



Better than Goodenough? Evaluating new computational techniques for finding diagnostic structure in children's drawings

Clint A. Jensen¹ · Timothy T. Rogers¹ · Karl S. Rosengren²

Accepted: 18 March 2024
© The Psychonomic Society, Inc. 2024

Abstract

In her 1926 book *Measurement of Intelligence by Drawings*, Florence Goodenough pioneered the quantitative analysis of children's human-figure drawings as a tool for evaluating their cognitive development. This influential work launched a broad enterprise in cognitive evaluation that continues to the present day, with most clinicians and researchers deploying variants of the checklist-based scoring methods that Goodenough invented. Yet recent work leveraging computational innovations in cognitive science suggests that human-figure drawings possess much richer structure than checklist-based approaches can capture. The current study uses these contemporary tools to characterize structure in the images from Goodenough's original work, then assesses whether this structure carries information about demographic and cognitive characteristics of the participants in that early study. The results show that contemporary methods can reliably extract information about participant age, gender, and mental faculties from images produced over 100 years ago, with no expert training and with minimal human effort. Moreover, the new analyses suggest a different relationship between drawing and mental ability than that captured by Goodenough's highly influential approach, with important implications for the use of drawings in cognitive evaluation in the present day.

Keywords Human figure drawing (HFD) · Convolutional neural networks (CNN) · Children's drawings · Child development · Intelligence

Introduction

The possibility that children's drawings can serve as a non-verbal assessment of their mental life has a long and broad history including research conducted in a variety of cultures (e.g., USA: Brown, 1897; England: Cooke, 1885; Germany: Kerschensteiner, 1905; France: Luquet, 1912; Italy: Ricci, 1887; as cited in Goodenough, 1926). Encapsulating both the spirit and the viewpoint of this early work, the German researcher Max Verworn (1908) proposed that “the child draws what he *knows* rather than what he *sees*.” Inspired by these authors, but aiming to increase empirical rigor, subsequent researchers adopted a more structured approach to the analysis of drawings (Goodenough, 1926; Harris, 1963;

Kellogg, 1969; Koppitz, 1968; Machover, 1949; Naglieri et al., 1991). Perhaps the most influential system was the work of Florence Goodenough (1926) presented in her book, *Measurement of Intelligence by Drawings*. Reflecting a contemporary interest in measuring mental ability and hewing to Verworn's dictum, Goodenough created the Draw-a-Man Test (DAMT), which applied then-recent advances in statistical analysis to the assessment of structure in children's drawings, and used the results to estimate a child's intelligence quotient (IQ) – launching a legacy of inquiry and reconsideration that continues today.

Goodenough's key innovation was the creation of a checklist-based scoring system for measuring the features appearing in a child's drawing. Any drawing recognizable as an attempted human figure could be scored by counting up the total number of features identifiable in the drawing. Because older children tended to receive higher scores, a given child's score could be used to estimate a “mental age,” which, following Binet's original work, could be compared to their chronological age to derive an estimate of IQ. The checklist approach has become standard in intervening years,

✉ Clint A. Jensen
cjensen5@wisc.edu

¹ Department of Psychology, University of Wisconsin–Madison, Madison, WI, USA

² Department of Brain and Cognitive Science and Department of Psychology, University of Rochester, Rochester, NY, USA

with subsequent innovations mainly involving changes to the number or kind of features scored (71 male – 73 female features, Goodenough-Harris Drawing Test; Harris, 1963; 30 developmental – 30 emotional indicator features, Human Figure Drawing test; Koppitz, 1968; 10 features, Draw-a-Child; McCarthy, 1972; 64 features, Draw-a-Person; Naglieri, 1988) and the cognitive or psychological traits the measure is purported to capture (Burn & Kaufman, 1970; Fury et al., 1997; Hammer, 1958; Harrison et al., 2007; Machover, 1949; Naglieri, 1988; Naglieri et al., 1991). Human figure drawing and accompanying checklists are now a common component in many front-line cognitive assessment batteries (see Rueda et al., 2020, for a review) and also a common tool for research in cognitive development (de Lemos & Doig, 1999; Denver Developmental Screening Test II [DDST- II]; Frankenburg et al., 1992; Karmiloff-Smith, 1990; Zee et al., 2020) and behavioral genetics (Arden et al., 2014; Oliver & Plomin, 2007; Rimfeld et al., 2019; Saudino et al., 1998).

Yet checklist-based metrics are limited in at least three respects. First, the features included rely on the designer's implicit or explicit hypotheses about what *matters* in the drawing a child creates – variation in drawings not anticipated by the scoring system cannot be captured. Second, the way human figures are depicted can vary cross-culturally, over time, or in different demographic groups, creating potential for bias in any given system depending on which features are/are not included. Third, children's drawings may contain richer latent structure than a single checklist score can capture, with important implications for how such drawings may be used in cognitive assessment.

To address these limitations, Jensen et al. (2023) recently developed new computational techniques to measure latent multivariate structure in 305 drawings collected from children aged 3–10 years. These techniques eschewed checklists completely, relying instead on methods from machine vision as well as non-expert human judgments of perceptual similarity and image quality. The resulting metrics encoded remarkably reliable information about the child's age, perceived gender, gross and fine motor abilities, and performance on several subscales of a standardized behavior screening inventory (the Ages and Stages Questionnaire [ASQ]; Squires & Bricker, 2009). Such findings suggest that multivariate structure can be discovered in drawings without reliance on checklists, and that such structure may contain more information about the developing child than previously recognized.

The current paper uses the same techniques to re-analyze the 95 drawings appearing in Goodenough's pioneering book, re-evaluating her observations about the relation between drawings and age, gender, and IQ. We focused on these drawings and characteristics essentially as a convenience sample: the images and child characteristics were all

published in the original book. The drawings are not a random sample of the thousands Goodenough originally collected and analyzed, but were selected by her as representative examples and outlying exceptions to aid in training for the scoring system. Nevertheless, the analysis serves two important purposes. First, it allows us to assess whether the new techniques are robust to changes in culture and drawing conventions – we can ask whether these new methods encode information about age, gender, and cognitive status for drawings collected over 100 years ago, just as they do for contemporary drawings. Second, it provides a case study of how conclusions drawn from the standard checklist approach might change when alternative metrics are used in their place.

The rest of this *Introduction* provides a brief overview of Goodenough's work and the resulting dataset, then describes the new measures developed by Jensen et al. (2023). We then report the results of our re-analyses, and in the *General discussion* consider the implications of these results for the use of drawing in assessment going forward.

Goodenough's legacy

Goodenough believed that drawings could allow researchers to overcome assessment barriers arising from a child's country of origin, culture, language, or verbal ability. As with earlier work (as cited in Goodenough, 1926; Kerschensteiner, 1905; Lamprecht, 1906; Lobsein, 1905; Rouma, 1913), she studied a diverse sample of children from different cultures and SES backgrounds, and so sought a subject for drawing that would be of universal appeal and equal familiarity to all children. She noted a tendency amongst children to draw *people* when not provided with specific instruction, and observed that the human form can be recognized and constructed even from simple figural gestures (e.g., lines and enclosed shapes). Further, given the same task, young children can achieve a recognizable form while older children remain challenged in the range of what can be included and the detail of their drawings (e.g., mouths, lips, eyes, eye lashes, etc.). The task she settled on asked children to draw a *man* – a choice intended to reduce differences associated with gendered appearance norms such as clothing or hair styles.

Beginning in 1920, Goodenough collected drawings and demographic data (e.g., age, gender, advancement in schooling, heritage, and artistic ability) from almost 4,000 children in kindergarten through fourth grade in the USA. Through an iterated age-stratified sample-and-survey approach, Goodenough and colleagues used this dataset to develop and refine a checklist that would allow trained raters to identify key features appearing in human-figure drawings across different developmental periods. The resulting Draw-a-Man test (DAMT) distinguishes drawings that can/cannot be

recognized as human figures, then lists 51 features that may appear in the recognizable drawings (see Online Supplementary Materials (OSM)). To administer the test, children from 4 to 12 years of age are asked to draw a single picture of a man as accurately as possible, and raters then indicate which of the 51 features they can discern in the resulting drawing. Goodenough advocated for this age range given her finding that children older than 12 years and adults often draw a stereotyped figure that is devoid of many features (e.g., a stick figure), resulting in a lower score. After learning how to administer the DAMT checklist through reference to the drawings used in the current research, a trained rater can generate a single valued numeric score for any human figure drawing.

By examining score distributions at each age level across the full sample, Goodenough proposed a normative transform that would yield a given child's "mental age" from their drawing, via the checklist score. The mental age was then divided by chronological age and multiplied by 100 to yield an estimate of IQ. DAMT mental age estimates from drawing showed strong correlations with both academic performance and mental age estimates yielded by the gold-standard IQ test of the time, the 1916 version of the Stanford-Binet test, at each studied age level (see Table 1). The observation was remarkable not least because the measure was derived, not from a complicated battery of tasks developed in the lab and unfamiliar to the child, but from a universally practiced activity that children often undertake for fun.

Today variants of the approach Goodenough pioneered remain in widespread use in clinical, educational, and research settings (e.g., Bruininks-Oseretsky Test of Motor Performance II, [BOT-II]; Bruininks & Bruininks, 2005; de Lemos & Doig, 1999; Denver Developmental Screening Test II [DDST- II]; Frankenburg et al., 1992; Fury et al., 1997; Harrison et al., 2007; McCarthy Scales of Children's Abilities [MSCA]; McCarthy, 1972). One contemporary example,

Table 1 Correlation between "mental age" estimates from the 1916 Stanford-Binet test and the Draw-a-Man test (DAMT) at different ages. Reproduced from Table 9 (Goodenough, 1926, p. 50). Computed for 334 participants

Age	<i>n</i>	Mental age standard deviation			
		Stanford-Binet	DAMT	<i>r</i>	<i>SE</i>
4	25	10.3 mo.	11.5 mo.	.863	.034
5	94	11.0 mo.	12.3 mo.	.699	.035
6	65	17.3 mo.	18.5 mo.	.832	.025
7	63	16.4 mo.	19.2 mo.	.716	.042
8	27	14.6 mo.	18.3 mo.	.557	.092
9	37	20.4 mo.	24.6 mo.	.728	.053
10	23	22.9 mo.	24.8 mo.	.849	.041

r represents the Pearson product-moment correlation

using a more distilled checklist based on the MSCA, is the Twins' Early Development Study (TEDS) – a large-scale longitudinal study of twins born in the United Kingdom between 1994 and 1996 and assessed at 2, 3, 4, 7, 9, 10, 12, 14, 16, 18, and 21 years of age (Oliver & Plomin, 2007; Rimfeld et al., 2019; Saudino et al., 1998). Researchers working with data from this sample found that checklist-derived scores of human figure drawings taken at age 4 years correlated not only with their measured intelligence at 4 years old, but also predicted variation in standard general IQ (*g*) measured in the same participants a decade later at age 14 years ($r = .20$; Arden et al., 2014).

