Looking represents choosing in toddlers: Exploring the equivalence between multimodal measures in forced - choice tasks

Hagihara, Hiromichi; Ienaga, Naoto; Terayama, Kei; Moriguchi, Yusuke; Sakagami, Masa-Aki

Hagihara, Hiromichi ...[et al]. Looking represents choosing in toddlers: Exploring the equivalence between multimodal measures in forced - choice tasks. Infancy 2021, 26(1): 148-167

Received: 27 April 2020 | Accepted: 29 October 2020

DOI: 10.1111/infa.12377

RESEARCH ARTICLE

INFANCY
THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES
WILEY

Looking represents choosing in toddlers: Exploring the equivalence between multimodal measures in forced-choice tasks

Hiromichi Hagihara[1,2] | **Naoto Ienaga**[3] | **Kei Terayama**[4,5,6,7] | **Yusuke Moriguchi**[8] | **Masa-aki Sakagami**[1]

[1]Graduate School of Human and Environmental Studies, Kyoto University, Kyoto, Japan

[2]Japan Society for the Promotion of Science, Tokyo, Japan

[3]Graduate School of Science and Technology, Keio University, Yokohama, Japan

[4]Graduate School of Medical Life Science, Yokohama City University, Yokohama, Japan

[5]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

[6]Medical Sciences Innovation Hub Program, RIKEN, Cluster for Science, Technology and Innovation Hub, Yokohama, Japan

[7]Graduate School of Medicine, Kyoto University, Kyoto, Japan

[8]Graduate School of Letters, Kyoto University, Kyoto, Japan

Correspondence
Hiromichi Hagihara, Graduate School of Human and Environmental Studies, Kyoto University, Yoshida-nihonmatsu-cho, Sakyo-ku, Kyoto 606-8501, Japan.
Emails: hiromichi.h@gmail.com or hagihara.hiromichi.75a@kyoto-u.jp

Funding information
This research was supported by JSPS KAKENHI Grant Numbers JP18J21948,

Abstract

In the two-alternative forced-choice (2AFC) paradigm, manual responses such as pointing have been widely used as measures to estimate cognitive abilities. While pointing measurements can be easily collected, coded, analyzed, and interpreted, absent responses are often observed particularly when adopting these measures for toddler studies, which leads to an increase of missing data. Although looking responses such as preferential looking can be available as alternative measures in such cases, it is unknown how well looking measurements can be interpreted as equivalent to manual ones. This study aimed to answer this question by investigating how accurately pointing responses (i.e., left or right) could be predicted from concurrent preferential looking. Using pre-existing videos of toddlers aged 18–23 months engaged in an intermodal word comprehension task, we developed models predicting manual from looking responses. Results showed substantial prediction accuracy for both the Simple Majority Vote and Machine Learning-Based classifiers, which indicates that looking responses would be reasonable alternative measures of manual ones. However, the further exploratory analysis revealed that when applying the created models for data of toddlers who did not produce clear pointing responses,

JP20H05002, and Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research.

the estimation agreement of missing pointing between the models and the human coders slightly dropped. This indicates that looking responses without pointing were qualitatively different from those with pointing. Bridging two measurements in forced-choice tasks would help researchers avoid wasting collected data due to the absence of manual responses and interpret results from different modalities comprehensively.

KEYWORDS

language development, machine learning, pointing, preferential looking, two-alternative forced-choice task

1 | INTRODUCTION

The two-alternative forced-choice (2AFC) paradigm has been leveraged for a long time in various researches on cognitive development such as numeric skills, false-belief understanding, and prosocial characters (Fantz, 1964; Hamlin et al., 2010, Onishi & Baillargeon, 2005; Southgate, Senju, & Chibra, 2007; Starkey et al., 1983; Wagner & Johnson, 2011). As measures to estimate cognitive abilities, manual responses such as pointing, touching, or reaching have been widely utilized.

Especially in the field of language development, well-known examples of the 2AFC tasks using children's arm responses are the Forced-Choice Pointing (FCP) paradigm (Fernandes et al., 2006; Fisher, 1996; Maguire et al., 2008; Noble et al., 2011) or the Computerized Comprehension Task (Friend & Keplinger, 2003, 2008). In these methods, children watch two juxtaposed pictures or video clips and are asked to choose one of them that matches auditory instruction by manual responses. These methods have various advantages (Noble et al., 2011). First, methods using manual measures require no expensive specialized equipment, so it can be administered easily. Second, such measures can be easily coded even when manual (not automated) coding is adopted since children produce an overt, volitional response. Third, this method provides direct, unambiguous, and less noisy indices that are analyzed and interpreted easily. This characteristic of indices is also advantageous since there is less room for analytical arbitrariness such as conducting statistical analysis in haphazard manners or summarizing data in a self-serving manner. Fourth, manual measures are applicable to a broader age group of children and adults, hence long-term developmental differences can be investigated.

