Title
Naive human judges can accurately predict expertise in children's block building. Can embedded motion sensors do just as well?

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Davis, Emory
Wang, Zihan
Malpani, Anand
et al.

Publication Date
2023

Peer reviewed

Naïve human judges can accurately predict expertise in children's block building. Can embedded motion sensors do just as well?

E. Emory Davis[1] (emorydavis@jhu.edu), Zihan Wang[1] (zwang215@jhu.edu), Anand Malpani[2] (amalpan1@jhu.edu), Gregory Hager[3] (hager@cs.jhu.edu), Amy Lynne Shelton[4] (ashelton@jhu.edu), and Barbara Landau[1] (landau@jhu.edu)

[1]Department of Cognitive Science
[2]Malone Center for Engineering in Healthcare
[3]Department of Computer Science
[4]School of Education
Johns Hopkins University
Baltimore, MD 21218

Abstract

Motion quality can differentiate experts from novices in fields like surgery (Ershad et al., 2018). We extend approaches used by researchers in that field to examine the relationship between motion and skill in a children's block-building task. We ask whether the relationship between these two variables is detected equally well by humans and machines—in this case, motion sensors embedded in the blocks. We investigate whether adults' judgments about motion quality and children's overall building skill reflect children's actual construction ability, and whether data from embedded motion sensors predict children's skill as well as adult judgments do. We find that human raters outperform the motion sensor data. Our findings raise questions about how people form such intuitive judgments of expertise, and how automated judgments of skill can be enhanced to more accurately predict expertise in block building and other similar cognitive tasks.

Keywords: block building, spatial skills, motor task, automated assessment

Introduction

Spatial skills such as block construction are ubiquitous among young children and are some of the first activities practiced by children in informal settings. These skills have been shown to predict many aspects of academic and career achievement (Kell et al., 2013; Sorby, 1999; Stannard et al., 2001; Verdine et al., 2014; *inter alia*), and they are often evaluated using standardized tests that require in-person assessment using specialized materials, such as the Test of Spatial Ability (TOSA) or the Differential Ability Scales (DAS) pattern matching task. In this paper, we look to another source of evaluation, building on observations from a very different domain – robotic surgery – to explore the possibility that children's block-building skill can be judged quite accurately by untrained observers by simply watching as the child builds. We contrast this simple, intuitive means of evaluating a builder's expertise with one that is derived from physical measurements of the builder's motions using sensors embedded in individual blocks. Sensor-derived metrics of motion have been shown to correlate with expertise in robotic surgery (Malpani et al., 2015), and we ask here whether these metrics do as well as, better than, or worse than human intuitive judgments of expertise. To preview, we find that naïve observers' intuitive judgments of a child's 'expertise' in block building predict metrics of actual building success better than the sensor-derived motion data measured as the build unfolds.

The idea that block-building expertise (or expertise in any complex perceptual or cognitive task involving action) can be evaluated automatically, by objective measures of ongoing motion, is appealing. Certainly, automation of performance assessment for various spatial and motor tasks and abilities is increasingly becoming a reality. Developmental researchers have developed a variety of automatic motion tracking tools to evaluate different aspects of motor development in children (e.g., Ossmy et al., 2022). In sports, human line judges are being replaced by computers in professional tennis, and artificial intelligence tools are being developed for judging Olympic sports like diving. While many of these computer-based assessments are proposed as a bias-free, more reliable, and scalable alternative to human judgments, computers sometimes fail to meet the human standard. For example, a trained physician using a robotic surgery device can be evaluated as a 'novice' or 'expert' on the basis of the information about the motions carried out while the surgeon manipulates the wands of the robotic surgery system from sensors that measure these motions (Ershad et al., 2018), demonstrating that motion data can provide some insight into expertise in skilled motor tasks. But these studies have also shown that humans are better able to distinguish differences between surgeons with varying levels of expertise than the motion data from the robotic device (Ershad et al., 2018). Research on crowd-sourcing assessments of robotic surgery skill has demonstrated that laypeople with no medical experience are remarkably good at judging robotic surgical skill, simply by viewing the actions taken by the surgeon as they manipulate the wands. In fact, they are as good at assessing skill as expert surgeons who view trainees as they carry out surgeries, and they surpass machines (Chen et al., 2014; Ershad et al., 2016, 2018; Kowalewski et al., 2016; Malpani et al., 2015; White et al., 2015). These crowd-sourced surgical assessments have looked not just at estimates of overall skill (Malpani et al., 2015), but also