Limitations with the standard checklist approach

Despite the ubiquity of checklist-based assessments, there are reasons to believe that such metrics provide an impoverished and somewhat arbitrary indication of structure in children's drawings. For example, though the DAMT clearly captures some information about overall drawing quality and level of detail (Fig. 1A), it also glosses over potentially important similarities and differences amongst images. Figure 1B shows examples of images that receive the same DAMT score but otherwise incorporate quite different parts and configurations, while Fig. 1C shows examples of images that receive different DAMT scores yet express similar features and configurations. Together these observations suggest that children's drawings may possess a latent structure beyond what can be captured by checklists, which in turn may reflect aspects of a child's cognitive, behavioral, and motor abilities.

Contemporary approaches

To test the possibility that checklists may fail to capture important aspects of the latent structure in children's drawings, we recently developed three novel approaches for characterizing latent structure in children's drawings that provide alternatives to checklist-based metrics (Jensen et al., 2023). The approaches leverage innovations in machine vision and techniques for exploiting human perceptual similarity judgments. When applied to human figure drawings collected in a contemporary sample of children, the metrics together captured aspects of cognitive, motor, and behavioral functioning far beyond that expressed by a contemporary standardized checklist. Given these recent findings and the historical legacy concerning how children's drawings may serve as an indicator of intelligence, the current paper assesses whether the new metrics, when applied to Goodenough's (1926) classic dataset, likewise capture important information about participant demographics and IQ, despite the century separating the collection of these drawings and the new approaches we describe.

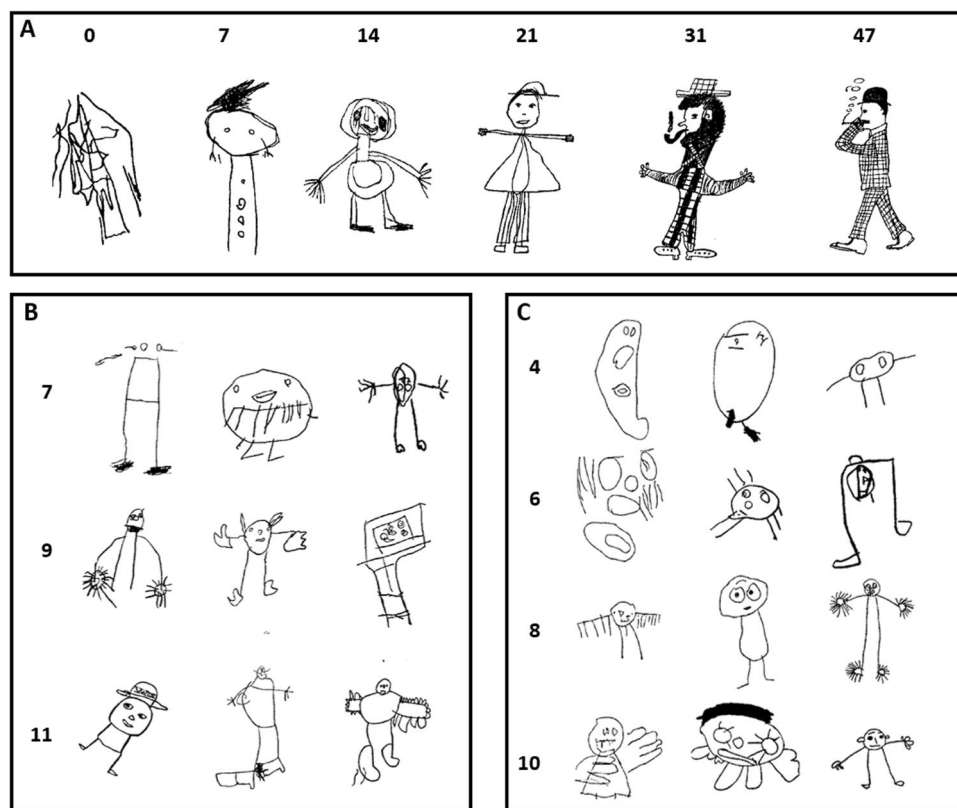


Fig. 1 Example drawings with different Draw-a-Man Test (DAMT) scores. Reproduced from Goodenough (1926). (A) These images show representative examples of drawings spanning the range of DAMT scores in the original dataset. The score for each drawing is indicated above it. (B) These images highlight variation among drawings receiving the same DAMT score, indicated at the left of each row (7, 9, and 11). Though items in the same row all received the same score, they contain quite different parts and configurations. (C)

These images illustrate visually apparent similarities across sets of drawings that received different DAMT scores. For instance, images in the first row (DAMT score = 4) are represented by circular shapes that suggest either the face alone or form that encompasses both the face and body/trunk, with appendages extending from that central form. A similar pattern can be seen across images that receive higher scores in second, third, and fourth rows (DAMT scores = 6, 8, and 10, respectively)

Such an exercise is illuminating for several reasons. First, Goodenough's scale was derived from extensive, careful research, yielding a detailed scoring system based on her observations from thousands of drawings. The resulting DAMT requires expert training and nontrivial human labor to score the drawings. More current checklist-based drawing assessments have greatly reduced the richness of Goodenough's original approach for pragmatic reasons by focusing on the presence of a more limited set of features in drawings. In contrast, the metrics we have developed require no expert training and minimal human effort. If they capture as much or more information about the developing child, they open the possibility for highly efficient and richer evaluation of cognitive change over childhood via drawing.

Second, drawing conventions change with culture and over time, in ways that may affect checklist scores (Cox, 1998; Cox et al., 2001; De La Serna et al., 1979; La Voy et al., 2001). Our prior work analyzed drawings produced recently by children aged 3–10 years, finding that latent

structure could predict socio-demographic characteristics of the artist including age and gender as well as scores on standardized measures like sub-components of the Ages-and-Stages Questionnaire (Jensen et al., 2023). If the same metrics likewise predict participant characteristics in a sample of drawings collected over 100 years ago, this would suggest that (a) they are robust to changes in culture and drawing convention, and (b) assessment of drawings need not rely on the particular features encompassed in Goodenough's (or any other) checklist. Conversely, if the DAMT better predicts socio-demographic or other characteristics of the children from the original study, this would suggest that the checklist captures important structure not expressed in the newer measures.

Third, Goodenough's work was impactful because it offered a simple and familiar non-verbal means of estimating a child's IQ, laying the foundation for the use of drawings in cognitive assessments more broadly. Many cognitive evaluation tools rely on tasks that are unfamiliar, not particularly

motivating, and/or reliant on language, making it difficult to disentangle factors underlying individual variability within and across ages. By contrast, drawing is an activity that most children enthusiastically engage in outside of the lab that, like language production, requires coordination/integration of perception, attention, semantic knowledge, executive function, and motor control. New metrics that encompass or expand on the structure captured by the DAMT may alter our understanding of the relationship between verbal and nonverbal aspects of cognition.

Overview of contemporary tools

Approach 1: Machine-derived latent feature vectors

The first approach characterizes the structure of drawings using features acquired by deep neural network image classifiers. When trained on color photographs of objects, such models demonstrate both human-level categorization performance (Krizhevsky et al., 2012; Yu et al., 2022) and a degree of generalizability to visual media outside their training sets, including drawings (Fan et al., 2018). The current study used VGG-19, a deep convolutional image classifier neural-network pretrained to assign each of ~14M ImageNet photographs to one of 1,000 possible mutually exclusive categories, based on its utility within both visual cognition and neuroscience (Fan et al., 2018; Jensen et al., 2023; Jha et al., 2023; Long et al., 2021; Simonyan & Zisserman, 2014); however, the principles of our approach could be extended to other model varieties.

In our procedure, each drawing in a dataset is converted to a bitmap that provides the input to the pretrained network. The resulting activation patterns across the 4,096-unit penultimate model layer (i.e., the last layer before classification) are extracted and serve as vector-based representations of each drawing. We then compute the cosine similarity of each vector pair and from the resulting matrix apply classical multidimensional scaling to extract d of latent coordinates for each image so that distances within the d -dimensional space approximate those within the high-dimension model representation space. In this way, each image is re-represented as a *machine-derived latent feature vector* of dimension d . The full workflow is shown in Fig. 2A.

Approach 2: Using crowd-sourced human perceptual judgments to identify structure

The second approach exploits human perception to find latent structure in drawings through use of crowd-sourced triadic comparisons. In each trial of this task, participants view a *sample* image above two *options* images (Fig. 2B) and are instructed to “Choose the image below that is most similar to the image above.” The criterion for “most similar”

is left unspecified so that participants are free to use whatever basis of similarity they wish. From many such judgments a matrix can be constructed indicating how often each pair of images was selected as “most similar” across all triplets in which they appeared. An ordinal embedding algorithm can then be applied to situate each image within a low-dimensional space such that Euclidean distances between two items express the likelihood that they are selected as “more similar” relative to some third item (Jamieson et al., 2015). Following this approach, the embeddings constitute a low-dimensional representation that approximates the continuous perceptual similarity relationships identified by the human participant raters. The result is a *human-derived latent feature vector* for each image, with the vector space capturing perceptual similarities amongst drawings. Jensen et al. (2023) found that human- and machine-derived vector spaces may express quite different information about the same set of drawings. Machine-based representations make use of complex visual features learned from classifying millions of photographs, which can express high-order visual structure that can be difficult for naïve human judges to discern or explicitly rate. Conversely, human perception is enriched by additional semantic knowledge about depicted items: human raters can not only classify images as depicting people, but understand the parts that comprise images as well as their respective names and functions (e.g., that a series of lines represent a human head that is wearing a hat, and that two different figures each share this feature). Thus, human- and machine-derived vector spaces may capture distinct information about the similarity relations amongst drawings, which in turn may be differentially useful for different assessment tasks.

