of systematic score variance it accounts for is construct-irrelevant:

$$\text{CIV} \;=\; \frac{R^2(\delta, \alpha)\;\sigma_\delta^2}{\sigma_\theta^2 + \sigma_\delta^2}$$

The expression is the product of two proportions. The first, $R^2(\delta, \alpha)$, is the share of between-category variance in apparent difficulty that attractor strength explains. The second, $\sigma_\delta^2/\left(\sigma_\theta^2 + \sigma_\delta^2\right)$, is the share of total systematic score variance that sits on the category facet rather than the child facet. Their product is an estimate of the share of systematic variance in the machine score associated with the scorer's label geometry rather than with the child or the concept. It should be read as an estimate under the additive model of §3.5, not as an exact partition.

3.6 Relation to prior work on matrix balancing

The estimator in §3.2 is not new, and no claim of methodological novelty is made for it. Balancing a contingency table by iterative proportional fitting dates to Deming and Stephan (1940) and Stephan (1942); its application to confusion matrices is standard in remote sensing (Congalton, 1991; Hardin and Shumway, 1997) and has been used in speaker verification (Nagineni and Hegde, 2010). Most directly, Erbani et al. (2025) apply IPF bi-

normalisation to classifier confusion matrices in order to separate *class similarity* from *distribution bias*, establish the $D^l M D^r$ scaling form via the RAS method, and give it a latent-space geometric interpretation. On the machine-learning side, correcting a skewed label prior by an additive logit offset is likewise established (Alexandari et al., 2020; Shin et al., 2024; Wang et al., 2024), and the accuracy gains reported in §4.5 are of a kind that literature already documents.

Three differences none the less separate the present use of the estimator from that prior work, and they are what the contribution rests on.

The matrix is different. Erbani et al. balance the *confusion matrix*: hard counts accumulated from arg-max predictions, $M = \sum_k E_{y_k \hat{y}_k}$. The object balanced here is the *aggregate posterior mass matrix*, in which every drawing contributes its full distribution over labels rather than a single vote. The distinction matters for measurement: a drawing that places 0.34 on the target and 0.33 on an attractor is scored identically to one placing 0.99 on the target under a confusion matrix, but the two carry very different information about how close the scorer came to a different verdict.

The estimand differs. For Erbani et al. the diagonal scalings are a nuisance to be divided out so that class similarity becomes visible; their bi-normalisation is explicitly designed to satisfy *class distribution invariance*, $\mathrm{bi}(AMB) = \mathrm{bi}(M)$ for any positive diagonal A, B, that is, to be blind by construction to precisely the quantity measured here. Their residual is our nuisance and our estimand is their nuisance. The two analyses are complementary rather than overlapping.

The attribution is different, and the design forces it. Erbani et al. attribute column inflation to *distribution bias* arising from imbalanced training or test sets. That explanation is unavailable here: the VGG-19 corpus is balanced by age and category by construction, and the CLIP results are reproduced on a balanced subsample (§4.1–4.5). Column inflation in a balanced test set cannot be test-set imbalance. It is more plausibly attributable either to the scorer's training prior (for a zero-shot vision–language model, the label prior of web-scale pretraining) or in the placement of decision boundaries itself. Either way it is a property of the scorer that the assessment designer did not choose and would not ordinarily inspect, which is what makes it a validity concern rather than a data-collection artefact.

We also introduce the accuracy-preserving null of §3.4, which we believe is new: existing work on label bias asks whether correction improves accuracy, not whether the observed geometry exceeds what label-exchangeability would produce at fixed discriminability.

---

4. Results

4.1 Machine scorers are far from label-neutral (RQ1)

Both scorers depart sharply from neutrality, with magnitude depending strongly on architecture. Under VGG-19 the most absorbing label takes 1.17× its neutral share, the

most donating 0.72×. Under CLIP the range is 0.14× to 4.11×: the label *snail* absorbs over four times the aggregate posterior mass it emits (Fig. 5).

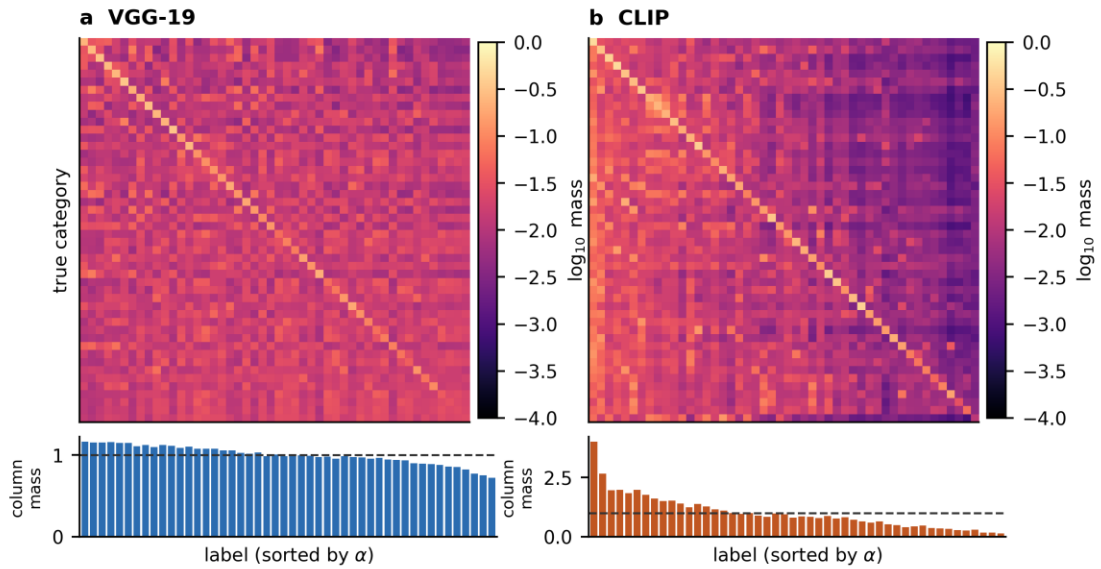


Fig. 5 Aggregate posterior mass matrices.

In Fig. 5, rows are true category and columns assigned label, both ordered by attractor strength, with colour giving $\log_{10}$ mean posterior mass. Panel **a** shows VGG-19 and panel **b** CLIP on the balanced subsample; the lower panels give the column mass absorbed by each label, with the dashed line marking neutrality.

Against the accuracy-preserving null, observed spread lies far outside the null distribution (Fig. 2): VGG-19, SD(α) = 0.122 vs. null mean 0.087, z = 24.5; CLIP, 0.798 vs. 0.114, z = 428.3; both $p < 0.001$ (0/1,000 permutations). Total-variation imbalance shows the same pattern (z = 12.1 and 294.6). Cluster-bootstrap intervals exclude zero for 79.2% (VGG-19) and 93.8% (CLIP) of labels.