Methods using manual responses are generally suitable for children older than two years (Ambridge & Rowland, 2013), yet some studies adopted this method to toddlers from around 18 months (Friend & Keplinger, 2003, 2008; Gurteen et al., 2011; Hagihara & Sakagami, 2020). However, since methods using manual measures require a volitional response of arm movement, it is often uncertain how to address absent responses if observed, particularly when applying these methods to toddlers (Hendrickson et al., 2015). For example, Hendrickson et al. (2017) reported that, in a familiar word comprehension task, absent touching responses were seen for roughly one-third of trials in 16-month-olds and 10% in 22-month-olds. Hagihara and Sakagami (2020) also reported that from 52 participants aged 19–35 months, 9 toddlers (17%) were excluded due to the difficulty in coding pointing responses. As long as strictly adopting manual measures, absent responses are treated as missing values. However, are there any possibilities to avoid wasting the collected data?

In such cases, looking responses such as preferential looking can be available as alternative measures, if they are extractable and codable (e.g., from video recordings). In fact, there are many infant studies that use looking responses as measures of 2AFC tasks, such as the Intermodal Preferential Looking (IPL) paradigm (Bailey & Plunkett, 2002; Golinkoff et al., 1987, 2013; Yuan & Fisher, 2009) or the Looking-While-Listening paradigm (Fernald et al., 1998, 2008). For example, preferential looking paradigm leverages young children's tendency to look significantly longer at a stimulus that matches linguistic input than a distracter presented side-by-side (Tafreshi et al., 2014), and this method has been used to verify whether children can identify the correct referent of novel words (Chan et al., 2010; Gurteen et al., 2011; Horváth et al., 2015) or familiar words (Durrant et al., 2015; Mani & Plunkett, 2011; Valleau et al., 2018). These methods using looking responses are advantageous as they are applicable even for infants under 18 months (Imai et al., 2015; Mani & Plunkett, 2010; Smith & Yu, 2008) since they do not need children's volitional manual responses but only spontaneous looking ones. However, if looking responses are used as a dependent variable instead of manual ones to reduce the exclusion rate or missing data, how well can we treat the results of looking measurements as equivalent to those of manual ones?

Few studies have investigated children's looking and concurrent manual responses within the same 2AFC task because most infant and toddler studies used either of these measures selectively (Hendrickson et al., 2015). Moreover, some researchers remain skeptical about looking time as an appropriate index reflecting higher-order cognitive abilities (Haith, 1998) and it has often been reported that there were dissociations of results between modalities in research not only on language but also on other cognitive development (Abbot-Smith et al., 2017; Ahmed & Ruffman, 1998; Charles & Rivera, 2009; Gurteen et al., 2011; Ruffman et al., 2001; Winters et al., 2015). Hence, it is unknown to what extent looking and manual measurements can be interpreted analogously.

To our knowledge, Hendrickson and colleagues are the only researchers who conducted the FCP and the IPL paradigms simultaneously to toddlers (Hendrickson & Friend, 2013; Hendrickson et al., 2015, 2017). In Hendrickson and Friend (2013), 16- to 18-month-olds participated in the familiar word comprehension task. Toddlers' looking and manual responses were recorded via HD video cameras and were coded manually. Looking responses were categorized frame-by-frame into three (i.e., left fixation, right fixation, or away look), whereas manual responses were categorized into three for each trial (i.e., target touch, distractor touch, no touch). The results showed that toddlers looked significantly longer at the stimulus to which they subsequently reached regardless of stimuli type (i.e., target or distractor). Furthermore, on trials where reaching responses were not observed, toddlers looked at the target stimulus significantly longer than chance. Referring to recent connectionist studies (Munakata, 2001; Munakata & McClelland, 2003), Hendrickson and colleagues interpreted these findings as looking and manual responses reflect different levels of understanding for the word-referent association, that is, preferential looking can be observed even when representations of words are fragile, whereas reaching can be demonstrated for only robust representations. Their subsequent studies enhanced this view and further explored the way to detect children's knowledge status on a certain word by leveraging both looking and manual measures (Hendrickson et al., 2015, 2017). However, since they have mostly focused on the different interpretability between two modalities, the question of how well each measure can be interpreted as equivalent is still unexplored.