Approach 3: Crowd-sourced judgments of drawing quality

The final approach again makes use of crowd-sourced human judgments, but rather than estimating a perceptual similarity space, we instead use comparative evaluations to *rank* images by their perceived overall *quality*. On each task trial participants view a randomly sampled pair of images and must indicate “Which is the better drawing of a person” (Fig. 2C). From many such judgments we compute, for each image, the proportion of times it was selected as a *better drawing* amongst all the trials in which it appeared. This metric situates each drawing along a continuum from the worst to the best depictions based on their perceived quality, a value we refer to as a drawing’s *Quality-rank* (QR) score. Like checklist-based approaches, QR produces a single score for each image, but it does not require expert training, does not rely on specific hypotheses about which features of a drawing *matter* when characterizing a child’s cognitive status, and does not require an effortful and time-consuming evaluation process. An interesting empirical question, then,

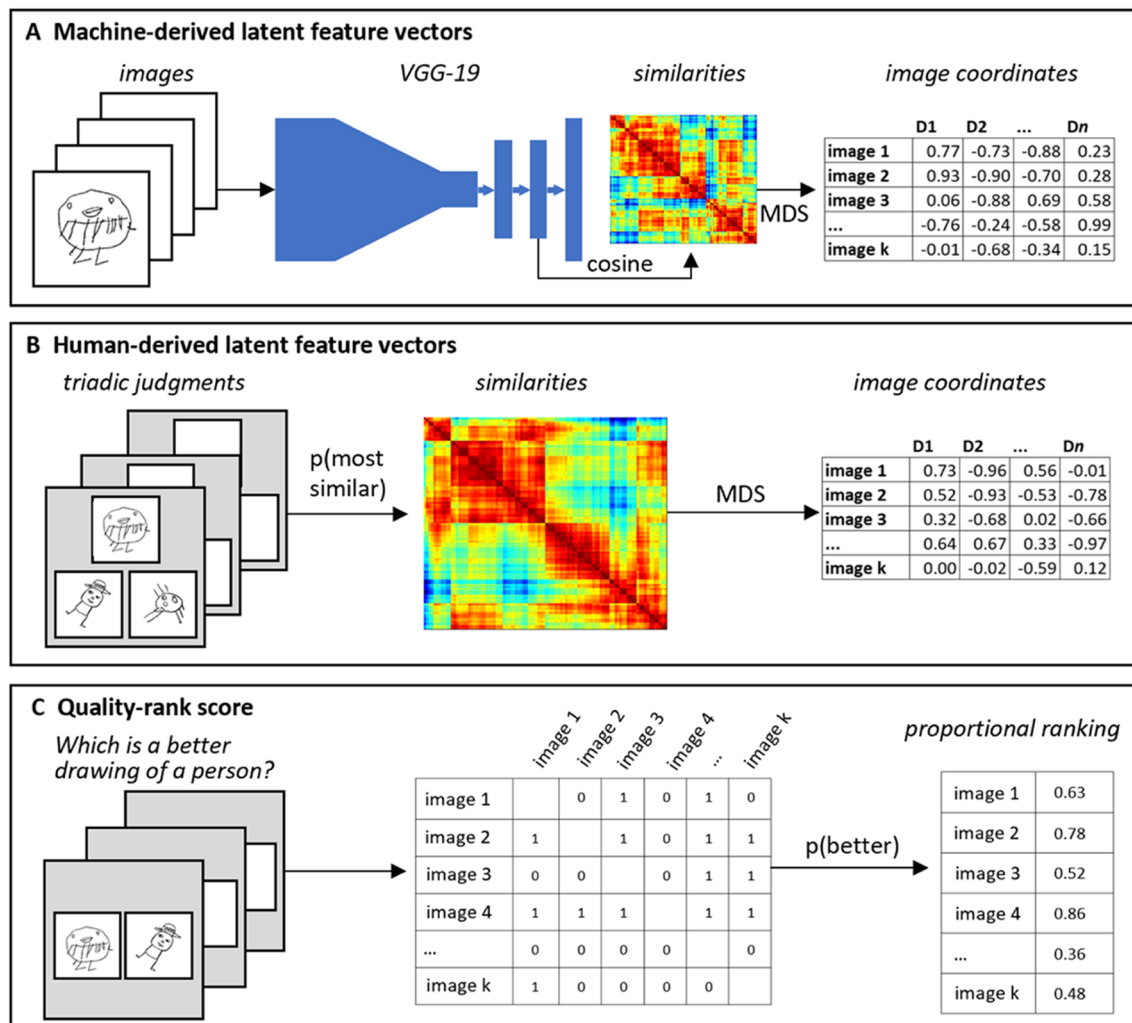


Fig. 2 Three methods for capturing structure in drawings. **(A)** The VGG-19 workflow uses each image as input for the neural network, extracting high-dimensional vectors from the penultimate layer, computing pairwise cosine similarities for all sketches, and reducing the resulting output to a set of coordinates for each image using machine-derived vector space (MDS). **(B)** The human workflow for triadic judgments collects participant decisions concerning how similar the images are to one another for a large set of triplets, with the prompt, “Choose the image below that is most similar to the image above.” Subsequently, the pairwise probability that two images are chosen

as most similar is computed, and again MDS is used to reduce the similarity matrix to a small number of coordinates for each drawing. **(C)** The human workflow for Quality-ranking judgments presents random pairings of images and has participants decide which of the two drawings is the “better drawing of a person” yielding a matrix where each row indicates which of the sketches was deemed to be a higher quality drawing. A proportion of times the individual images were selected as “better” is then computed, resulting in a one-dimensional ranking embedding

is whether QR, independent of *or together with* other metrics, can provide as accurate an estimate of a child’s demographic and cognitive status as does Goodenough’s original checklist or its more contemporary variant.

Together the application of these methods to a set of drawings yields, for each image, (a) a machine-derived latent feature vector, (b) a human-derived latent feature vector, and (c) a human-derived estimate of perceived drawing quality. These numeric descriptions of the images do not correspond to explicit, identifiable features of the kind appearing in checklists, but may nevertheless capture underlying structure

that reflects characteristics of Goodenough’s participants. The next section empirically tests this possibility.

Method

Development of the Draw-a-Man-Test (DAMT) in Goodenough’s original work

Though the original published work includes only the 95 drawings used in the present study, the DAMT scale was

developed and standardized through consideration of drawings from 3,593 children (4–10 years old) primarily living in New Jersey, USA. That large collection of children's drawings were used to develop and standardize the final version of the DAMT scale. The schools where the task was administered were selected to reflect a diverse sample of children who varied in age, educational achievement, race, ethnicity, and familial social status. An additional 2,034 children's drawings were used to validate the scale (652 – Rutherford, New Jersey; 479 – Fresno, California; and a combined set of 540 from exclusively white children in Mt. Pleasant, Tennessee; Chattanooga, Tennessee; Natchitoches, Louisiana; and 363 – Santa Clara, California). No differences based on locality were qualitatively observed, though the number of students who were considered advanced or behind their corresponding age to grade level did differ by sample.

The resulting DAMT consisted of a scale of two classes: Class A – the *preliminary stage* of drawings that cannot be identified as a human figure (e.g., scribbles, and lines that approach shapes), and Class B – for drawings that can be recognized as human figures. Class B comprises 51 features, five being reserved for figures in profile, and was designed to capture variance outside of tests of verbal ability, shape matching etc. Administration of the DAMT was straightforward: children between 4 and 12 years of age were asked (in a group setting) to draw a single picture of a man based on the following instructions:

On these papers I want you to make a picture of a man. Make the very best picture that you can. Take your time and work very carefully. I want to see whether the boys and girls in _____ school can do as well as those in other schools. Try very hard and see what good pictures you can make.

After explaining the task, the examiner was urged to walk around the room, and to encourage the children as a group, being careful not to praise any specific child's drawing, to identify any omissions or offer any suggestions about what content should be included.

Stimuli from Goodenough (1926) used in the current experiment

From her full collection, Goodenough included one drawing from 95 children for publication within the Series 1 and 2 training/test exercises in her (1926) book *Measurement of Intelligence by Drawings*. The drawings were selected to reflect characteristic elements of children's output as well as several figures identified as “bizarre” (p. 89). Each drawing included a figure note identifying a by-item account of the individual checklist credits Goodenough assigned based on the features deemed present within the DAMT scale. The DAMT Score, which constitutes a summed count of those

credits, was also included for all 95 drawings. The majority of drawings also included demographic information (e.g., gender, age, grade level), a DAMT-based estimate of the child's Mental Age, corresponding IQ score, and notes related to the individual child or drawing.

For the present study, original drawings from the Goodenough (1926) text were photographed using a digital camera. All images were cropped to remove notes or other markings not part of the drawing (e.g., the word “stomach” written within in the figure, figure borders, etc.) while maintaining the aspect ratio. The images were then contrast-normalized so that all pixels were either black or white to ensure minimal low-level visual differences between the captured images. All drawings were also centered and padded with white pixels to a uniform height of 450 pixels and width of 450 pixels.

Three of the 95 drawings depict two or more human figures side-by-side. To avoid human rater similarity judgments based on the number of figures within a drawing rather than the manner in which they were drawn, these images were edited so that each human figure was presented as a separate drawing while maintaining their scale and position on the page.

One drawing within the published collection includes a smaller impoverished and inverted figure beside a completed human figure. Goodenough made note that this specific additional figure was not scored in her original study and so that inverted figure was edited out of the image used within our study. This resulted in a total of 98 images from 95 children, ages (stated in year;month) ranging from 4;0 – 14;7 ($M = 8$, $SD = 2;9$, 47% female, 53% male).

Measures from Goodenough (1926) used in the current study

DAMT score

The DAMT scores were transcribed from the figure notes that accompany all drawings within the published Series 1 and 2 training/test exercises.

Demographic characteristics

We also recorded chronological age and reported gender of the child who produced each drawing. The number of children reported at each age level was roughly uniformly distributed.

Intelligence quotient (IQ) estimate

IQ score estimates included in the published notes were derived from the DAMT by first estimating a child's “mental age” based on normative scores; for instance, the mean score

for 8-year-olds was 22, so a drawing receiving a score of 22 suggested an approximate “mental age” of 8. To derive an age-standardized IQ, the “mental age” estimate was divided by the child’s chronological age and the result multiplied by 100. Thus, a child of 7 years scoring 22 on the DAMT had an estimated “mental age” of 8 (since 8-year-olds scored 22 on average), yielding an IQ of $100 * 8/7$ or 114. This approach to developmental IQ – computing norms for different age ranges, using these norms to estimate mental age from the child’s test score, then taking the ratio of mental age to chronological age – paralleled the approach originally developed by Binet, which was used in the contemporaneous Stanford-Binet test (Table 1).

Procedures for additional metrics collected in the current study

Draw-a-Child 12-item score

Each image in the dataset was scored using the contemporary Draw-a-Child checklist employed by the TEDS longitudinal study (Arden et al., 2014; Oliver & Plomin, 2007; Saudino et al., 1998), a tool that is conceptually similar to the DAMT but includes many fewer items (12 rather than 51). All drawings were scored by a trained rater indicating which of the 12 features were present in the drawings, resulting in a total summed score from 0–12 for each image.