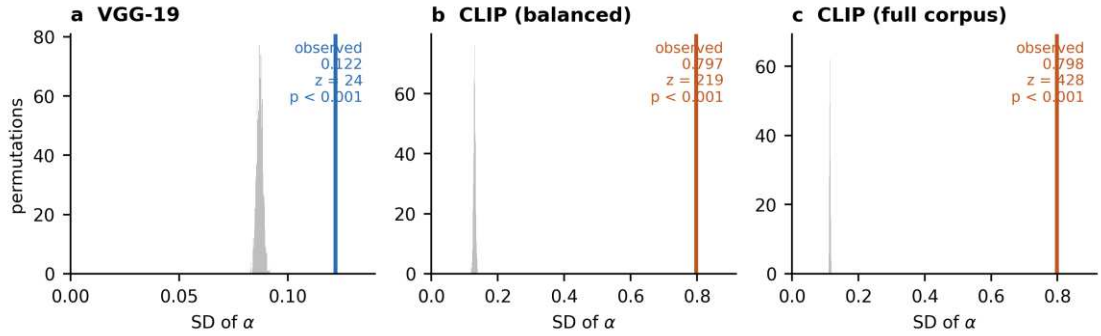


Fig. 2 Observed label geometry against the permutation null.

In Fig. 2, panels **a**, **b** and **c** show VGG-19, CLIP on the balanced subsample, and CLIP on the full corpus respectively. Grey histograms are the null distribution over 1,000 rank-preserving permutations; the coloured line marks the observed value.

The attractors are interpretable. Under VGG-19 they are simple closed contours: *bottle*, *book*, *chair*, *apple*, *bowl* and *clock*. Repellers are detail-bearing animate categories: *bird*, *dog*, *cat*, *cow* (Fig. 1a). This is consistent with under-specified drawings being absorbed into generic-shape labels.

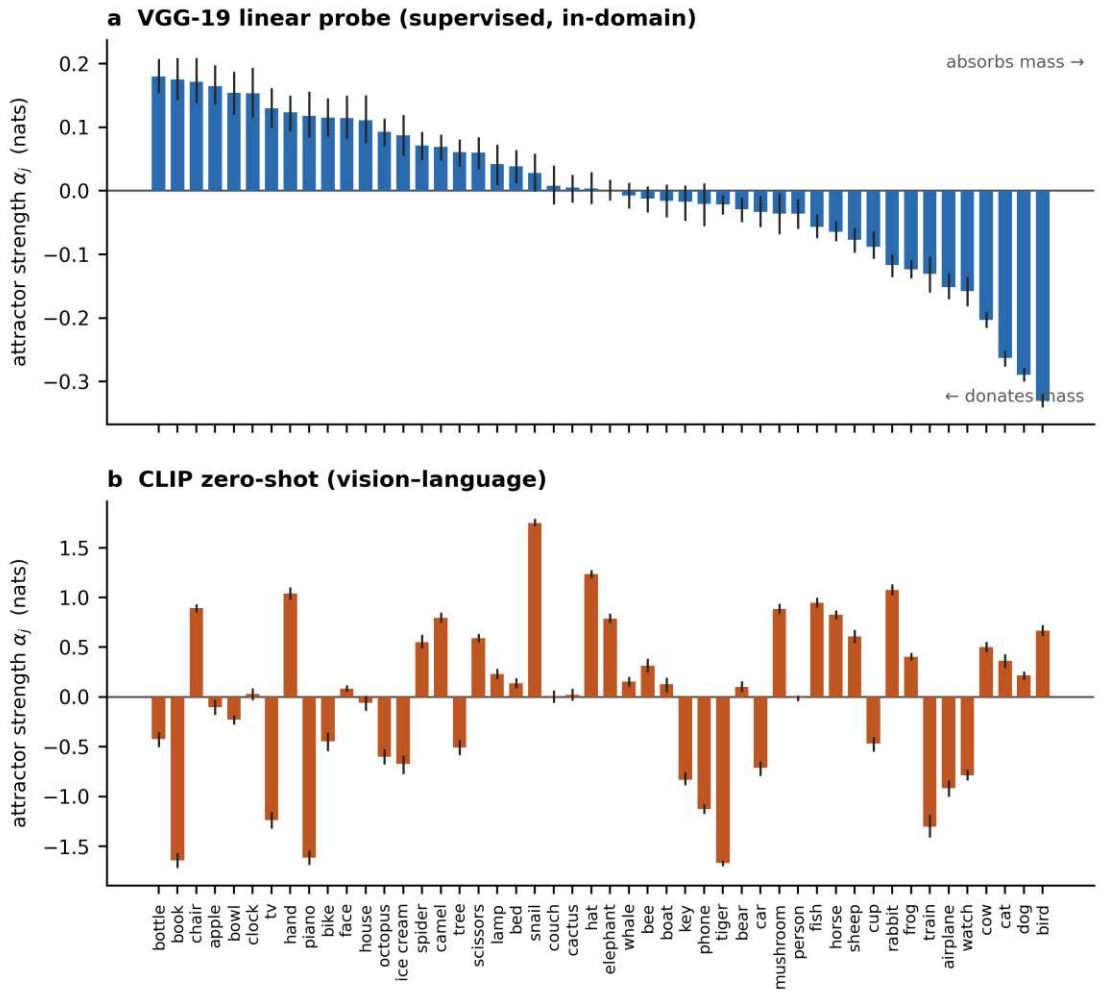


Fig. 1 Attractor strength across the 48-label inventory.

In Fig. 1, panel **a** shows the VGG-19 linear probe and panel **b** the CLIP zero-shot scorer. Error bars are 95% child-level cluster-bootstrap confidence intervals from 500 resamples, and labels are ordered by VGG-19 attractor strength in both panels.

4.2 Category difficulty is largely label geometry (RQ2)

The principal results for all three scorer configurations are reported in Table 2.

Table 2. Principal results.

	VGG-19 probe	CLIP (balanced)	CLIP (full)
Accuracy (raw)	0.3353	0.3162	0.2785

	VGG-19 probe	CLIP (balanced)	CLIP (full)
Accuracy (geometry-corrected)	0.3390	0.3744	0.3410
Δ accuracy	+0.0037	+0.0582	+0.0625
Scores flipped (%)	2.67	11.21	10.73
SD(α)	0.122	0.797	0.798
SD(α) under null	0.087	0.130	0.114
z	24.5	218.9	428.3
Column mass, min–max	0.72–1.17	0.14–4.02	0.14–4.11
σ_θ (child)	0.532	0.392	0.442
σ_δ (category)	1.291	1.026	1.072
$\sigma_\delta^2/\sigma_\theta^2$	5.88	6.85	5.87
$R^2(\delta \sim \alpha)$	0.784	0.524	0.528
CIV share	**0.670**	**0.457**	**0.451**
σ_δ reduction after correction	−26.2%	−45.3%	−46.9%
Developmental slope (raw)	0.0290	0.0158	0.0156
Developmental slope (corrected)	0.0292	0.0180	0.0199
Slope change	**+0.8%**	**+14.3%**	**+27.1%**

Two findings dominate. First, **between-category variance exceeds between-child variance by roughly a factor of six** ($\sigma_\delta^2/\sigma_\theta^2$ = 5.88 for VGG-19). In an instrument meant to index child development, the item facet carries six times the systematic variance of the person facet. Second, **attractor strength alone explains 78.4% of that category variance** under VGG-19 and 52.8% under CLIP (Fig. 3b,c), yielding construct-irrelevant shares of 67.0% and 45.1%.