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

ratings of stylistic descriptions of the quality of the surgeon's movements, such as whether they are smooth vs. rough or crisp vs. jittery (Ershad et al., 2018). This indicates that even untrained humans are picking up on something in a surgeon's movements that machines (currently) cannot.

Robotic surgery represents just one example of a highly trained cognitive and perceptual motor skill. Others include knife skills in cooking, brush or sculpting skills in art, dance techniques, maneuvers in gymnastics and figure skating, and even block-building skills in children—the focus of our paper. While we look towards the day when automated systems might be able to automatically detect expertise, here we take a first step by asking whether naïve adult judges can do as well or better than machine-generated measurements to predict the actual accuracy shown by a child as they build block structures.

To do this, we adapted methodology from the robotic surgery assessment literature, and combined that with existing data from a children's block-building task (Cortesa et al., 2018; Landau et al., 2022). We examined whether naïve raters' judgments about a child's block-building ability reflect the child's actual performance, and whether such judgments better predict performance than data derived solely from block-embedded sensors that record a range of properties of the motions carried out by the child as they build. We asked two specific questions. First, can naïve judges make accurate judgments of a child's block-building skill? Second, can the physical properties of motion detected by sensors implanted in the blocks produce data that do as well as our human judges in predicting children's performance? To answer these questions, we asked adults to watch video clips of children building a Duplo block structure, with no information about the target structure or how well the child actually performed in duplicating the structure. In Experiment 1, for each video clip, a group of adults was asked to rate the quality of the children's movements (were the movements fluid? fast? jerky?) and then to estimate the child's overall skill level. In Experiment 2, a different group of adults watched the videos and only estimated the child's overall skill, without judging the quality of their motions. We asked how accurately the adults' judgments of overall skill level (both with and without motion judgments) mapped onto a number of building performance metrics. Finally, we compared the accuracy of these judgments of performance to those from sensor-derived measurements of motions made by children while building.

Experiment 1

Methods

Participants Participants were 45 undergraduate students at Johns Hopkins University. A university ethical review board approved all study procedures, and participants provided informed consent and received course credit for their participation.

Design, Stimuli and Procedure All participants watched 8 25-second video clips of children executing a block copying task. After viewing each video, participants were asked first to rate the qualities of the child's motions while building, and then to rate the child's overall block-building skill.

In the original building task (see Cortesa et al., 2018; Shelton et al., 2022; Landau et al., 2022), 34 children were provided with 6 different Duplo block models and asked to build a copy of each one. The models consisted of 4, 6, or 8 blocks; half of the models were symmetrical and half asymmetrical. Children's building was recorded using an overhead camera; these videos were then coded and analyzed step-by-step, resulting in several different measures of children's performance on the task, which we use as the measures of children's block-building skill in the present study (see Results below). For more detail on the block building task, coding process, and the metrics used to assess building, see Shelton et al. (2022) and Landau et al. (2022).

Children's movement of the blocks during the task was captured by inertial measurement units (IMUs) embedded in the blocks provided to the children to build their copy. The IMUs contained an accelerometer and gyroscope and measured six aspects of the movement of the blocks: average acceleration (average magnitude of the block's acceleration), average jerk (the rate of change of acceleration; higher values indicate jerkier movements), number of peaks in acceleration per second (indicates indecisiveness in planning), number of peaks in jerk per second, average angular velocity (indicates motion efficiency), and average angular acceleration (indicates planning in orienting the blocks).