Machine-derived latent feature vectors

We used a standard implementation of the VGG-19 architecture pre-trained to classify photographs of real objects in the ImageNet database (Deng et al., 2009; Simonyan & Zisserman, 2014). We chose VGG-19 as the image encoder as this was the model used in prior work (Jensen et al., 2023). The model was coded in Python 3.6 using TensorFlow (1.13.1) libraries. The input for the model consisted of the pre-processed drawings, which were rescaled to the dimensions of the model input layer ($3 \times 224 \times 224$). Activation patterns for each layer of the model were computed in a feed-forward pass, and the resulting vector output from the model’s penultimate layer were extracted for each drawing. Cosine similarities were computed for all vector pairs, and the resulting matrix was decomposed using classical multidimensional scaling to yield embeddings in two dimensions, so that each image was represented as two coordinates. All code for replicating our analyses is available at <https://github.com/ClintJensen/DrawingsProject>.

Human-derived metrics

Data collection from human participants was approved by the Institutional Review Board for Education and Social/Behavioral Science at the University of Wisconsin-Madison

(Protocol no. 2013–09999, “Learning meanings of words and objects”). Sixty-seven undergraduate participants were recruited for the triadic comparison task and 73 for the QR task, both via the SONA system. They took part in the study using a computer in their own space, via the “Salmon” online interactive cloud-based platform that facilitates collection of comparative similarity judgments for visual and verbal stimuli (Sievert, 2019). Prior to the study, participants provided informed consent and completed a reCAPTCHA verification. Participation was compensated via course credit. For both tasks participants were asked to judge 200 stimuli but were permitted to stop at any point.

Triadic comparison task

On each trial participants pressed a left or right arrow key to indicate which of two drawings at the bottom of their screen was most similar to a target image at the top. Participants who had a mean response time of 1 s or less, or who chose the same response key significantly more frequently than expected given the random positioning of items on the screen, were omitted from the analysis. Human-derived latent features were then computed from the remaining sample of 46 participants who together contributed 9,037 judgments.

From these data, a random 10% of trials were selected as hold-outs to evaluate the embeddings. Embeddings were then estimated in 1–5 dimensions from the remaining data using Crowd Kernel, an algorithm that produces learned non-metric embeddings from discrete group-level comparative decisions (Tamuz et al., 2011). Each embedding was evaluated against the held-out triplets by considering, for each test trial, which of the two options was closest to the target in the embedding space, and comparing this modeled *choice* to the human decision on the same item. We then compared mean prediction accuracy across the embeddings. A two-dimensional space showed the best performance and was used in both the visualizations and the regression analyses that follow. Data were collected between the 6th of May, 2022 and the 5th of December, 2022.

Quality-rank score

On each trial participants viewed two drawings and judged via the left or right arrow key which was the better drawing of a person. We again removed participants with very rapid mean response times (under 1 s) or who chose the same response key reliably more often than expected given the randomization of images on the screen. QR scores were then computed from the remaining 51 participants, who together produced a total of 10,075 judgments. The QR score for each image was calculated as the number of times a drawing was selected as the better depiction of a person divided by

the total number of times it appeared, producing a proportion in the range 0–1. Data were collected between the 1st of December, 2022 and the 3rd of December, 2022.

Results

Comparing human- and machine-derived feature spaces

Figure 3 shows the two-dimensional (2D) machine- and human-derived latent feature vectors for all drawings, with color indicating participant age. By inspection both spaces capture age information, but also express somewhat different similarities amongst the images, as observed previously for contemporary drawings (Jensen et al., 2023). For instance, the machine space situates drawings with open round shapes toward the extreme left and tall thin drawings toward the right, whereas the human space organizes drawings partly based on their semantic content, with unrecognizable and simple images in the top left, more complex figures drawn in profile in the bottom right, and those shown face-on in the top right.

To quantify whether the human- and machine-derived embeddings capture different similarity relationships among the drawings, we used regression to predict coordinates in one embedding space from those in the other. Separately for each dimension in the target space, we fit a linear model to predict an image's location from its two coordinates in the original space and their interaction. We then used the two fitted models to predict both coordinates for all images in the target space and evaluated the accuracy of the predictions using the square of the Procrustes correlation between

predicted and true coordinates. This metric indicates how much variation of distances in the target space is explained by regression from coordinates in the origin space, analogous to r^2 . Predicting human coordinates from machine embeddings accounted for 38% of the total variation distance, reliably better than chance ($p < 0.001$) but with over 60% of the variance unexplained. Predicting machine coordinates from human embeddings accounted for 31% of the variance, also reliably better than chance ($p < 0.001$) but with over 65% of the variation distance unexplained. Thus, while the two embedding spaces are not completely unrelated, each express substantially different similarity relations amongst the drawings.

Comparing measures of drawing quality

We next considered relations amongst different measures of drawing quality. Figure 4 plots the relationship between drawing scores from the original 51-item DAMT and two contemporary measures of drawing quality: the total score from the 12-item Draw-a-Child checklist, and the QR score derived from non-expert crowd-sourced human judgments. Both cases showed a clear nonlinear relationship with the original measure. The red curves in the figure show the best-fitting exponential curve predicting original scores from each new metric. Predictions from the 12-item checklist accounted for 72% of the variance in the original DAMT scores, but with prediction error clearly increasing for drawings with higher scores – the variance around the regression line is much wider toward the right end of the range. In contrast, the QR score, despite arising from non-expert judgments without any special attention to which features do/do not appear in the drawing, accounted for a remarkable 83%

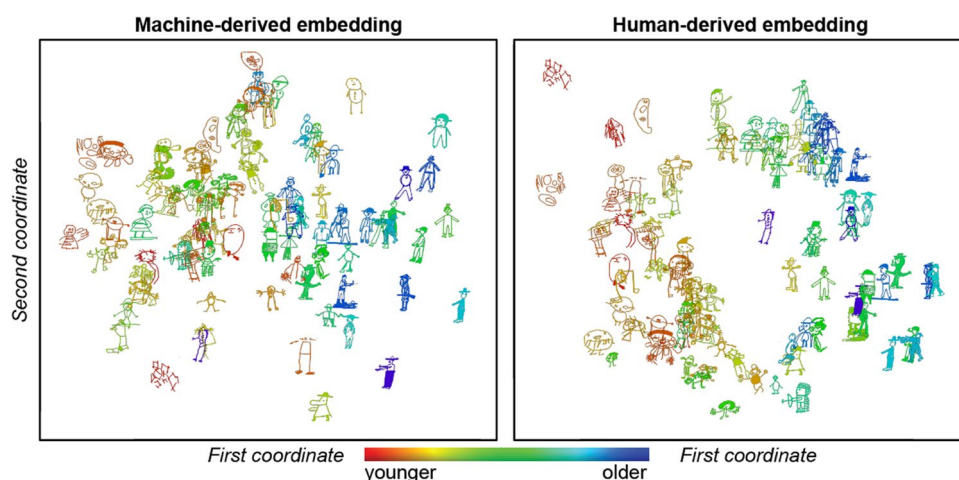


Fig. 3 Two-dimensional embeddings derived from VGG-19 (left) or from human triplet judgments (right). Each sketch from Goodenough (1926) is placed according to its coordinates in the corresponding two-dimensional space. Colors show the age of the participant

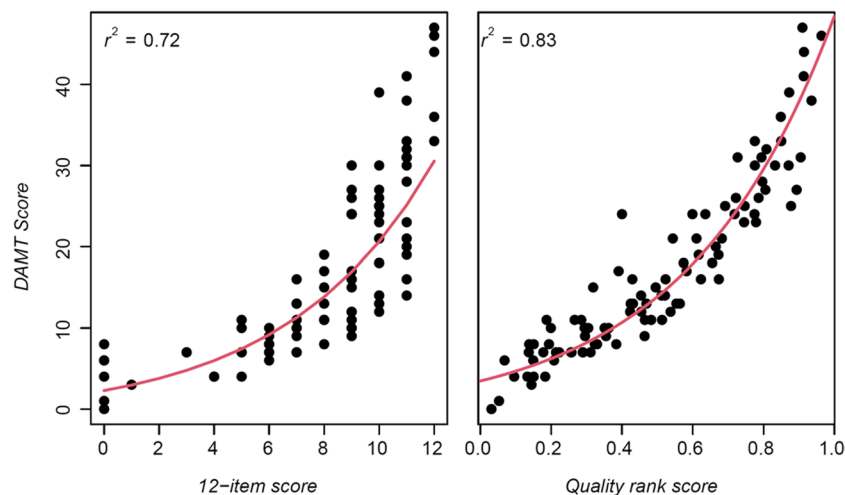


Fig. 4 Regressions predicting Draw-a-Man-Test (DAMT) score from two alternative measures of drawing quality. The 12-item score, a short human figure drawing checklist metric in current widespread use, shows a reliably exponential relationship with the original DAMT score, but with increasing dispersion around the regression

of the variation in the original DAMT scores, with roughly equal scatter around the regression line across the full range. This result suggests that QR scores better capture information from Goodenough's original more detailed DAMT checklist, despite reflecting nothing more than crowd-sourced, subjective judgments of drawing quality.

Gender differences

Prior to Goodenough, large-scale studies of children's drawings had generally reported better performance for boys than girls (Kerschensteiner, 1905; Lamprecht, 1906; Lobsein, 1905; Rouma, 1913; cf. Goodenough, 1926). Her results contradicted this pattern, showing a marked advantage among participants identified as female – a pattern that has since been replicated in several checklist-based studies of human figure drawing (Cox, 1993; Cox et al., 2001; Goodenough & Harris, 1950; Harris, 1963; Lange-Küttner, 2011; Lange-Küttner et al., 2002; Naglieri, 1988; Picard, 2015). The causes of this effect are unclear, but many checklists include items that are culturally gendered, and the resulting scores may in turn be influenced by drawing conventions that reflect such socially derived influences. For instance, the DAMT includes points for depicting hair and clothing, items that persist in the 12-item Draw-a-Child test. Common drawing conventions often signal gender differences via presence/absence of hair or through selection of triangular versus straight bodies that suggest presence/absence of skirts or dresses. Goodenough asked children to draw a *man* specifically to avoid such confounds, but it remains

unclear to what extent checklists deployed in her work and subsequent studies include items where children identifying/presenting as male or female may have differing degrees of knowledge or familiarity with the cultural conventions that underpin the unique selection of features represented across checklist-based assessments.

One possible hypothesis is that previously documented gender differences in drawing scores mainly arise as an artifact of the particular features the assessment instrument designer has included in the checklist – features that are culturally or socially gendered may be differentially likely to be included in drawings produced by children who identify as boys versus girls. Alternatively, it may be that such differences arise not as an artifact of the features included in a given checklist, but because children with differing gender identities tend to depict the adult male figure in systematically different ways. To adjudicate between these positions, we assessed whether contemporary metrics computed on the 98 drawings in the dataset carry latent information about participant gender, beyond any information encoded by the total DAMT score. If gender differences arise solely as an artifact of the features included in the checklist, it should not be possible to decode participant gender from the new metrics.