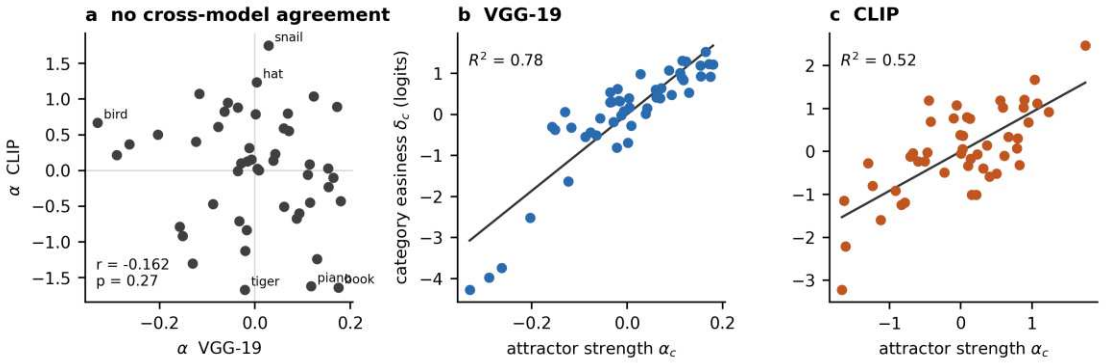


Fig. 3 Cross-scorer agreement and category easiness.

In Fig. 3, panel **a** plots attractor strength under VGG-19 against CLIP across the 48 categories, with six categories labelled for reference. Panels **b** and **c** plot estimated category easiness against attractor strength for VGG-19 and CLIP, with ordinary least squares fits.

Applying the correction reduces category-difficulty dispersion by 26.2% (VGG-19) and 46.9% (CLIP), while leaving child-ability dispersion essentially unchanged (0.532 → 0.521; 0.442 → 0.419). Much of what appears as "some categories are harder to draw" is in fact "some labels absorb more posterior mass."

4.3 Attractor structure does not replicate across scorers (RQ3a)

If α_c indexed a real property, the intrinsic difficulty of drawing a chair, it would agree across scoring architectures. It does not (Fig. 3a):

$$r(\alpha^{\mathrm{VGG}}, \alpha^{\mathrm{CLIP}}) = -0.162, \quad 95\%\ \mathrm{CI}\ [-0.427, 0.128], \quad p = 0.27$$

Spearman ρ = −0.165 (p = 0.26). **25 of 48 categories (52%) reverse their role** between attractor and repeller: indistinguishable from a coin flip. *Book* is a strong attractor under VGG-19 (α = +0.17) and among the strongest repellers under CLIP (α = −1.58); *snail* is near-neutral under VGG-19 (+0.03) and the single strongest attractor under CLIP (+1.76).

Downstream, estimated category easiness correlates only r = 0.265 across scorers, and raw category accuracy only r = 0.445. Crucially, α is highly stable *within* a scorer across age bands (CLIP: r = 0.988–0.998 with the global estimate; VGG-19: 0.63–0.89), confirming it is a property of the model rather than of the children sampled.

4.4 Human recognisers calibrate how much label geometry is normal (RQ3b)

A natural objection is that one scorer may simply be right and the other wrong. The corpus permits a direct test. In the recognition games, child recognisers judged 2,160 of these drawings in a four-alternative forced choice within fixed category sets. We restricted each machine posterior to the same four alternatives and renormalised, so that humans and machines faced an *identical decision problem, over an identical label set, on identical drawings*, and computed α on the resulting 4 × 4 mass matrices. The resulting magnitudes are given in Table 3.

Table 3. Label-geometry magnitude.

Choice set	Human	VGG-19	CLIP	CLIP ÷ human
animals	0.145 [0.105, 0.185]	0.159 [0.119, 0.203]	0.345 [0.286, 0.404]	2.4× [1.7, 3.3]
big animals	0.072 [0.042, 0.101]	0.096 [0.051, 0.142]	0.911 [0.829, 0.998]	12.6× [8.8, 21.9]
objects	0.111 [0.087, 0.135]	0.093 [0.045, 0.145]	0.614 [0.552, 0.678]	5.5× [4.4, 7.1]
vehicles	0.137 [0.097, 0.176]	0.113 [0.063, 0.173]	0.601 [0.521, 0.681]	4.4× [3.3, 6.2]

Note. SD(α) with 95% bootstrap confidence intervals, 2,000 resamples.

Figure 6 displays these intervals; panel b plots the ratio to the human benchmark on a log scale, where parity is the dashed line.

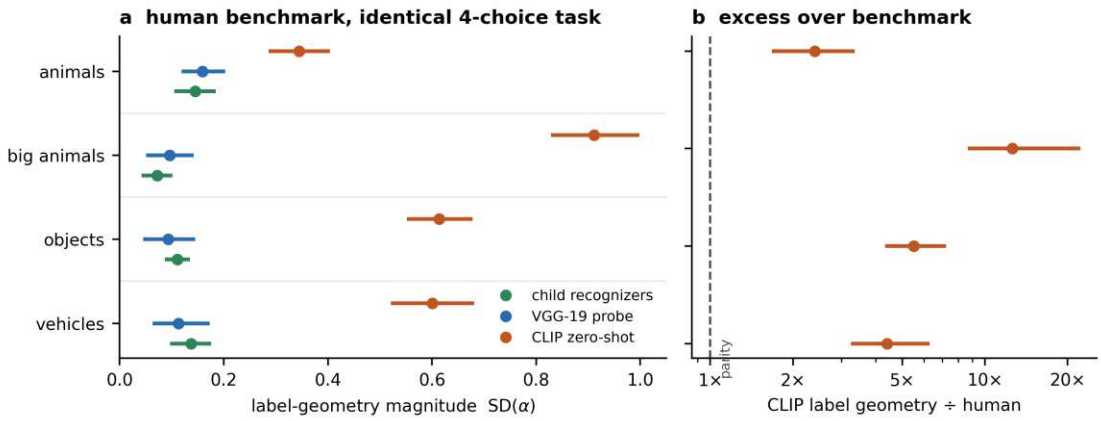


Fig. 6 Label geometry of humans and machine scorers.