Fifty-three videos from the block copying study were used in the current study. Videos were only selected if they lasted at least 25 seconds from the first blocks connection until the end of the child's attempt to copy the model (when they said they were 'done'). We did not include any videos of children building the 4-block models, because the large majority of those videos did not meet this criterion. All videos were edited to begin at the point when the child connected the first two blocks of the construction and proceeded for 25 seconds following that point. Videos were cropped to show the build area where the child was working, but to occlude areas of the table outside of the construction area. A white rectangular shape was inserted digitally into the video frame to cover the part of the construction area that contained the model being copied (Figure 1). Masking the model was done to allow the viewer to evaluate the builder's movements without also evaluating their construction accuracy.

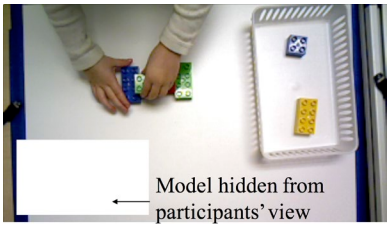

Figure 1: Example of a video clip shown in the motion and skill rating task.

Participants sat at a desk in front of a desktop computer, and an experimenter assisted with presenting the 8 video clips, 2 for each of the symmetrical and asymmetrical 6- and 8-block models. The selection of video clips was semi-randomized to ensure each video was rated by at least 7 participants. Video presentation was ordered so that participants did not see two videos of the same model in a row. Video order was randomized across participants.

Participants were instructed as follows: "You will be watching videos of children building with blocks; it does not matter what they were building. We would like you to pay attention to the way they were building with the blocks, particularly the quality of their movements. After watching each video, you will be asked to evaluate how well a variety of words describe the child's movements in this video. The videos are cropped so they may start or end suddenly, and that is OK." Following each 25-second video, participants were shown a questionnaire displayed on the computer screen. For each video, participants rated motions on the following 20 words: fast, controlled, rough, frantic, careful, fluid, careless, precise, relaxed, confused, anxious, confident, smooth, slow, cautious, quick, unsure, calm, sluggish, and hesitant.[1] These words were displayed in the questionnaire in randomized order. Participants rated on a scale of zero to ten, where zero indicates that the word is not at all descriptive of the child's movements and ten indicates that the word is very descriptive. After the motion ratings, participants were asked to guess the age of the child in the video. Next, they were asked to rate the child's construction ability compared to other children his/her age, on a scale of one to ten, with one indicating "poor, much worse than peers" and 10 indicating "excellent, much better than peers." Finally, participants were asked to guess whether the child was male or female.

Analysis Sixteen videos were excluded from final analysis for having fewer than 7 raters each, and another 13 videos were excluded because there was no motion data from the IMUs for those trials. This resulted in a total of 24 videos (featuring 18 child builders) included in our analyses, with 7 ratings per video from 33 participants.[2] There were 6 videos of the 6-block symmetric model, 7 videos of the 6-block asymmetric model, 5 videos of the 8-block symmetric model, and 6 videos of the 8-block asymmetric model.

Statistical analyses were conducted in R (R Core Team, 2021). Linear mixed-effects models were conducted using the *lme4* package (Bates et al., 2015). The ratings for the motion description words and the overall estimates of builder skill were normalized (z-scored), and we used these normalized ratings in all analyses.

Our goal was to address the question of whether naïve human raters can better assess a child's actual skill in block building than data from embedded sensors measuring properties of the children's motions during building. Accordingly, our analyses first examined the relationship between human judgments of overall skill and motion quality and children's actual building performance. Then, we analyzed the relationship between the motion data obtained from the IMUs and children's actual building performance. We also looked at whether there was any overlap between the adult judgments and the IMU data.

For the children's building performance, we used five building performance metrics from the analyses of builder skill reported in Landau et al. (2022) and Shelton et al. (2022). Those metrics were Accuracy (the proportion of correct states in a build path – i.e., those that could lead to a correct final copy), Excess Steps (the number of steps taken by the child that exceeded the minimum number required to build a correct copy), average Action Duration (in seconds, the average amount of time taken to make a block placement), Layering (the extent to which a child built their copy layer-by-layer from the bottom up, an approach associated with better building outcomes), and Stability (the proportion of stable states created in a build path, also associated with better building outcomes).