With this idea in mind, we fit and evaluated a series of logistic regression models predicting perceived participant gender from metrics derived from the drawings and including age as a covariate. A model fit using age alone did not predict gender reliably better than chance, nor did models fit using age and either of the three drawing-quality measures

(DAMT, 12-item score, QR score), as simple effects or with an interaction. Thus, there was no suggestion of a confound between age and gender, nor of overall better drawing quality for either reported gender within this sample of drawings. We then applied stepwise binomial regression to evaluate whether human- or machine-derived latent feature vectors capture information about participant gender, alone or in combination with age and QR score.

For human-derived vectors, the stepwise procedure retained participant age and both image coordinates, as well as some second-order interactions, but discarded QR score. The fit of this model was significantly better than a model fit using only age ($p < 0.04$ by Chi-square). We evaluated the model accuracy by computing confidence limits on the area under the receiver-operator curve (area-under-curve or AUC), a threshold-independent metric of binary classifier accuracy that adopts a value of 0.5 for null models and 1.0 for perfect discrimination. The binomial classifier fit to human-derived embeddings had a mean AUC of 0.66 with 95% confidence interval (CI) of 0.57–0.78, suggesting somewhat better than chance classification.

For machine-derived vectors, the selected model retained both coordinates and the QR score as well as some second-order interactions, but did not retain age. It fit reliably better

than a model fit only with QR score ($p < 0.0001$ by Chi-square), and also showed higher accuracy than the model fit with human-derived embeddings (mean AUC = 0.73, 95% CI: 0.63–0.83). Thus, the machine-derived embeddings, together with the QR score, reliably capture information about participant gender.

To visualize this information, we used the best-performing regression model (fit using QR score and machine-derived latent features) to estimate the probability that each drawing was produced by a girl, then plotted these probabilities against participant age. The result is shown in Fig. 5, which also highlights eight drawings Goodenough chose as representative of gender differences she noticed in her study. The regression correctly classified all but one of Goodenough's examples. It is clear that the machine vectors exploit structure that is difficult for the human eye to discern, and thus difficult to capture explicitly in a checklist. Since neither the machine-derived embeddings nor the QR score contain any information about the features appearing in the checklist metrics, this finding suggests that participants reported as male versus female in this sample differ systematically in how they depict the adult male form, in ways that go beyond the particular features included in a given checklist.

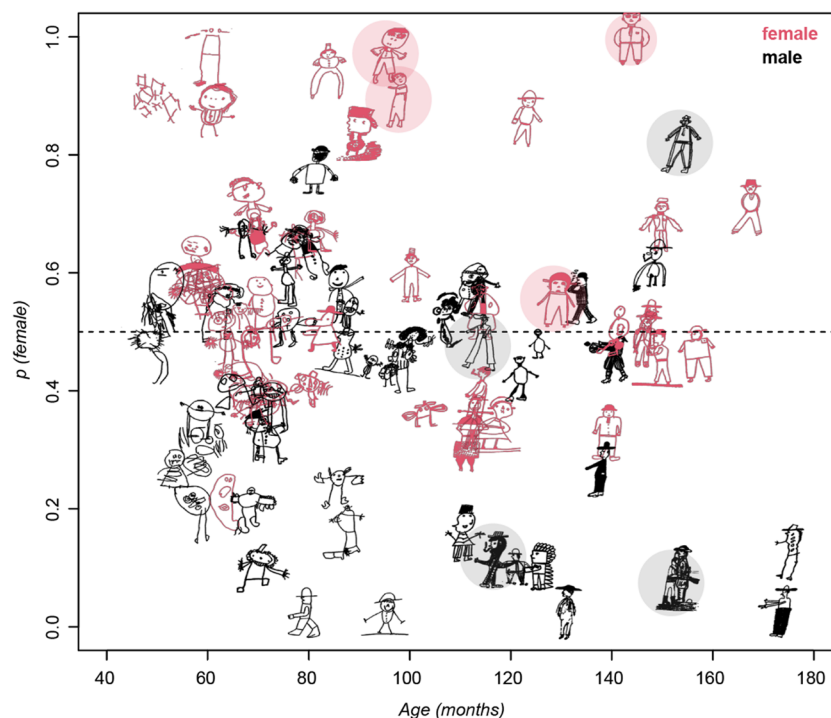


Fig. 5 Predicted probability that a participant is female based on drawing characteristics, plotted against age. Probabilities on the y-axis were generated by applying the best-fitting logistic regression model, which used Quality-rank score and machine-derived latent

features as predictors, to image data. The colored circles indicate images Goodenough highlighted to exemplify her ideas about gender differences in human figure drawings (red for female, gray for male)

Age differences

While it is no surprise that older children generally produce drawings that receive higher DAMT scores, this predictive relationship between DAMT score and age played a critical role in Goodenough's effort to create a proxy for IQ suited to children at different ages. Her DAMT IQ score is simply the ratio of a child's "mental" to chronological age, with the mental age estimate derived from DAMT scores normed for each age group. This conception of IQ clearly depends upon the aggregate accuracy of the model that predicts a child's age from information in her drawings – if the relationship between drawing metric and chronological age changes, so too will the estimate of IQ derived from the drawing (e.g., the Flynn effect, but see Lynn, 2013, for reconsideration of the effect's attribution). We therefore compared how well age can be predicted from classic and novel drawing metrics.

We began with the three measures of overall drawing quality: the DAMT score, the 12-item checklist score, and the QR score, with results shown in Fig. 6. All three measures showed an exponential relationship with participant age, with the best-fit curve accounting for 60% of the variance for DAMT, 45% for the 12-item checklist, and 58% for the QR score ($p < 0.001$ for contrast to null hypothesis in all contrasts). Thus the 12-item checklist score shows a looser relationship to participant age than either alternative method, which show a similar relationship to participant age.

Do DAMT and QR scores capture the same predictive relationship? To answer this we assessed whether the fit of each model was reliably improved by adding the other metric and its interaction to the regression. Adding QR score to the DAMT model reliably improved model fit ($r^2 = 0.63$ vs. 0.60 , $F(94) = 3.7$, $p < 0.03$), as did adding DAMT to the QR model ($r^2 = 0.63$ vs. 0.58 , $F(94) = 6.0$, $p < 0.004$). Thus, each metric captures unique variance in age.

Next, we assessed whether human- and machine-derived latent feature vectors carry additional information about participant age, beyond DAMT and/or QR scores. For each

latent feature type, we computed a stepwise regression predicting participant age from four independent variables and all possible interactions: participant reported gender, QR score, and coordinates on each dimension of the latent space (either human- or machine-derived). We included reported gender to capture the possibility that the relationship between drawing metrics and age may differ between participants identified as boys versus girls in the sample. As before, the stepwise procedure evaluated fits of increasingly complex models and selected a final model based on the lowest Akaike information criterion (AIC).

For both human- and machine-derived spaces, the stepwise procedure retained all four predictors and some second- and third-order interactions. The model fit using machine-derived latent features, however, did not explain more variance in age than that fit only to age, gender, and QR score (65%, $F(85) = 1.5$, $p = \text{n.s.}$), while adding DAMT to the regression reliably improved model fit (to 68%, $F(84) = 8.1$, $p < 0.01$). Thus machine-based vectors do not capture age information beyond that expressed by QR score, nor do they capture the relationship to age expressed by DAMT score. In contrast, models fit using human-derived latent features accounted for 76% of the variance in participant age, reliably more than the model fit without these predictors ($F(85) = 5.8$, $p < 0.0001$). Adding DAMT score to the regression as a simple effect did not reliably improve model fit – thus QR and human-derived features together fully express the relationship between DAMT score and age while also capturing additional variance. In other words, novel metrics based on naïve human judgments can more accurately predict participant age than does the DAMT score without missing any information, potentially suggesting a different approach to IQ estimation as explored below.

Predicting the original IQ estimate

We next assessed how well the original IQ score can be estimated via linear regression using the covariates of age and

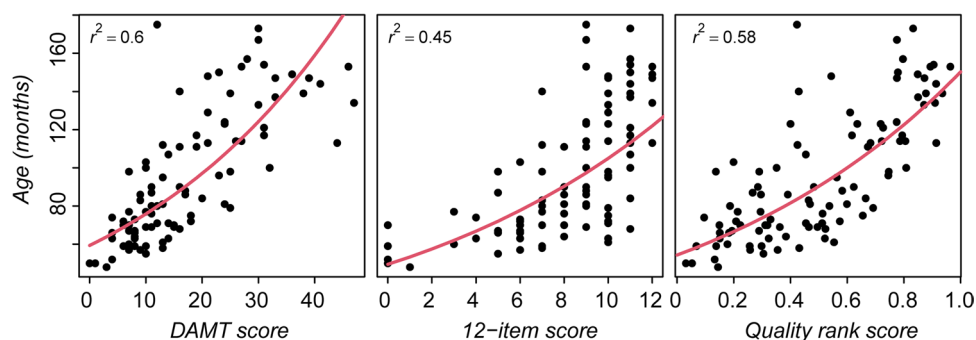


Fig. 6 Regressions predicting participant age from different measures of drawing quality. All three show an exponential relationship, but the original Draw-a-Man-Test (DAMT) score and the Quality-rank (QR) score account for more variation than does the 12-item checklist

gender together with both classical and contemporary drawings metrics. Table 2 provides an overview of the results.

Participant age alone accounted for 7% of the variance in IQ, reliably better than a null model ($p < 0.01$), and the inclusion of an interaction with gender did not account for significant additional variance. A stepwise regression fit to age, gender, and DAMT score retained all three terms and their interactions in the selected model, accounting for 89% of the variance in IQ. This result is not surprising given that the IQ metric was derived directly from DAMT score, but it suggests that, whatever the age-normalizing transform Goodenough applied to derive this estimate, linear models can reconstruct the estimate with quite good but not perfect accuracy.

We then considered whether contemporary measures of drawing quality could predict the original IQ estimate with comparable accuracy. A stepwise regression using the 12-point checklist retained this predictor, age, and their interaction, discarding sex. The resulting model explained only 52% of the variance in IQ. Models fit with QR score fared somewhat better: the stepwise procedure retained just the simple effects of age and QR, which accounted for 70% of variance in IQ.

Next, we assessed whether the human- or machine-derived feature spaces capture further information about the original IQ estimate. When included with age and gender, the stepwise procedure selected all predictors and multiple high-order interactions for both spaces, but in neither case did model fits exceed the fit of a model based solely on age and QR score (see Table 2). When DAMT scores were added to the stepwise procedure, however, fit improved dramatically. For both human and machine spaces, the procedure

returned a model with several high-order interactions that explained 96–97% of the variance in IQ – reliably more than the model based on age, gender, and DAMT scores alone ($F(66) = 8, p < 0.0001$), despite the greater number of estimated model parameters. This result suggests that both human- and machine-derived spaces capture information relevant to relating raw DAMT score to Goodenough's IQ estimate, beyond that already expressed by the DAMT score, age, and gender.