In Fig. 6, panel **a** compares child recognisers with both machine scorers on identical four-alternative decisions, by choice set. Panel **b** gives the ratio of CLIP label geometry to the human benchmark on a log scale, with the dashed line marking parity. Bars are 95% bootstrap confidence intervals from 2,000 resamples.

Three conclusions follow, and they refine rather than overturn the preceding analysis.

First, **human recognisers also exhibit label geometry**. Attractor strength is not zero for humans, so $\alpha > 0$ is not by itself evidence of a defective scorer; human confusion has structure, as it should.

Second, **the supervised in-domain probe is human-comparable in magnitude**: VGG-19 intervals overlap the human interval in all four choice sets. Third, **the zero-shot vision–language model is not**: CLIP exceeds the human benchmark by 2.4× to 12.6×, and every ratio interval excludes parity.

This supplies the calibration the construct-irrelevance argument requires. The defensible quantity is not α itself but α **in excess of the human benchmark**, and it explains the asymmetry in §4.5: the scorer whose geometry matches human recognisers barely distorts the developmental trajectory (+0.8%), whereas the scorer exceeding it by roughly an order of magnitude distorts it substantially (+27.1%).

Pattern agreement, as opposed to magnitude, is weak and low-powered. Across the 16 categories, human and VGG-19 attractor patterns do not align ($r = 0.05$, $p = 0.86$; $\rho = 0.17$, $p = 0.54$) and human and CLIP align only marginally ($\rho = 0.50$, $p = 0.047$). With four categories per choice set these correlations rest on three free parameters apiece and should be read as indicative; the magnitude comparison, resting on tens of thousands of trials, is the well-powered test.

4.5 Correcting geometry changes developmental conclusions (RQ4)

For the supervised in-domain probe the developmental slope is essentially unaffected (0.0290 → 0.0292, +0.8%, against SE = 0.0014): the geometry distorts individual scores (2.67% flip; 338 gains, 256 losses) without much biasing the age trend (Fig. 4a).

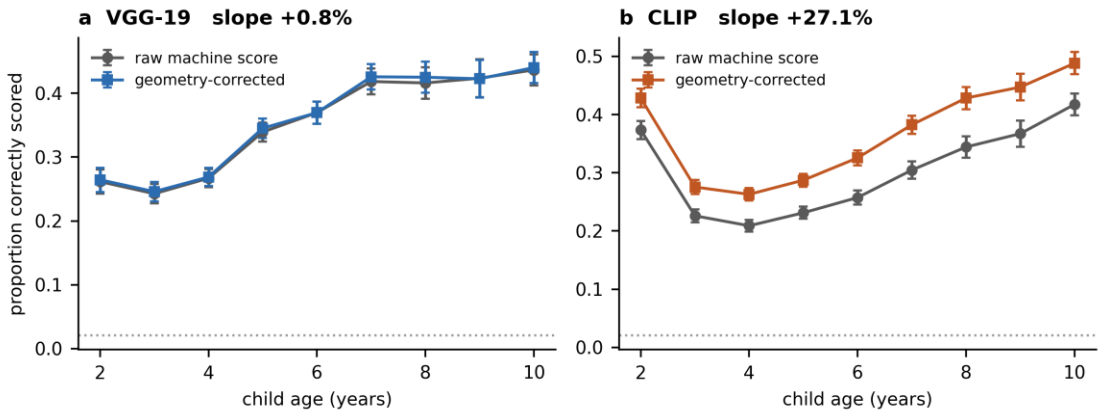


Fig. 4 Developmental trajectories before and after correction.

In Fig. 4, panel **a** shows VGG-19 and panel **b** CLIP on the full corpus. Error bars are 95% binomial confidence intervals and the dotted line marks chance performance (1/48).

For CLIP the effect is substantial. The slope steepens by 14.3% (balanced) and 27.1% (full, 0.0156 → 0.0199 against SE = 0.0010), and 10.7% of scores flip, overwhelmingly in the direction of recovery (3,206 gains vs. 846 losses). Raw CLIP scores therefore **understate measured developmental gain by roughly a quarter** (Fig. 4b).

More seriously, raw CLIP scores place 2-year-olds (0.380) well *above* 4-year-olds (0.218): a developmentally implausible inversion that persists in the balanced subsample and is therefore not a composition artefact. The correction does not eliminate it. Section 4.8 supplies a partial account: the sparse, rapidly executed marks characteristic of 2-year-olds are exactly the production profile most readily absorbed, but absorption explains only part of the effect, and we do not claim to have resolved it. A plausible remaining contributor is survivorship in who engages with the task at that age: the 2-year-olds who complete a prompted drawing at a museum kiosk are unlikely to be representative of 2-year-olds generally. We flag the inversion as a caution about reading raw zero-shot scores developmentally at the youngest ages, not as a finding about 2-year-olds.

Age-differential functioning is **uniform**, not non-uniform: the slope of category easiness on α does not trend with age (VGG-19 $p = 0.35$; CLIP $p = 0.44$). The bias operates as an approximately constant logit offset. This is nonetheless consequential, because a constant logit offset maps into unequal probability changes at different baseline levels, which is exactly why the developmental slope shifts.

4.6 The correction is not circular (robustness)

Attractor strength is estimated from the same posteriors it subsequently corrects. If the gains of §4.5 were in-sample overfitting, they would not survive cross-fitting. We therefore split **children** (not drawings) into five folds, estimated α on the out-of-fold children only, and applied it to the held-out fold, so that no drawing is ever corrected by an α that saw that child. We add a control in which the same α values are randomly shuffled across labels, which preserves the magnitude of the adjustment but destroys its label assignment.

Held-out retention of the accuracy gain was 98.8% (VGG-19), 98.9% (CLIP balanced) and 100.3% (CLIP full); the shuffled-α control reduced accuracy by 0.4, 3.7 and 2.2 percentage points respectively; attractor strength correlated 0.996–1.000 across folds; and the developmental slope change was reproduced out of sample (CLIP full +27.4% held-out vs +27.1% in-sample). Paired tests over children: $t = 3.9$, $p = 9.5 \times 10^{-5}$ (VGG-19); $t = 27.4$, $p = 1.8 \times 10^{-158}$ (CLIP full). Full figures are given in Supplementary Table S1.

The correction transfers essentially without loss: held-out retention is 98.8–100.3%, and the developmental slope change is reproduced out of sample (CLIP full: +27.4% held-out vs. +27.1% in-sample; Fig. 7a). Attractor strength is near-perfectly stable across folds (mean pairwise $r = 0.996$–1.000; Fig. 7b), which is what one expects of a parameter that belongs to the scorer rather than to the sample of children.