Results

How good are adults' judgments of children's block-building skill? We began by addressing the question of whether adults' judgments of children's construction ability reflect how well children did on the building task. We called the adults' overall judgments of skill 'Ability Guess'. Specifically, we looked at whether adult raters' Ability Guess scores and their ratings for individual motion descriptors predicted children's performance. We conducted individual linear mixed-effects models for each of the five performance metrics (Accuracy, Excess Steps, Action Duration, Layering, and Stability), with the performance metric as the dependent variable, Ability Guess and block model as main (fixed) effects, and child participant as a random slope. Ability Guess significantly predicted Stability ($\beta = 0.003$, $SE = 0.001$, $p < 0.05$), and was a marginally significant predictor of Layering ($\beta = 0.02$, $SE = 0.01$, $p = 0.07$). A Pearson's correlation with Holm correction for multiple comparisons also showed significant correlations between Ability Guess and all five performance metrics (see Figure 2); a higher Ability Guess was correlated with greater Accuracy, Layering, and Stability, while a lower Ability Guess was correlated with fewer Excess Steps and shorter Action Duration.

[1] The words used for this evaluation were based on task in which we asked a separate set of adult participants ($N = 7$) to provide free-response descriptions of the construction movements of child builders in a set of 19 different construction videos. We then chose the 20 most common words generated in this free-response task, after eliminating variants of the same word (e.g., hesitated, hesitant), and close-synonym words (e.g., swift, fast, speedy).

[2] Some videos were rated by more than 7 participants during an initial phase of data collection in which we sought to discover how many ratings per video were necessary to establish consistency in the ratings. After this initial phase, we determined that only 7 ratings were necessary to achieve sufficient consistency, and we collected 7 ratings per video after that. For videos rated by more than 7 participants, we included the ratings only from the first 7 participants in this analysis.

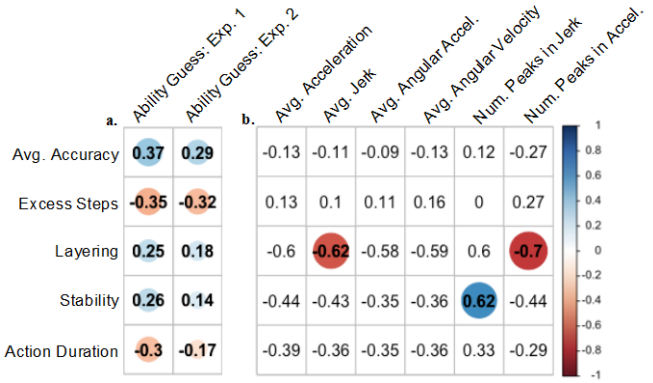

Figure 2: Correlation of children's building performance metrics with (a) adults' judgments of children's overall skill (Ability Guess) in Experiments 1 and 2 and (b) data from IMUs embedded in the blocks. Cells contain *r* values; colored circles indicate significant correlations.

Given that adults' judgments of overall construction ability reflected children's performance, could adults' observations of specific aspects of children's motions also predict their building performance? Examining the ratings for all 20 motion quality descriptors, we discovered that many of the descriptor ratings were strongly colinear. To identify a smaller number of relevant dimensions of motion quality and thereby reduce the number of predictors in our models, we conducted a cluster analysis using the *mclust* package in R (Scrucca et al., 2016). The optimal model was a Gaussian finite mixture model with four diagonal, equally shaped clusters of varying volume. We labeled these clusters based on the overall motion quality described by the descriptors in the cluster (see Figure 3): these were Fluidity, Confidence, Hesitancy, and Roughness. Then we took the average of the normalized ratings for all descriptors in each cluster to get a single value for that cluster (e.g., the average of the ratings for Fast, Quick and Confident for the Confidence cluster) per adult participant and video.

We then examined whether the average ratings for each descriptor cluster predicted the children's building performance. We used linear mixed-effects models with each performance metric as the dependent variable, the average ratings for each cluster as main (fixed) effects, and child participant as a random slope. Lower Hesitancy ratings predicted greater Accuracy (β = -0.06, SE = 0.03, $p < 0.05$), whereas higher Hesitancy ratings predicted more Excess Steps (β = 2.07, SE = 0.93, $p < 0.05$). Higher Fluidity ratings predicted greater Stability (β = 0.008, SE = 0.004, $p < 0.05$). The overall findings indicate that children who were judged to be faster and more fluid builders actually made fewer mistakes and created more stable structures while building.