Finally, we evaluated how well the original IQ estimate could be reconstructed using only the novel metrics. We fit stepwise regressions predicting IQ from age, gender, QR score, and either the human- or machine-derived latent features. Models fit with human-derived features explained 81% of variance in IQ, reliably more than age, sex, and QR alone ($F(73) = 2.1, p < 0.02$). Models fit with machine-derived features explained 80% of variance, with the stepwise procedure selecting more predictors overall, so that the resulting model did not explain reliably more variance (given the number of parameters) than a model based only on age, gender, and QR score.

An alternative measure of IQ

In this final analysis, we consider whether estimates of IQ in this sample systematically change if we follow Goodenough's general approach – taking the ratio of estimated age from the drawing score to chronological age – but using latent structure and regression to predict the participant's age from their drawing in place of Goodenough's DAMT score and age-assigned bins. If this *adjusted estimate* leads to a qualitatively similar distribution of scores, highly correlated

Table 2 Results of stepwise regression predicting original IQ measure from various drawing metrics. "Full model" shows all predictors considered with asterisks indicating interactions. "Retained" shows which predictors were kept following stepwise selection. Asterisks indicate models where all simple effects and some interactions were included, with the number showing the highest-order interaction term in the selected model. "Contrast" shows the simpler model against

which each more complex model was evaluated to assess whether additional factors reliably improved fit. The remaining columns show the F and p values for the corresponding contrast. h1/h2 = the two human-derived coordinate vectors; m1/m2 = the two machine-derived coordinate vectors; damt = Draw-a-Man Test; qrs = Quality-rank score

	full model	retained	r^2	contrast	F (df)	$p <$
<i>covariates</i>	age	NA	0.07	null	7 (96)	0.01
	age* sex	NA	0.07	age	0.2 (94)	0.8
<i>drawing quality</i>	age * sex * damt	age * sex * damt	0.89	age* sex	162 (90)	0.0001
	age * sex * 12pt	age* 12pt	0.52	age	43 (94)	0.0001
	age * sex * qrs	age+ qrs	0.7	age	202 (96)	0.0001
<i>feature vectors</i>	age * sex * h1 * h2	***	0.64	age* sex	13 (84)	0.0001
	age * sex * m1 * m2	****	0.36	age* sex	3 (82)	0.001
<i>vectors and damt</i>	age * sex * damt * h1 * h2	*****	0.97	age * sex * damt	8 (66)	0.0001
	age* sex* damt * m1 * m2	*****	0.96	age * sex * damt	8 (66)	0.0001
<i>vectors and qrs</i>	age * sex * qrs * h1 * h2	****	0.81	age * sex * qrs	2.1 (73)	0.02
	age * sex * qrs * m1 * m2	****	0.8	age * sex * qrs	1.5 (70)	0.11

with the original, that would suggest that IQ estimated from drawings is robust to the precise technique for connecting drawing properties to age norms. If the adjusted estimate differs systematically and qualitatively from the original, that would instead suggest that improved methods for connecting drawing features to participant age can significantly impact the estimate of participant IQ, with implications that extend beyond this analysis of historical data.

The adjusted score relies on the same strategy of measuring IQ as the ratio between “mental age” estimated from a drawing and chronological age. Goodenough’s method involved computing means and variances of DAMT scores for participants binned by their chronological age in years, then assigning a “mental age” to children by placing them in the most likely *bin* based on their individual DAMT score. The preceding analyses showed that QR score and human-derived features together predict participant age significantly more accurately than does the DAMT score itself. Thus, to derive a contemporary estimate of IQ for Goodenough’s participants, we generated new “mental age” scores for each participant by applying the best age-predicting regression model to corresponding drawing metrics, dividing this predicted age by true chronological age, and multiplying by 100. We then compared the resulting scores and distributions to the original IQ estimate.

The left panel of Fig. 7 shows the relationship between the two measures for all drawings. Most lie above the identity line (diagonal), indicating a higher adjusted estimate

relative to the original. The mean estimated IQ from the adjusted method is significantly higher than the original estimate ($p < 0.001$ by paired *t*-test). Thus, Goodenough’s original metric may have significantly under-estimated IQ in this sample, a conclusion consistent with a comparison of the two score distributions (right panel). For Goodenough’s estimate, the 95% CI for the mean spans 89–96, reliably below the expected value of 100 ($t(97) = -3.8$, $p < 0.001$). By contrast the 95% CI of the mean for the adjusted estimate span 98–105, encompassing the expected value of 100. The adjusted measure suggests an IQ 9 points higher on average than the original measure.

The scatterplot also suggests a potential gender difference amongst participants with particularly high scores on the original measure. To test this, we fit regression models predicting contemporary IQ from original IQ alone, or with gender and its interaction included. Original IQ explained 39% of the variance in adjusted IQ ($p < 0.0001$ vs. null model), but this increased to 47% when gender was included, a reliable improvement ($F(94) = 6.2$, $p < 0.003$). The gender effect was attributable to participants with scores larger than 120 on the original metric: the same analysis performed only on participants with scores of 120 or lower showed no reliable difference between models fit with/without gender as a predictor. The result suggests that idiosyncrasies of the DAMT, while under-estimating IQ generally, may have over-estimated IQ for high-performing girls in this sample.

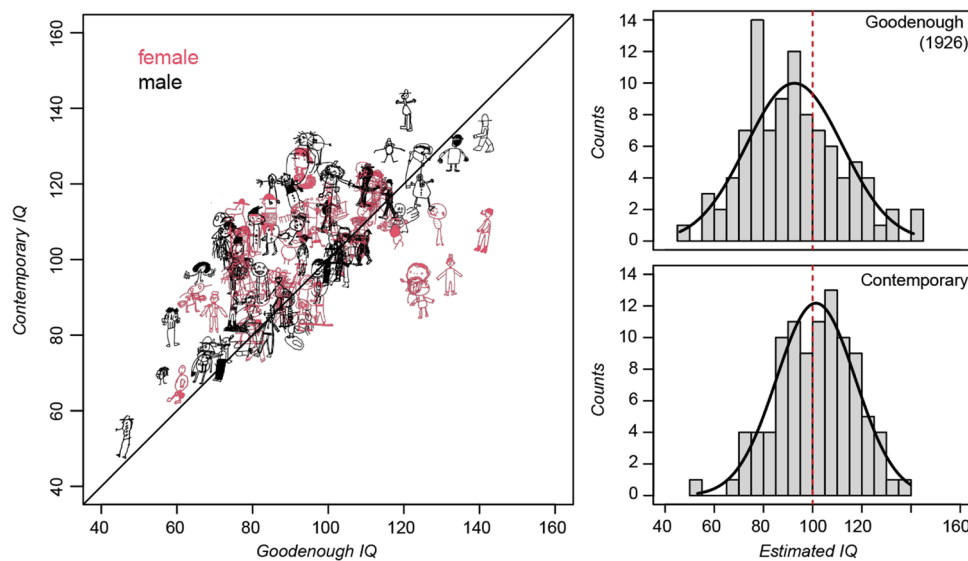


Fig. 7 Relationship between original and contemporary estimates of IQ from drawing. The left panel shows all drawings as a scatterplot colored by participant gender. Most drawings lie above the diagonal identity line, suggesting that Goodenough’s measure systematically under-estimated IQ. The plot also suggests a strong gender difference amongst participants with a high estimate on the original meas-

ure. Right panels show the distribution of scores in this sample with the best-fitting normal curve superimposed. The red line shows the expected mean value of 100. Goodenough’s metric shows a lower median and mean as well as a higher variance than the contemporary score

Perhaps more importantly, the analysis also shows that no more than about 45% of the variance in adjusted IQ is explained by the original metric. The new metrics suggest that latent structure in drawings changes more systematically over development than the DAMT can reveal (as evidenced by better prediction of participant age); exploiting this better predictive accuracy then yields a substantially changed picture of participant IQ in this classic dataset.

General discussion

In the preface to *Measurement of Intelligence by Drawings*, Florence Goodenough described her work as an experiment with a specific objective: to better characterize how children’s intellectual development may be assessed by their drawings. At the time of her writing, children’s drawings had already been the subject of several “big data” studies: Kerscheneister (1905), for instance, collected almost 100,000 drawings under standardized conditions! In a thorough review of prior work, Goodenough noted how earlier efforts would have benefited from a more objective and structured approach that incorporated modern statistical techniques. The current project was undertaken in this spirit, aiming to incorporate contemporary insights from computational cognitive science to characterize structure in children’s human-figure drawings, and applying the approach to Goodenough’s original data.

Consistent with a prior study of contemporary drawings (Jensen et al., 2023), we found that three novel metrics – human- and machine-derived latent feature spaces, and crowd-sourced rankings of drawing quality – together encode information about participant age, gender, and IQ. In all cases the new metrics captured substantially more information than did a 12-item checklist measure widely used in contemporary cognitive evaluation (Arden et al., 2014; Oliver & Plomin, 2007; Rimfeld et al., 2019; Saudino et al., 1998). In predicting age and gender, these new computational techniques fully captured information encoded in the DAMT score and explained additional variance as well. Though the new metrics alone could not fully recover Goodenough’s original IQ estimates without inclusion of the DAMT score, an alternative analysis leveraging the higher accuracy of regression models to predict participant age suggested that the original estimates may have significantly mischaracterized IQ in this sample (understood as the ratio of estimated “mental age” to chronological age).

Also consistent with the prior study, we found that human- and machine-derived latent feature vectors encoded complementary information about participant characteristics: machine-derived features fared better at predicting participant gender, while human-derived features better captured age information. This difference reflects the initial

motivation for considering the two approaches. Deep neural networks learn complex features via training on millions of photographs. The resulting image representations express rich and highly abstract structure, as evidenced by the success such models show in generalizing to new images (Krizhevsky et al., 2012; Yu et al., 2022) and capturing similarity relations between photographs and drawings of the same item (Fan et al., 2018). Yet the features they exploit and the resulting similarity structures they encode can be difficult for human beings to explicitly discern – thus deep networks provide one avenue for deriving structured representational spaces different from those derived from human judgments. In contrast, human-derived spaces can express the overall Gestalt similarities that people “see” in drawings, potentially encompassing many factors beyond what neural networks currently capture: knowledge about parts, categories, functions, orientations, and so on, that can all jointly contribute to perceived similarity. The current results suggest that both kinds of spaces can carry distinct, complementary information about participant characteristics, and so can contribute to the use of drawing in cognitive assessment.