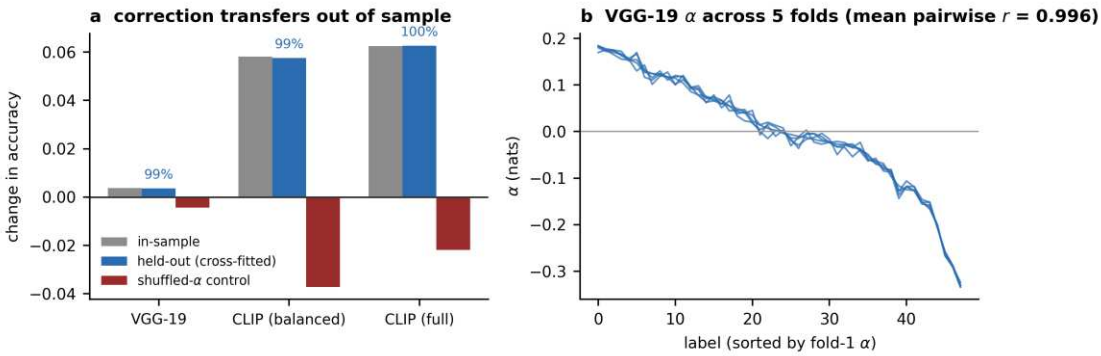


Fig. 7 Cross-fitted validation of the correction.

In Fig. 7, panel **a** compares the in-sample accuracy gain with the cross-fitted held-out gain and a magnitude-matched shuffled-label control; percentages give held-out retention. Panel **b** shows attractor strength estimated independently within each of five child-level folds.

The shuffled-α control is the more informative comparison: applying an adjustment of identical magnitude but arbitrary label assignment **reduces** accuracy in every case (−0.4 to −3.7 percentage points). The gain is therefore not an artefact of logit adjustment in general; it depends on the specific correspondence between labels and their attractor strengths.

4.7 Criterion validity: does correction move scores toward human judgment?

The evidence so far is negative: α does not replicate across scorers. A skeptic may reply that both scorers are simply wrong in different ways and that α is meaningless. We therefore test the positive prediction: if α captures construct-irrelevant variance, removing it should move machine scores *toward* an external criterion. Our criterion is human recognition of the same drawings.

At the **category level**, which is the level at which α operates, the correction improves agreement with human recognisers for both scorers (Table 4):

Table 4. Agreement with human recognition.

	raw r	corrected r	Δr	raw ρ	corrected ρ
VGG-19	+0.441 (p = .088)	**+0.529 (p = .035)**	+0.089	+0.479	+0.528
CLIP	+0.393 (p = .132)	**+0.511 (p = .043)**	+0.118	+0.421	+0.497

Note. Across the 16 categories of the four-alternative choice sets.

Both scorers move in the same direction and both cross conventional significance after correction. With 16 categories this is suggestive rather than conclusive, and we present it as such.

At the **individual-drawing level**, the correction does not help: mean Δr = +0.001 (VGG-19) and −0.031 (CLIP) against per-drawing human recognition rates, cross-fitted by child.

Per-scorer figures are given in Supplementary Table S3. Two things are visible there beyond the null. First, VGG-19 agrees with human recognisers far better than CLIP at the level of individual drawings (r = .30–.48 versus .05–.17), a further respect in which the supervised in-domain scorer is the more human-aligned instrument. Second, the direction of Δr is inconsistent within VGG-19 and uniformly slightly negative for CLIP, which is what a between-label constant should produce when evaluated within categories.

Two things are visible here beyond the null. First, VGG-19 agrees with human recognisers far better than CLIP does at the level of individual drawings (r = .30–.48 versus .05–.17), a further respect in which the supervised in-domain scorer is the more human-aligned instrument. Second, the direction of Δr is inconsistent within VGG-19 and uniformly slightly negative for CLIP, which is what a between-label constant should produce when evaluated within categories. This is not a failure but a scope boundary, and it follows from the construction: α is a between-label constant, so within a category it is very nearly a monotone transform of the raw score and cannot reorder drawings. **Geometry correction repairs comparisons between categories; it does nothing for the assessment of an individual child's drawing.** Practitioners should read the correction accordingly.

4.8 Mechanism: what is absorbed, and what is not contaminated

What gets absorbed. Defining each drawing's *attraction load* as the α-weighted destination of its off-target posterior mass, and regressing on production covariates standardised within category (N = 20,980):

Attraction load falls with stroke count ($\beta = -0.148$, $p < 1 \times 10^{-16}$) and drawing duration ($\beta = -0.113$, $p < 1 \times 10^{-16}$), and rises with ink coverage ($\beta = +0.042$, $p = 5.7 \times 10^{-8}$); shape-tracing skill contributes weakly ($\beta = -0.021$, p = .002). Together these account for R^2 = 0.047 of the variance in load (Supplementary Table S2).

Sparse, quickly executed, ink-dense drawings are drawn toward attractor labels. This coheres with the attractor identities under VGG-19: *bottle*, *book*, *chair*, *apple* and *bowl* are simple closed contours, and supplies the mechanism the diagnostic alone does not. The modest R^2 indicates that production effort shapes absorption without determining it.

A falsified prediction. We predicted that the machine would index production effort more heavily than human recognisers, which would name the contaminating construct as motor output. It does not. Comparing machine and human sensitivity to the same covariates on the same drawings (N = 1,206), using the machine's continuous target logit so that the comparison is like-for-like, gives the coefficients in Table 5:

Table 5. Sensitivity to production effort.

Predictor	machine β	human β	Δβ (machine − human), 95% CI
Shape-tracing skill	+0.147 (p = 1.4 × 10^{-5})	+0.160 (p = 1.8 × 10^{-6})	−0.013 [−0.087, +0.063]
Stroke count	+0.087	+0.122	−0.035 [−0.148, +0.081]
Drawing duration	−0.029	+0.045	−0.075 [−0.156, +0.009]
Ink coverage	+0.038	+0.065	−0.027 [−0.095, +0.043]

Note. Machine versus human; machine outcome is the continuous target logit.

Every difference spans zero. Machine and human scorers weight visuomotor production quality equivalently. (An earlier analysis using a binary machine outcome appeared to show a difference in tracing sensitivity, Δβ = −0.100, 95% CI [−0.173, −0.025]. That is an attenuation artefact of regressing a dichotomous outcome against a continuous one, and it does not survive the like-for-like specification. `mechanism.py` reports both specifications side by side so the artefact is inspectable rather than merely asserted.)

The dissociation. Taken with §4.4, this yields a dissociation between the two facets:

- **Person side: machine ≈ human.** No detectable difference in how machine and human scorers weight production effort.
- **Label side: machine ≠ human.** CLIP's label geometry exceeds the human benchmark by 2.4× to 12.6×, and attractor structure does not replicate across scorers at all.

The contamination is *localised to the label space*. The most obvious alternative explanation, that machine scores are simply differently sensitive to how much effort a child expended, is ruled out by direct test.