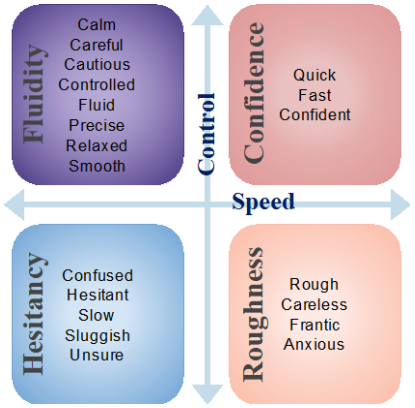

Figure 3: The four clusters of motion description words identified by our cluster analysis. The clusters varied on the two dimensions of control and speed.

Do motion quality judgments predict skill ratings? We next examined the relationship between adult raters' assessment of specific aspects of motion quality and their assessment of children's overall skill at block building. If adults used their observations of specific aspects of children's motions to judge children's overall ability, then adults' motion descriptor ratings should predict the overall skill ratings. We conducted a linear regression with Ability Guess as the dependent variable and the average ratings for each of the four clusters as predictors.[3] Higher Ability Guess ratings were predicted by higher average ratings for the Fluidity cluster (β = 0.53, SE = 0.14, $p < 0.01$) and Confidence cluster (β = 0.37, SE = 0.11, $p < 0.01$). Ability Guess was not predicted by ratings for the Hesitancy cluster (β = -0.13, SE = 0.12, $p > 0.1$) or Roughness cluster (β = -0.18, SE = 0.12, $p > 0.1$). This indicates that there is a relationship between adults' detailed judgments of motion quality and their estimates of children's overall skill. In particular, adults who judged children as having quicker and smoother movements also considered them better builders overall than children judged to have slower or rougher movements.

How good are IMU (motion) data in predicting children's block-building skill? We analyzed the relationship between children's movements during building, measured by the sensors (IMUs) embedded in the blocks, and children's building performance in order to determine whether objective measurements of movement are associated with building performance in any way. Since adult judgments predicted children's performance, would IMU sensor measurements also predict children's performance in the block-building task?

For this analysis, we used the IMU data from the entire duration of the task (rather than just the 25 seconds shown to adults for rating purposes), since the children's performance

[3] We attempted to conduct a linear mixed-effects model for this analysis, but the inclusion of random slope for participant resulted in a model with singular fit.

data are based on measurements across the entire build.[4] For each of the six types of motion data measured by the IMUs (average acceleration, average jerk, average angular velocity, average angular acceleration, number of peaks in acceleration per second, and number of peaks in jerk per second), we calculated the average across all of the blocks used in the build (e.g., the average acceleration for all 6 or 8 blocks, depending on model size). We then conducted individual linear regression analyses for each of the five performance metrics (Accuracy, Excess Steps, Action Duration, Layering, and Stability). In each model, the performance metric was the dependent variable, and the six IMU metrics were the predictors. We found that none of the IMU metrics from the full task predicted children's performance on the task (all *p*s > 0.1). This was especially surprising for Action Duration, which we expected to be predicted by at least average acceleration. A Pearson's correlation with Holm correction for multiple comparison showed few significant relationships between the IMU metrics and the building performance measures (Figure 2). Layering was significantly correlated with average jerk and the number of peaks in acceleration, and Stability was significantly correlated with the number of peaks in jerk. However, it is not clear how to interpret these sparse significant correlations, or how these specific IMU properties relate to Layering and Stability.

Are adult judges perceiving movement information also conveyed by IMU metrics? Given the relationships between adults' judgments about both movement quality and builder skill and children's performance on the task, we also investigated whether there was a relationship between adult raters' judgments and children's movements during the task as measured by the IMUs. Are human judges and machine measurements picking up on any of the same aspects of motion?