The concordance of current results (using a curated historical sample of human figure drawings) with the prior study (using recent children’s drawings collected across differing media types; Jensen et al., 2023) suggests that the utility of these metrics is not restricted to drawings collected in a particular time and place. In addition to being distant in time, Goodenough’s data were collected across a range of cultural, linguistic, racial, and socioeconomic communities. While many of the drawings are clearly identifiable as belonging to a particular era – note all the hats, guns, and tobacco! – nevertheless, the same contemporary techniques can capture demographic and cognitive information in drawings regardless of the era in which they were collected.

We note that the new metrics were computed without any expert knowledge and with comparatively little human labor. Machine-derived embeddings can be computed automatically from any image, raising the possibility of generating such spaces even for very large image datasets. The human-derived metrics were generated from crowd-sourced comparative judgments, using about 10,000 trials for both QR score and latent feature vectors. A human worker can complete 200 trials in about 10 min – thus the total human time invested for each metric was about 8 h and 20 min for 98 images. A trained rater takes about 15 min to score a drawing using the 51-item DAMT, or a total of 24 h and 30 min for 98 drawings. Given the tight relationship between QR and DAMT scores, this means that essentially the same information can be recovered with one-third the human effort, and without special knowledge or training. Moreover, because the forced-choice decisions can be collected in parallel, the required data collection can be completed in just an hour or two.

Implications for contemporary use of drawings in cognitive assessment

Some aspects of Goodenough's work can seem archaic by contemporary standards, and it is worth considering what implications, if any, the current analysis has for the use of children's drawing in cognitive assessment. For instance, Goodenough presumably hewed to the contemporaneous understanding of gender as a binary natural kind, intimately connected to biology. Her observation of higher DAMT scores amongst girls versus boys of a given age spurred a broad tradition of research seeking to understand the causes and implications of these differences (Cox, 1993; Cox et al., 2001; Goodenough & Harris, 1950; Harris, 1963; Lange-Küttner, 2011; Lange-Küttner et al., 2002; Naglieri, 1988; Picard, 2015), which in turn is part of a legacy of developmental research rife with claims and counter-claims about innate gender differences, or the absence thereof, in various different aspects of cognition (e.g., Hyde, 2016; Hyde & McKinley, 1997). Today gender identity is viewed as a non-binary and potentially multi-factorial construct, shaped by psychology and culture, and independent of, but with a complex relationship to, biology (Hyde et al., 2019; Thorne et al., 2019). Given this more nuanced view, what is the point of asking whether children identified by their parents as male or female in 1926 differ in the way they produce drawings?

One reason is that, despite extensive prior work focusing on gender differences in human figure drawing, it has not been clear to what extent such differences hinge on the particular features raters are explicitly asked to score in checklists, or on the cultural norms of a given time and place. Prior work showed that both machine- and human-derived embeddings of contemporary human-figure drawings carried reliable information about participants' identified gender (Jensen et al., 2023); however, in that dataset, (a) participants were at liberty to draw whatever gender they liked regardless of their own relative masculinity and/or femininity, and were likewise able to include or omit any gender-aligned distinctions in their drawings, and (b) the human-derived embeddings were collected from participants belonging to the same time and culture as the children who produced the drawings, and so might be particularly sensitive to such signifying. The current results show that drawings collected in a culturally and temporally distal population, from children instructed specifically to depict an adult male, nevertheless possess latent structure reliably different in children identified as male versus female. The features encoding this structure were extracted from a machine vision model trained to classify modern photographs and applied in the current study to drawings collected over a 100 years ago from children selected specifically to capture then-present cultural differences. Human raters judging image quality and similarity were at a long temporal and cultural remove from the children who

produced the drawings; nevertheless these judgments carried information about participant gender. The result suggests that children of differing gender identities depict the human form in systematically different ways that do not depend on the particular choices of a checklist designer or close knowledge of the child's cultural/temporal norms.

The current analyses have little to say about the causes of the observed differences. It is possible that they arise in part or in whole from acquired gender norms that shape how children choose to depict human figures. Goodenough suggested, for instance, that boys may depict a figure engaged in activities that were culturally gendered in 1926, such as smoking a pipe or fighting a battle. Such activities are difficult to depict in a face-on figure, potentially leading boys to choose profile-based orientations more often than girls. In our study of more recently collected drawings, depictions in profile were exceedingly rare, as was the inclusion of accessories such as ties, pipes, hats and guns (Jensen et al., 2023).

It is also the case that drawing, like language, requires the integrated coordination of several cognitive faculties, including perception, conceptual knowledge, spatial abilities, memory, planning/executive function, and motor control. These abilities develop along diverse trajectories that reflect the interplay of experience and biology, leading to potentially highly complex profiles of change over time (Case & Okamoto, 1996; Elman, 1996; Goswami & Bryant, 2012; Karmiloff-Smith, 2009). The early emergence of such differences in children's drawings may provide an avenue for adjudicating or better understanding some of these patterns, especially if it is possible to bring methods like those we have described to bear on datasets collected across many different eras and cultures. If such differences solely reflect the socio-cultural milieu, together with learned attitudes about gender expression, the properties of drawings that "work" to differentiate drawings produced by male, female, or non-binary-identifying children should not generalize well across cultural/temporal cohorts.

In a similar vein, although Goodenough endeavored to establish an instrument that could reflect cognitive abilities in a largely nonlinguistic and culturally unbound manner by collecting drawings from a sample of children that were both more closely connected to their ancestral heritage and spanned social and economic strata, it is reasonable to question the broad generalizability of the DAMT and subsequent checklist-based approaches to the assessment of human figure drawings. Given that drawing conventions and practices vary across cultural and temporal contexts (Cox, 1993, 1998; Cox et al., 2001; Kellogg, 1969), it is hard to conceive of a feature-based checklist that would both include enough informative detail and successfully exclude potentially biased additions, allowing for a single valued score that could accurately be assigned to drawings produced by the full diaspora of children living in the USA in the 1920s,

let alone all the world's children living in different points in history. Those limitations have inspired both a legacy of updates to the checklist approach that Goodenough pioneered (e.g., Harris, 1963; Kellogg, 1969; Koppitz, 1968; Machover, 1949; Naglieri et al., 1991), and encouraged our own reconsideration of both the DAMT, the focus of this paper, and human figure drawing assessment processes based on checklists more generally (Jensen et al., 2023). Indeed, although the present study demonstrates that our contemporary approaches – human- and machine- derived embeddings temporally uncoupled from the period in which the drawing was created – can uncover latent structure within a relatively small historical dataset children's drawings, it remains to be explored whether the novel approaches and findings described would generalize to non-Western and/or larger historical samples that more fully reflect the diversity of ways that children represent the human figure through the act of drawing.

Finally, it is worth noting that the images in this dataset are not a random sample of those collected but were chosen to exemplify particular observations that Goodenough thought useful for training raters. Since Goodenough was interested in the possibility of gender differences, it is possible that she selected images exemplifying distinctions that she believed to be present in the sample, and that the current methods succeed only due to that sampling confound. To us this seems unlikely, given that the male and female participants in this sample did *not* differ reliably on Goodenough's own metric (the DAMT) – the gender differences Goodenough reported were observed instead in the fuller and more representative sample, suggesting that the images included in the book may have downplayed such differences, if anything. Nevertheless, future work should look to more representative datasets to adjudicate this question.

Implications for using drawings to understand patterns of cognitive change in development

Goodenough emphasized drawing as a means of estimating mental function because it is near-universally practiced by children regardless of their language, culture, gender, race, education, socio-economic status, or geographic situation. Thus, she viewed the activity as providing a more level playing field for assessment than other metrics that might depend to a greater degree upon the idiosyncrasies of one's circumstances, language, or cultural milieu. Of course, children differ widely in the opportunities they have to engage in drawing activities, the materials to hand, extent of parental support and demonstration, and explicit instruction – the playing field for drawing is hardly level in that sense. However, the same can be said for measures based on linguistic competence, reasoning, mathematics, and essentially all other cognitive/behavioral tasks commonly used in assessment.

We believe drawing to be particularly useful in the context of understanding diversity in the development of children's cognition. Consider that the verbal components of contemporary measures (such as the Weschler Intelligence Scale for Children [WISC]; Wechsler, 1939; Weiss et al., 2019) rely on tasks like explaining a word's meaning and comprehending spoken or written text – activities that children are likely to engage in as part of everyday life and schooling. In contrast, common non-verbal IQ measures include tasks such as pictorial analogy, image pattern completion, block-copying, timed replication of block arrangements or geometric designs, working memory span, etc. – tasks intended to tap certain cognitive abilities but which the child is unlikely to have encountered, let alone practiced extensively or used in daily life (see, e.g., the review of non-verbal IQ measures by DeThorne & Schaefer, 2004). In this sense, common IQ measures may disadvantage less-verbal children: only the verbal subscales incorporate tasks familiar and well-practiced to the child. Drawing, in contrast, is a non-verbal task that shares many properties with language: it draws upon perception, memory, knowledge, planning, sequential coordination, goal-monitoring, cognitive control, and motor function, is used to communicate meaning, and is an activity familiar to most children that they are likely to pursue in their own daily life. Thus, drawing incorporates many cognitive and functional aspects of language without the language, potentially providing a more useful and ecologically valid measure of nonverbal cognition.

The central problem for realizing this possibility has been how best to measure the rich structure appearing in the drawings that children produce. The checklist approach Goodenough pioneered (and its modern variants) boils this down to a single number – essentially a one-dimensional measure of the drawing's quality. Our results show that multivariate structure uncovered by human and machine vision relates to characteristics of the child in ways not expressed by a single quality-based metric.

This observation has critical implications for the use of drawings in studies designed to understand patterns of change, independence, mutual influence, and genetic effects in cognitive faculties over the course of development. Such studies typically rely on understanding patterns of covariation and independence over time via, for instance, structural equation modeling, across tasks thought to tap different core cognitive abilities. A key goal is to estimate whether a given predictor explains significant variance in an outcome when the effects of other possible causes have been partialled out. We know from studies like TEDS (Arden et al., 2014) that a drawing checklist score obtained early in life predicts significant variance in IQ later in life, over and above other factors. If, however, richer drawing metrics explain greater variation in outcomes (including IQ but also other important indicators of cognitive functioning), this means there is less residual variance, and

different patterns of residual variation, remaining for other candidate causal factors to explain. In other words, replacing a one-dimension drawing score with the richer metrics we have described can potentially alter, not just the putative relationship between drawing and other abilities, but the full pattern of predictive relationships observed across a suite of different assessments. Testing this possibility will, of course, require introduction of the new measures into much richer datasets, but the current study shows that, because drawings leave a trace of the child's state that persists through time, such introduction can be done post hoc on any dataset that includes copies of the original drawings. In the present study, we followed upon prior work using a well-studied (Fan et al., 2018; Jensen et al., 2023; Jha et al., 2023; Long et al., 2021) convolutional neural network, VGG-19 (Simonyan & Zisserman, 2014). Future work could consider whether alternative machine vision models capture similar or somewhat different varieties of structure.