One incidental result qualifies the benchmark honestly. Human recognition itself depends on shape-tracing skill (β = +0.142, $p = 6 \times 10^{-7}$), the very visuomotor confound that motivated the original corpus design. Human recognisers are not a clean criterion; they are a *differently contaminated* one. This is the correct reading of §4.4: humans calibrate what magnitude of label geometry is normal, not what a perfect measurement would look like.

5. Discussion

5.1 What the results establish

We frame what follows as a caution for practitioners rather than as a challenge to published developmental findings, and the evidence dictates that choice. The supervised in-domain scorer used in the source corpus performs *well* by every criterion introduced here: its label geometry is statistically indistinguishable in magnitude from that of child recognisers, and correcting it shifts the developmental slope by less than one percent. Conclusions previously drawn with such a scorer are not called into question by this work. What is called into question is the growing practice of substituting zero-shot general-purpose models into the same measurement role.

Accuracy is not a validity argument. Reporting accuracy for a machine scorer is not a validity argument. Two scorers here achieve broadly comparable accuracy on the same corpus (0.335 vs. 0.279 against a chance rate of 0.021) while differing by nearly an order of magnitude in label geometry, disagreeing entirely about which categories are attractors, and implying developmental trajectories that differ by 27%. No accuracy-based diagnostic would surface any of this.

Attractor strength is cheap to compute, it requires only the posterior matrix and the target labels, not model internals, gradients, or retraining, and it is model-agnostic, applying to any scorer that emits a distribution over labels. Given the human benchmark of §4.4, it also admits a defensible reference level rather than an arbitrary threshold.

5.2 What the dissociation buys

The single most useful thing this paper offers a practitioner is not a number but a distinction: **contamination in machine scoring can sit on the person side or the label side, and these require different remedies.**

Person-side contamination is what the fairness literature in educational measurement has taught us to look for. Scores differ across groups of test takers; the diagnosis is differential functioning; the remedy is rebalancing data or adjusting for group membership. Label-side contamination is different in kind. It does not require any group of children to be treated differently, in our balanced designs every age group faces the identical category set, and it is invisible to every person-side diagnostic. It arises because a classifier, unlike a human rater, must commit its judgment to a fixed, finite label inventory whose geometry it did not choose and the assessment designer never inspected.

The person-side null of §4.8 strengthens the construct-irrelevance interpretation, though it does not establish it conclusively. Had machines and humans differed in their sensitivity to production effort, the cross-model divergence in α could have been explained away as two scorers indexing effort differently. They do not differ. The divergence is confined to the label space, which is precisely the facet no rater-based instrument has and every classifier-based instrument does.

This suggests a reframing. When an assessment is automated, the intuition that the machine will reproduce the human rater's biases (more consistently, perhaps, but recognisably) is wrong in a specific and consequential way. On the dimension where we expected the machine to differ, it does not: it weights a child's production effort exactly as human recognisers do. On a dimension with no human counterpart at all, it diverges sharply. **Automating a rater does not inherit that rater's biases; it substitutes new ones.** The practical implication is that validation transferred from a human-rated instrument to its automated successor does not carry over, because the two are vulnerable in different places.

5.3 Scope of the claim

We do not show that machine scoring of children's drawings is invalid. The supervised in-domain probe performs well by this criterion: its label geometry is statistically indistinguishable in magnitude from that of child recognisers facing the same task, and correcting it changes the developmental conclusion by less than one percent. That is a positive result, and it should be read as such.

What we show is narrower and more actionable: a quantifiable share of machine-score variance is scorer-arbitrary; that share is far larger for zero-shot than for supervised in-domain scorers; and it is large enough in the zero-shot case to distort developmental inference. The practical recommendation follows directly: prefer supervised in-domain scorers for developmental measurement, report α against a human benchmark where one is obtainable, and treat cross-model non-replication of item difficulty as a validity red flag rather than a curiosity.

5.4 Relation to existing work

Our finding that CLIP carries strong label bias is consistent with work documenting skewed label priors in web-pretrained vision–language models and the effectiveness of logit adjustment as a remedy (Agarwal et al., 2021; Alabdulmohsin et al., 2024; Shin et al., 2024; Wang et al., 2024). Our contribution is neither the estimator nor the observation that label bias exists (both are established, see §3.6), but to show what that bias *does to a measurement*: to give it a psychometric interpretation, quantify its share of construct-irrelevant variance, benchmark it against human performers of the same task, and trace its consequences for a developmental claim.

Relative to the algorithmic fairness literature in education, which locates bias in person-level attributes (Loukina et al., 2019; Johnson & McCaffrey, 2023; Suk & Han, 2024), label geometry is an *item-side* bias that survives perfect person-group balance, as the age × category balanced designs used here indicate. The two are complementary, not competing: a scoring system can be unbiased across every demographic group and still measure the wrong thing because its label space has sinks.

Relative to representational-geometry work, which has repeatedly found that internal geometry is model-specific in ways accuracy conceals (Mehrer et al., 2020; Storrs et al., 2021; Kornblith et al., 2019), our contribution is to show that this model-specificity can propagate into a substantive developmental conclusion.

5.5 Limitations

Several constraints bound these conclusions. **Corpus.** A single corpus from a single cultural setting and one drawing paradigm; the museum-kiosk sample is not representative, and generalisation to other populations and to clinical instruments such as the DAP is untested. **Method.** The correction is validated out of sample (§4.6) but working from posteriors alone, we cannot separate feature-space geometry from decision-boundary placement; the correction is a post-hoc rebalancing rather than a retrained model, and a retrained scorer might distribute mass differently rather than simply better. **Human benchmark.** The comparison rests on 16 categories in four-alternative sets; magnitude comparisons are well powered but pattern comparisons are not, and child recognisers are themselves developing observers rather than a gold standard. **Unexplained finding.** The 2-year-old inversion under CLIP is described but not fully explained, though §4.8 supplies a partial account: the sparse, rapidly drawn marks typical of 2-year-olds are exactly the profile most readily absorbed. **Criterion power.** The category-level criterion test rests on 16 categories and should be replicated on a wider inventory. **Scope of the estimand.** Attractor strength is defined relative to a corpus; a different category inventory would yield different values, so α characterises a scorer-on-a-task, not a scorer in the abstract. **Two scorers only.** The non-replication result rests on a comparison between two architectures. Two points cannot characterise the distribution of label geometries across scorers, and a wider survey might reveal families of models that agree with one another. The claim supported here is that agreement is not automatic, not that disagreement is general. **Measurement model.** The variance decomposition assumes an additive Rasch structure with no discrimination or guessing parameters and no person-by-item interaction. A model admitting varying discrimination could apportion variance differently, and the reported CIV share should be read as conditional on this specification. **Directional uncertainty in the criterion test.** Correction improved category-level agreement with human recognition for both scorers, but with 16 categories the confidence intervals on those correlations are wide and the improvement, while consistent in direction, is not individually significant for either scorer before correction. **Causal language.** Where the text describes label geometry as distorting or absorbing, this denotes a statistical association established by intervention on the scoring rule rather than a causal claim about drawing behaviour or child development. **The human benchmark is not a gold standard.** Human recognition depends measurably on visuomotor production quality (§4.8), the very confound the source corpus was designed to separate from conceptual knowledge. Human recognisers calibrate what magnitude of label geometry is ordinary; they do not define what a valid measurement would look like.