For this analysis, we used just the data recorded during the 25 seconds shown in the video clips to ensure that we only included motion data from block movements that the adult raters actually saw (for purposes of their rating). We conducted our analysis using linear mixed-effects models with Ability Guess as the dependent variable and the six averaged IMU metrics as predictors, with participant (adult rater) as a random effect. We found no relationship between the Ability Guess given by adults and any of the IMU metrics (all *p*s > 0.1).

The motion data from the IMUs did predict some of the motion descriptor judgments, however. We conducted linear regression models with the averaged ratings for the motion word clusters as the dependent variable, and each of the averaged IMU metrics as predictors. Higher adult ratings for descriptors in the Confidence cluster were significantly predicted by faster average acceleration (β = 136.71, *SE* = 36.99, p < 0.01), lower average angular velocity (β = -0.62, *SE* = 0.25, p = 0.01), lower average jerk (β = -1.32, *SE* = 0.56, p = 0.02), and higher number of peaks in acceleration (β = 1.04, *SE* = 0.52, p < 0.05). By contrast, higher adult ratings for the Hesitancy cluster were predicted by slower average acceleration (β = -102.37, *SE* = 34.65, p < 0.01) and higher average angular velocity (β = 0.57, *SE* = 0.23, p = 0.01). Together, these findings suggest that adults' judgments of Confidence and Hesitancy could be linked to specific aspects of motion – perhaps speed – that were picked up by the IMUs. However, ratings for the Fluidity and Roughness clusters were not predicted by any IMU metrics, showing the limitations of the IMU data for predicting adult ratings of children's motions.

In sum, both adults' overall estimates of children's skill and adults' detailed movement ratings predicted aspects of children's building performance, as we found relationships between Ability Guess and the motion description word judgments as well as our five performance metrics – Accuracy, Excess Steps, Action Duration, Layering, and Stability. This is especially interesting given that some measures, like Accuracy, were not detectable from the video clips, since our adult participants did not know what the target block model looked like. We also found a relationship between adults' detailed judgments of builder movements, as measured by their ratings of motion description words, and their judgments of a builder's overall skill. Finally, motion data on children's movement of the blocks while building, as measured by IMUs inserted in the blocks, predicted some aspects of adults' ratings of the quality of children's motions, but did not have a strong relationship with either the overall skill estimates given by adults or children's actual performance on the task.

Experiment 2

Given the relationship between adults' motion quality judgments and their estimate of children's construction ability, we wondered whether adults' judgments of the children's overall ability were influenced by their detailed judgments about the qualities of the children's movements. We conducted a second, follow-up experiment to explore this possibility.

Methods

Participants Participants were 28 undergraduate students at Johns Hopkins University. A university ethical review board approved all study procedures, and participants provided informed consent and received course credit for their participation.

Design, Stimuli, and Procedure The set-up of Experiment 2 was the same as Experiment 1, except that participants were only asked about the child's age, construction ability, and gender, and were not asked to rate any motion words.

[4] To determine if the motion data from the 25-second sample was representative of the full task, we looked at the average jerk and average acceleration of each block and found that the motion data from the 25-second sample and full task were highly correlated, with Pearson's *r* ranging from 0.64-0.95 with an average value of 0.85.

Participants viewed and rated the same 24 videos as in Experiment 1.

Results

We compared the average Ability Guess score for each construction video from Experiment 1 to the average Ability Guess score for each video in Experiment 2, where adults only provided overall skill estimates. The Ability Guess scores in the two experiments were highly correlated (Pearson's $r = 0.71$, $p < 0.01$). We also analyzed the overall skill assessments in each experiment using a linear mixed-effects model with Ability Guess as the dependent variable, experiment as a main effect (treatment coded), and participant (adult rater) as a random slope. There was no significant difference in the Ability Guess ratings for the two experiments ($\beta = -0.14$, $SE = 0.14$, $p = 0.34$). Judgments of motion quality thus appear to predict, but not uniquely determine, judgments of children's overall skill.