Finally, following Goodenough, we have focused on whether/how information contained in children's drawings might be used in cognitive assessment – but we do not mean to suggest that drawings alone are sufficient to this task or that other metrics are inherently flawed. Instead we believe that structure in drawings can provide a complementary source of information to other common metrics that may better capture how patterns of developmental change across verbal and nonverbal aspects of cognition. It remains an empirical question whether richer characterization of structure in children's drawings capture patterns of variation that elude other measures; answering this question will require application of methods like those we have developed to much larger longitudinal datasets.

Conclusion

Human figure drawings have long been used as a simple, fun, and ecologically valid tool for estimating the development of cognition over childhood. Variants of the checklist method pioneered by Florence Goodenough remain in widespread use today in the assessment of drawings, and are able to capture important information about a child's mental abilities in a manner that taxes many cognitive faculties but does not rely on language. The current work shows, however, that drawings encode rich, multi-dimensional structure beyond what checklists can capture, and that such structure relates more systematically to a child's demographic and mental characteristics than checklists have suggested. These patterns extend across time and culture, can be efficiently measured with contemporary computational/behavioral methods, and offer a new window into patterns of cognitive change over time in the developing child.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13421-024-01557-0>.

Acknowledgements We thank Michelle Golden, Katrina Nicholas, and Maria Schapfel for their help in image processing.

Data availability The data and materials for all experiments are available at: <https://github.com/ClintJensen/DrawingsProject>.

References

- Arden, R., Trzaskowski, M., Garfield, V., & Plomin, R. (2014). Genes influence young children's human figure drawings and their association with intelligence a decade later. *Psychological Science*, 25, 1843–1850. <https://doi.org/10.1177/0956797614540686>
- Bruininks, R. H., & Bruininks, B. B. (2005). *Bruininks-Oseretsky test of motor proficiency* (2nd ed.). Pearson Assessment.
- Burn, R. C., & Kaufman, S. H. (1970). *Kinetic family drawings (K-F-D): An introduction to understanding children through kinetic drawings*. Burner/Mazel.
- Case, R., & Okamoto, Y. (1996). The role of conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development*, 61, v–265. <https://doi.org/10.2307/1166077>
- Cox, M. V. (1993). *Children's drawings of the human figure*. Psychology Press.
- Cox, M. V. (1998). Drawings of people by Australian Aboriginal children: The inter-mixing of cultural styles. *Journal of Art & Design Education*, 17(1), 71–79.
- Cox, M. V., Koyasu, M., Hiranuma, H., & Perara, J. (2001). Children's human figure drawings in the UK and Japan: The effects of age, sex and culture. *British Journal of Developmental Psychology*, 19(2), 275–292.
- De La Serna, M., Helwig, L., & Richmond, B. (1979). Cultural impact on human figure drawings. *Social Behavior and Personality: An International Journal*, 7(1), 29–32.
- de Lemos, M., & Doig, B. (1999). *Who Am I?: Developmental Assessment*. The Australian Council for Educational Research Ltd.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.
- DeThorne, L. S., & Schaefer, B. A. (2004). A guide to child nonverbal IQ measures. *American Journal of Speech-Language Pathology*, 13(4), 275–290.
- Elman, J. L. (1996). *Rethinking Innateness: A connectionist perspective on development* (10th ed.). MIT press.
- Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, 42(8), 2670–2698.
- Frankenburg, W. K., Dodds, J., Archer, P., Shapiro, H., & Bresnick, B. (1992). The Denver II: a major revision and restandardization of the Denver Developmental Screening Test. *Pediatrics*, 89(1), 91–97.
- Fury, G., Carlson, E. A., & Sroufe, A. (1997). Children's representations of attachment relationships in family drawings. *Child Development*, 68(6), 1154–1164.
- Goodenough, F. (1926). *Measurement of intelligence by drawings*. World Book.
- Goodenough, F. L., & Harris, D. B. (1950). Studies in the psychology of children's drawings: II 1928–1949. *Psychological Bulletin*, 47(5), 369–433. <https://doi.org/10.1037/h0058368>
- Goswami, U., & Bryant, P. (2012). Children's cognitive development and learning. *The Cambridge primary review research surveys* (pp. 141–169). Routledge.

- Hammer, E. F. (1958). *The clinical application of projective drawings*. Carl C Thomas.
- Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity*. Harcourt Brace Jovanovich.
- Harrison, L. J., Clarke, L., & Ungerer, J. A. (2007). Children's drawings provide a new perspective on teacher–child relationship quality and school adjustment. *Early Childhood Research Quarterly*, 22(1), 55–71.
- Hyde, J. S., & McKinley, N. M. (1997). Gender differences in cognition: Results from meta-analyses. In P. J. Caplan, M. Crawford, J. S. Hyde, & J. T. E. Richardson (Eds.), *Gender differences in human cognition* (pp. 30–51). Oxford, England: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195112917.003.0002>
- Hyde, J. S. (2016). Sex and cognition: Gender and cognitive functions. *Current Opinion in Neurobiology*, 38, 53–56.
- Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, 74(2), 171–193. <https://doi.org/10.1093/acprof:oso/9780195112917.003.0002>
- Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. (2015). NEXT: A system for real-world development, evaluation, and application of active learning. *Advances in Neural Information Processing Systems*, 28, 2656–2664.
- Jensen, C. A., Sumanthiran, D., Kirkorian, H. L., Travers, B. G., Rosengren, K. S., & Rogers, T. T. (2023). Human perception and machine vision reveal rich latent structure in human figure drawings. *Frontiers in Psychology*, 14, 1029808.
- Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive Science*, 47(1), e13226. <https://doi.org/10.1111/cogs.13226>
- Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, 34(1), 57–83.
- Karmiloff-Smith, A. (2009). Nativism versus neuroconstructivism: Rethinking the study of developmental disorders. *Developmental Psychology*, 45(1), 56.
- Kellogg, R. (1969). *Analyzing children's art*. National Press Books.
- Kerschensteiner, G. (1905). *Die Entwicklung der zeichnerischen Begabung: Neue Ergebnisse auf Grund neuer Untersuchungen [The development of the talent for drawing: New results due to new investigations]*. Gerber.
- Koppitz, E. M. (1968). *Psychological evaluation of children's human figure drawings*. Grune and Stratton.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.
- La Voy, S. K., Pedersen, W. C., Reitz, J. M., Brauch, A. A., Luxenberg, T. M., & Nofsinger, C. C. (2001). Children's drawings: A cross-cultural analysis from Japan and the United States. *School Psychology International*, 22(1), 53–63.
- Lamprecht, K. (1906). *Les dessins d'enfants comme source historique [Children's drawings as a historical source]*. Hayez.
- Lange-Küttner, C. (2011). Sex differences in visual realism in drawings of animate and inanimate objects. *Perceptual and Motor Skills*, 113(2), 439–453.
- Lange-Küttner, C., Kerzmann, A., & Heckhausen, J. (2002). The emergence of visually realistic contour in the drawing of the human figure. *British Journal of Developmental Psychology*, 20(3), 439–463.
- Lobsien, M. (1905). Kinderzeichnung und Kunstkanon [Children's drawing and canon of art]. *Zeitschrift für Pädagogische Psychologie*, 7, 393–404.
- Long, B., Fan, J., Chai, Z., & Frank, M. C. (2021). Parallel developmental changes in children's drawing and recognition of visual concepts. Preprint at PsyArXiv <https://doi.org/10.31234/osf.io/5yv7x>
- Lynn, R. (2013). Who discovered the Flynn effect? A review of early studies of the secular increase of intelligence. *Intelligence*, 41(6), 765–769.
- Machover, K. (1949). *Personality projection in the drawing of the human figure: A method of personality investigation*. Carl C Thomas.
- McCarthy, D. (1972). *McCarthy scales of children's abilities (MSCA)*. Psychological Corporation.
- Naglieri, J. A. (1988). *Draw A Person: A Quantitative Scoring System*. Psychological Corporation.
- Naglieri, J. A., MacNeish, T. J., & Bardos, A. (1991). *Draw a person: Screening procedure for emotional disturbance; dap: sped*. Pro-Ed.
- Oliver, B., & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics*, 10, 96–105.
- Picard, D. (2015). Sex differences in scores on the draw-a-person test across childhood: Do they relate to graphic fluency? *Perceptual and Motor Skills*, 120(1), 273–287.
- Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., Dale, P. S., Eley, T. C., & Plomin, R. (2019). Twins early development study: A genetically sensitive investigation into behavioral and cognitive development from infancy to emerging adulthood. *Twin Research and Human Genetics*, 22(6), 508–513.
- Rouma, G. (1913). *Le langage graphique de l'enfant [The graphic language of the child]*. F. Alcan & Lisbonne.
- Rueda, F. J. M., Noronha, A. P. P., Dos Santos, A. A., Jesuino, A. D. S. A., Zuanazzi, A. C., Ferraz, A. S., Costa, A. R. L., & Otoni, F. (2020). Drawing the human figure: Systems most used in cognitive assessment of children. *Psico*, 51(1), e31313–e31313.
- Saudino, K. J., Dale, P. S., Oliver, B., Petrill, S. A., Richardson, V., Rutter, M., ..., & Plomin, R. (1998). The validity of parent-based assessment of the cognitive abilities of 2-year-olds. *British Journal of Developmental Psychology*, 16(3), 349–362.
- Sievert, S. (2019). Salmon documentation. *GitHub*. <https://github.com/stsievert/salmon>
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*.
- Squires, J., & Bricker, D. (2009). *Ages and Stages Questionnaire: A parent-completed child monitoring system* (3rd ed.). Brookes Publishing Company.
- Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on Machine Learning* (pp. 673–680).
- Thorne, N., Yip, A. K. T., Bouman, W. P., Marshall, E., & Arcelus, J. (2019). The terminology of identities between, outside and beyond the gender binary—A systematic review. *International Journal of Transgenderism*, 20(2–3), 138–154.
- Verworn, M. (1908). *Zur Psychologie der primitiven Kunst: ein Vortrag [On the psychology of primitive art: A lecture]*. G Fisher.
- Wechsler, D. (1939). *The measurement of adult intelligence*. Williams & Wilkins.
- Weiss, L. G., Saklofske, D. H., Holdnack, J. A., & Prifitera, A. (2019). *WISC-V: Clinical Use and Interpretation*. Elsevier Science.
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). *Coca: Contrastive captioners are image-text foundation models*. <https://arxiv.org/abs/2205.01917>
- Zee, M., Rudasill, K. M., & Roorda, D. L. (2020). “Draw Me a Picture” student-teacher relationship drawings by children displaying externalizing, internalizing, or prosocial behavior. *The Elementary School Journal*, 120(4), 636–666.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.