5.6 Implications

For researchers using machine scores as developmental measures, the minimum standard we would propose is: report α and its null-referenced z; report the variance decomposition $\sigma_\delta^2/\sigma_\theta^2$; and, where a second scorer is available, report cross-model agreement in item difficulty. Where human judgments on a subset of stimuli can be collected, benchmarking α against them converts an uninterpretable number into a calibrated one. None of these requires new data collection beyond what a scoring pipeline already produces.

References

Abell SC, Wood W, Liebman SJ (2001) Children's human figure drawings as measures of intelligence: The comparative validity of three scoring systems. J Psychoeduc Assess 19:204–215. Agarwal S, Krueger G, Clark J et al (2021) Evaluating CLIP: Towards characterization of broader capabilities and downstream implications. arXiv:2108.02818. Alabdulmohsin I, Wang X, Steiner AP et al (2024) CLIP the bias: How useful is balancing data in multimodal learning? International Conference on Learning Representations. Alexandari A, Kundaje A, Shrikumar A (2020) Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. International Conference on Machine Learning, 222–232. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014) Standards for Educational and Psychological Testing. Washington, DC: AERA. Baraheni N, Heidarabady S, Nemati S, Ghojazadeh M (2018) Goodenough-Harris Draw-A-Man Test as a substitute of Ages and Stages Questionnaires for evaluation of cognition. Iran J Child Neurol 12:94–102. Bean AM, Kearns RO, Romanou A (2025) Measuring what matters: Construct validity in large language model benchmarks. Advances in Neural Information Processing Systems. Brualdi RA, Parter SV, Schneider H (1966) The diagonal equivalence of a nonnegative matrix to a stochastic matrix. J Math Anal Appl 16:31–50. Casabianca JM, McCaffrey DF, Johnson MS, Alper N, Zubenko V (2025) Validity arguments for constructed response scoring using generative artificial intelligence applications. arXiv:2501.02334. https://doi.org/10.48550/arXiv.2501.02334 Chen G, Wu AD, Zhang C (2026) Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect predictive bias in automated essay scoring. Assess Writ. Chuang C, Jampani V, Li Y, Torralba A, Jegelka S (2023) Debiasing vision-language models via biased prompts. arXiv:2302.00070. Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 37:35–46. Cox MV (1993) Children's Drawings of the Human Figure. Hove, UK: Lawrence Erlbaum. De Boeck P (2008) Random item IRT models. Psychometrika 73:533–559. https://doi.org/10.1007/s11336-008-9092-x De Boeck P, Wilson M (2004) Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. New York: Springer. Dehdashtian S, Wang L, Boddeti VN (2024) FairerCLIP: Debiasing CLIP's zero-shot predictions using functions in RKHSs. arXiv:2403.15593. Deming WE, Stephan FF (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann Math Stat 11:427–444. https://doi.org/10.1214/aoms/1177731829 Downing SM (2002) Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Adv Health Sci Educ 7:235–241. https://doi.org/10.1023/A:1021112514626 Erbani J, Ben Mokhtar S, Portier P, Egyed-Zsigmond E, Nurbakova D (2025) On the normalization of confusion matrices: Methods and geometric interpretations. arXiv:2509.04959. Freiesleben T (2026) Establishing construct validity in LLM capability benchmarks requires nomological networks. arXiv:2603.15121. Goodenough FL, Harris DB (1950) Studies in the psychology of children's drawings: II. 1928–1949. Psychol Bull 47:369–433. Goyal A, Shah KN, Rao KVV (2026) Zero-shot vision-language models for classroom engagement recognition: A benchmark study of prompt sensitivity and cross-dataset generalisation. arXiv:2606.21861. Hardin PJ, Shumway JM (1997) Statistical significance and normalised confusion matrices. Photogramm Eng

Remote Sens 63:735–739. Harris DB (1963) Children's Drawings as Measures of Intellectual Maturity: A Revision and Extension of the Goodenough Draw-a-Man Test. New York: Harcourt, Brace & World. Imuta K, Scarf D, Pharo H, Hayne H (2013) Drawing a close to the use of human figure drawings as a projective measure of intelligence. PLoS ONE, 8(3), e58991. https://doi.org/10.1371/journal.pone.0058991 Johnson MS, McCaffrey DF (2023) Evaluating fairness of automated scoring in educational measurement. In Advancing Natural Language Processing in Educational Assessment (p 142–164). New York: Routledge. Jolley RP (2010) Children and Pictures: Drawing and Understanding. Chichester: Wiley-Blackwell. Kane MT (2013) Validating the interpretations and uses of test scores. J Educ Meas 50:1–73. https://doi.org/10.1111/jedm.12000 Knight PA (2008) The Sinkhorn–Knopp algorithm: Convergence and applications. SIAM J Matrix Anal Appl 30:261–275. https://doi.org/10.1137/060659624 Koppitz EM (1968) Psychological Evaluation of Children's Human Figure Drawings. New York: Grune & Stratton. Kornblith S, Norouzi M, Lee H, Hinton G (2019) Similarity of neural network representations revisited. International Conference on Machine Learning, 3519–3529. Kriegeskorte N, Kievit RA (2013) Representational geometry: Integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412. Kriegeskorte N, Mur M, Bandettini PA (2008a) Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci 2:4. https://doi.org/10.3389/neuro.06.004.2008 Kriegeskorte N, Mur M, Ruff DA (2008b) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–1141. Kriegeskorte N, Wei X (2021) Neural tuning and representational geometry. Nat Rev Neurosci 22:703–718. Kumar VS, Boulanger D (2021) Automated essay scoring and the deep learning black box: How are rubric scores determined? Int J Artif Intell Educ 31:538–584. Lee S, Cai Y, Meng D, Wang Z, Wu Y (2024) Prompting large language models for zero-shot essay scoring via multi-trait specialization. arXiv:2404.04941. Lin Z (2025) A validity-guided workflow for robust large language model research in psychology. arXiv:2507.04491. https://doi.org/10.31234/osf.io/xw98v Lin Z (2025) From prompts to constructs: A dual-validity framework for LLM research in psychology. arXiv:2506.16697. Litman D, Zhang H, Correnti R, Matsumura LC, Wang E (2021) A fairness evaluation of automated methods for scoring text evidence usage in writing. International Conference on Artificial Intelligence in Education, 255–267. Long B, Fan JE, Huey H, Chai Z, Frank MC (2024) Parallel developmental changes in children's production and recognition of line drawings of visual concepts. Nat Commun 15:1191. https://doi.org/10.1038/s41467-023-44529-9 Long B, Wang Y, Christie S, Frank MC, Fan JE (2024) Developmental changes in drawing production under different memory demands in a U.S. and Chinese sample. Dev Sci. PMID 37768614 Lottridge S, Ormerod C, Jafari A (2023) Psychometric considerations when using deep learning for automated scoring. In Advancing Natural Language Processing in Educational Assessment (p 15–30). New York: Routledge. Loukina A, Madnani N, Zechner K (2019) The many dimensions of algorithmic fairness in educational applications. Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications, 1–10. Mehrer J, Spoerer CJ, Kriegeskorte N, Kietzmann TC (2020) Individual differences among deep neural network models. Nat Commun 11:5725. Menon MV (1967) Reduction of a matrix with positive elements to a doubly stochastic matrix. Proc Am Math Soc 18:244–247. Nagineni S, Hegde RM (2010) On line client-wise cohort set selection for speaker verification using iterative normalization of confusion matrices. 18th European Signal Processing Conference, 576–580. Naglieri JA