We looked at whether adult raters' Ability Guess scores also predicted children's actual performance in this experiment. We conducted linear mixed-effects models with each of the five performance metrics as the dependent variable, Ability Guess and block model as main (fixed) effects, and child participant as a random slope. Adults' Ability Guess scores significantly predicted children's Accuracy ($\beta = 0.01$, $SE = 0.006$, $p < 0.05$), Excess Steps ($\beta = -0.56$, $SE = 0.23$, $p < 0.05$), Layering ($\beta = 0.02$, $SE = 0.01$, $p < 0.05$), and Stability ($\beta = 0.004$, $SE = 0.001$, $p < 0.01$). A Pearson's correlation with Holm correction for multiple comparisons also showed significant correlations between Ability Guess and all five performance metrics (see Figure 2), as in Experiment 1.

The results from Experiments 1 and 2 show that detailed judgments about motion quality were related to, but did not determine, estimates of overall building skill, since adults gave very similar estimates of construction ability whether they judged motion quality or not. In both experiments, the building performance metrics were correlated with adults' construction ability ratings.

Discussion

Our findings show that naïve human judges can accurately evaluate children's expertise in a block-building task simply by watching the children build, without knowledge of the target structure. Our findings also indicate that these judges outperformed information from embedded sensors (IMUs) in gauging children's skill on this commonplace yet complex activity. While both adult raters and the IMUs in the blocks picked up on some aspects of children's motions, no aspect of movement measured by the IMUs was a reliable predictor for children's performance on the building task in our regression analyses. In contrast, adults' judgments about children's motion quality and overall construction ability predicted children's actual building skill, even when adults were not directed to pay attention to motion. This indicates that human judges were picking up on qualities of motion that may not be easily measured by our IMUs.

Our finding that information from IMUs tracking children's motions while building did not predict children's ability as well as adults making judgments based on simply watching may not be surprising to some. After all, some might argue that the IMU data do not reflect the higher-order relationships that characterize 'expertise.' However, this is the point: expertise is not obviously the product of all and only properties of motion. But it is important to note that our results fit with findings from research on crowd-sourcing assessment of robotic surgery skills, which has shown that laypeople, with no medical or surgical expertise, can not only reliably distinguish novices from experienced surgeons as well as expert surgeons can, but can also make more fine-grained distinctions in skill levels than is possible from the measurements of a robotic surgery simulator (Ershad et al., 2018). Here, we have documented analogous findings using a complex task (block construction) as it is executed by young children – findings that are important for any consideration of how to evaluate children's expertise in this (or other) perceptual-motor domains.

Our findings not only highlight the ability of naïve adults to make overall judgements about a child's skill, but also suggest that there is still considerable progress to be made in automating skill assessment for everyday yet complex tasks such as block construction. In our study, automated measures were limited to those derived from embedded sensors in each block the child moved. We can imagine next steps, including refining the information gathered by the sensors, adding information derived from the videotaped sequences viewed by our adult judges, and using machine learning to train such a system to make more accurate predictions about children's actual building performance. However, such programs may still be in their infancy and early results expose the complexity of the problem at hand (e.g., Jones et al., 2021). It still remains the case that measuring children's spatial abilities using standardized tests or everyday tasks is arduous and time consuming. Being able to assess spatial abilities in reliable and scalable ways, using commonplace activities, could be a valuable tool for identifying children at risk for difficulties in STEM courses or other domains, and provide opportunities for intervention and assistance in improving spatial skills and the kinds of thinking and reasoning that depend on them. The findings from the current study and those on robotic surgery assessment show that there is still a need to determine what exactly humans perceive when they make judgments about a person's skill simply by watching them execute complex tasks such as the block construction task that we have studied here. What is clear is that whatever human perceivers extract in this process, it allows them to easily and rapidly assess skill and expertise in ways that simple machine measurements, such as those we have examined here, currently cannot.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Chen, C., White, L., Kowalewski, T., Aggarwal, R., Lintott, C., Comstock, B., Kuksenok, K., Aragon, C., Holst, D., & Lendvay, T. (2014). Crowd-Sourced Assessment of Technical Skills: A novel method to evaluate surgical performance. *Journal of Surgical Research*, *187*, 65–71.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Landau, B., & Shelton, A. L. (2018). Constraints and Development in Children's Block Construction. *CogSci 2018 Proceedings*, 244–249.

Ershad, M., Koesters, Z., Rege, R., & Majewicz, A. (2016). Meaningful Assessment of Surgical Expertise: Semantic Labeling with Data and Crowds. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 508–515. https://doi.org/10.1007/978-3-319-46720-7

Ershad, M., Rege, R., & Fey, A. M. (2018). Meaningful Assessment of Robotic Surgical Style using the Wisdom of Crowds. *International Journal of Computer Assisted Radiology and Surgery*, *13*(7), 1037–1048. https://doi.org/10.1007/s11548-018-1738-2

Jones, J. D., Cortesa, C., Shelton, A., Landau, B., Khudanpur, S., & Hager, G. D. (2021). Fine-Grained Activity Recognition for Assembly Videos. *IEEE Robotics and Automation Letters*, *6*(2), 3728–3735. https://doi.org/10.1109/LRA.2021.3064149

Kell, H. J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2013). Creativity and Technical Innovation: Spatial Ability's Unique Role. *Psychological Science*, *24*(9), 1831–1836. https://doi.org/10.1177/0956797613478615

Kowalewski, T. M., Comstock, B., Sweet, R., Schaffhausen, C., Menhadji, A., Averch, T., Box, G., Brand, T., Ferrandino, M., Kaouk, J., Knudsen, B., Landman, J., Lee, B., Schwartz, B. F., McDougall, E., & Lendvay, T. S. (2016). Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills Tasks. *Journal of Urology*, *195*(6), 1859–1865. https://doi.org/10.1016/j.juro.2016.01.005

Landau, B., Davis, E., Cortesa, C. S., Wang, Z., Jones, J. D., & Shelton, A. L. (2022). Young children's copying of block constructions: Remarkable constraints in a highly complex task. PsyArXiv: https://doi.org/10.31234/osf.io/tjb6f

Malpani, A., Vedula, S. S., Chen, C. C. G., & Hager, G. D. (2015). A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *International Journal of Computer Assisted Radiology and Surgery*, *10*(9), 1435–1447. https://doi.org/10.1007/s11548-015-1238-6

Ossmy, O., Kaplan, B. E., Han, D., Xu, M., Bianco, C., Mukamel, R., & Adolph, K. E. (2022). Real-time processes in the development of action planning. *Current Biology*, *32*, 190–199. https://doi.org/10.1016/j.cub.2021.11.018

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, *8*(1), 289–317. https://doi.org/10.32614/rj-2016-021

Shelton, A. L., Davis, E. E., Cortesa, C. S., Jones, J. D., Hager, G. D., & Khudanpur, S. (2022). Characterizing the Details of Spatial Construction: Cognitive Constraints and Variability. *Cognitive Science*, *46*(1), e13081. https://doi.org/10.1111/cogs.13081

Sorby, S. (1999). Developing 3-D Spatial Visualization Skills. *Engineering Design Graphics Journal*, *63*(2), 21–32.

Stannard, L., Wolfgang, C. H., Jones, I., & Phelps, P. (2001). A Longitudinal Study of the Predictive Relations Among Construction Play and Mathematical Achievement. *Early Child Development and Care*, *167*(1), 115–125. https://doi.org/10.1080/0300443011670110

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., Newcombe, N. S., Flipowicz, A. T., & Chang, A. (2014). Deconstructing Building Blocks: Preschoolers' Spatial Assembly Performance Relates to Early Mathematics Skills. *Child Development*, *85*(3), 1062–1076. https://doi.org/10.1111/cdev.12165.Deconstructing

White, L. W., Kowalewski, T. M., Dockter, R. L., Comstock, B., Hannaford, B., & Lendvay, T. S. (2015). Crowd-Sourced Assessment of Technical Skill: A Valid Method for Discriminating Basic Robotic Surgery Skills. *Journal of Endourology*, *29*(11), 1295–1301. https://doi.org/10.1089/end.2015.0191