(1988) Draw A Person: A Quantitative Scoring System. San Antonio, TX: Psychological Corporation. Naglieri JA, Maxwell S (1981) Inter-rater reliability and concurrent validity of the Goodenough-Harris and McCarthy Draw-A-Child scoring systems. Percept Mot Skills 53:343–348. Nili H, Wingfield C, Walther A et al (2014) A toolbox for representational similarity analysis. PLoS Comput Biol, 10(4), e1003553. Noortgate W, De Boeck P, Meulders M (2003) Cross-classification multilevel logistic models in psychometrics. J Educ Behav Stat 28:369–386. Radford A, Kim JW, Hallacy C (2021) Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748–8763. Rasch G (1980) Probabilistic Models for Some Intelligence and Attainment Tests (expanded ed.). Chicago: University of Chicago Press. (Original work published 1960) Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P (2003) A nonlinear mixed model framework for item response theory. Psychol Methods 8:185–205. Rodriguez PU, Jafari A, Ormerod CM (2019) Language models and automated essay scoring. arXiv:1909.09482. Rüschendorf L (1995) Convergence of the iterative proportional fitting procedure. Ann Stat 23:1160–1174. Schaller NJ, Ding Y, Horbach A, Meyer J, Jansen T (2024) Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, 210–221. Scott LH (1981) Measuring intelligence with the Goodenough-Harris Drawing Test. Psychol Bull 89:483–505. Seth A, Hemani M, Agarwal C (2023) DeAR: Debiasing vision-language models with additive residuals. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6820–6829. Shin C, Zhao J, Cromp S, Vishwakarma H, Sala F (2024) OTTER: Effortless label distribution adaptation of zero-shot models. Advances in Neural Information Processing Systems. Sinkhorn R (1964) A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann Math Stat 35:876–879. https://doi.org/10.1214/aoms/1177703591 Sinkhorn R (1967) Diagonal equivalence to matrices with prescribed row and column sums. American Mathematical Monthly 74:402–405. Sinkhorn R, Knopp P (1967) Concerning nonnegative matrices and doubly stochastic matrices. Pac J Math 21:343–348. https://doi.org/10.2140/pjm.1967.21.343 Soules GW (1991) The rate of convergence of Sinkhorn balancing. Linear Algebra Appl 150:3–40. Stephan FF (1942) An iterative method of adjusting sample frequency tables when expected marginal totals are known. Ann Math Stat 13:166–178. Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N (2021) Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. J Cogn Neurosci 33:2044–2064. Strommen EF (1987) Internal consistency and bias considerations of the Goodenough-Harris Draw-a-Person Test. Educ Psychol Meas 47. Suk Y, Han KT (2024) A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning. J Educ Behav Stat 49:151–172. Ter Laak J, De Goede M, Aleva A, Van Rijswijk P (2005) The Draw-A-Person Test: An indicator of children's cognitive and socioemotional adaptation? J Genet Psychol 166:77–93. Van Meerbeke AV, Sandoval-Garcia C, Ibanez M et al (2011) Validation study on human figure drawing test in a Colombian school children population. Span J Psychol 14:464–477. Wang X (2024) Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. Advances in Neural Information Processing Systems. arXiv:2410.19294 Willcock E, Imuta K, Hayne H (2011) Children's human figure drawings do not measure intellectual ability. J Exp Child Psychol 110:444–452. Xu Y, Vaziri-

Pashkam M (2021) Limits to visual representational correspondence between convolutional neural networks and the human brain. Nat Commun 12:2065.

Data availability

The datasets analysed during the current study are available in the kiddraw repository, https://github.com/brialorelle/kiddraw, released under a CC BY 4.0 licence (Long B, Fan JE, Huey H, Chai Z, Frank MC (2024) Nat Commun 15:1191). No new data were generated. All analysis code required to reproduce every reported result, figure and table from these public data is provided as Supplementary Software, together with a self-contained notebook that runs the complete pipeline from the public corpus without local installation.

Acknowledgements

The author thanks the creators of the kiddraw corpus for releasing the data, the derived classifier outputs and the human recognition judgements under an open licence, without which this secondary analysis would not have been possible.

Author contributions

Rafat Saleh Madani is the sole author of this work and carried out every stage of it. He conceived the study and formulated the research questions; developed the theoretical framing that treats classifier label geometry as a source of construct-irrelevant variance; devised the methodology, including the definition of attractor strength, the accuracy-preserving null and the design of the human benchmark comparison; specified and directed the full analysis pipeline; curated and prepared the data; interpreted all results and drew the substantive conclusions; produced the figures and tables; wrote the manuscript in its entirety and revised it through successive drafts; and managed the project throughout. Using CRediT taxonomy terms: Conceptualisation; Methodology; Software; Formal analysis; Investigation; Data curation; Writing, original draft; Writing, review and editing; Visualisation; Supervision; Project administration. The author read and approved the final manuscript and accepts full responsibility for the integrity of the analysis and the accuracy of all reported results.

Funding

No funding was received for conducting this study.

Competing interests

The author declares no competing interests.

Ethical approval

Ethical approval was not required as this study is a secondary analysis of a publicly available, fully de-identified dataset. No new data were collected from human participants and no identifying information was accessed. The original data collection was conducted under the ethical approvals reported by the corpus authors.

Informed consent

This study did not involve the recruitment of human participants. Informed consent for the original data collection was obtained by the corpus authors as reported in the source publication.

Use of artificial intelligence

The author used Anthropic's Claude to assist with analysis code, manuscript drafting and reference checking. The author reviewed all outputs and takes full responsibility for the content of this article.