As the first step to directly address this question, this study investigated how accurately toddlers' volitional pointing (i.e., left or right) could be predicted from preferential looking. If pointing were accurately predictable from preferential looking, then it can be claimed that these two measurements are related to each other and looking responses are, to some extent, reasonably used as alternative measures of manual ones. If the prediction accuracy were low, then this indicates that these two indices reflect different cognitive processes or have different robustness to noise irrelevant to choice. We utilized pre-existing video data where both looking and pointing measures could be coded for this study (Hagihara et al., 2020). In this data, 18- to 23-month-old toddlers participated in a 2AFC task evaluating whether the meanings of object words were affected by object-specific actions. Here, we particularly focused on exploring the temporal features of frame-by-frame coded preferential looking that would predict binary pointing responses most adequately. This study consisted of three phases (see Figure 1 for the schematic flow). In Phase 1, we created two types of models for the prediction of pointing from preferential looking: the Simple Majority Vote (SMV) and the Machine Learning-Based (MLB) models. In the former model, the proportion of total looks to juxtaposed stimuli (left or right) for each trial was calculated while changing the target time window and the dominant side was used as a prediction index. In the latter model, we adopted the decision-tree-based algorithm LightGBM (Ke et al., 2017). We chose this algorithm for several reasons such as the fact that it is known to be a state-of-the-art option for a relatively small number of input variables. We used this machine learning method because it was expected that certain particular time ranges and/or their combinations would yield higher prediction accuracy than would merely use the proportion of looks to either stimulus. In Phase 2, we conducted the validation test of the created models by putting another dataset that was not used in Phase 1 into each model. If it turned out that pointing and preferential looking were closely related, are features of looking with clear pointing the same as those without pointing? In Phase 3, we exploratorily applied the created models to data of toddlers who did not produce clear pointing to preliminarily investigate this question. Since there were no absolute correct answers (i.e., absent pointing responses), we adopted manual estimations of pointing responses from looking behavior as a pseudo-correct index in order to evaluate the applicability of the created models. If the agreement of pointing estimations between human coders and the models were equivalent to the prediction accuracy calculated in Phases 1 and 2, then it would be indicated that looking responses were similar regardless of executing manual responses or not. If the agreement dropped, then it would be suggested that looking responses without manual ones were qualitatively different from those with manual ones.

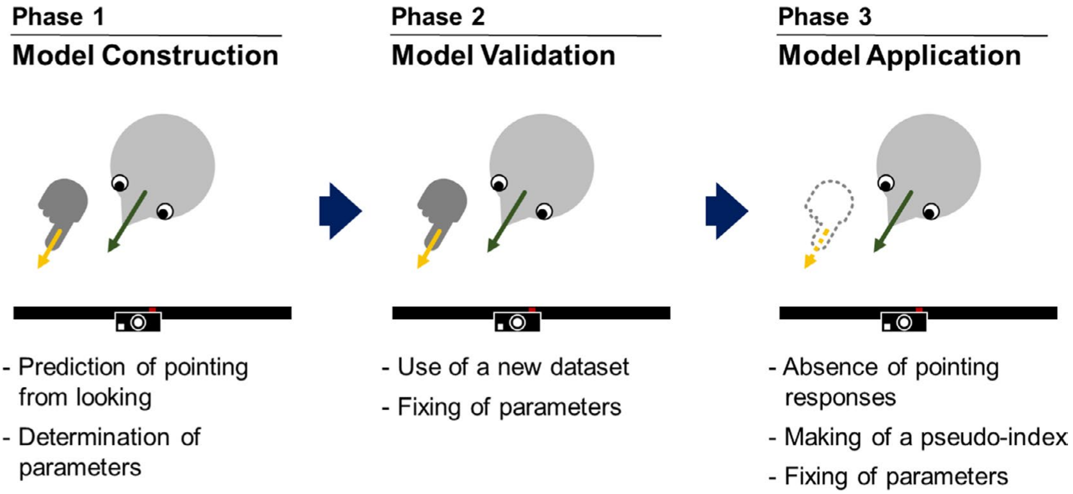


FIGURE 1 Schematic view of the flow of the present study.

2 | PHASE 1: MODEL CONSTRUCTION

2.1 | Methods

2.1.1 | Video recordings

For this study, we utilized pre-existing videos in which toddlers' familiar word comprehension was investigated using the FCP paradigm (Hagihara et al., 2020). Inter-rater reliability of pointing responses (left or right) was confirmed with 97.8% of agreement (kappa = 0.96). Of the participants who showed explicit pointing responses for 75% or more in all 16 experiment trials, 36 toddlers were selected and allocated to model construction ($n = 24$) and validation ($n = 12$) in this study so that age and gender were not biased. For Phase 1, we used videos of 369 trials with 24 monolingual Japanese toddlers aged 18–23 months (12 girls; $M = 21.1$, $SD = 1.7$). Fifteen trials (8 toddlers) were excluded for lack of explicit pointing responses. Each remaining participant responded clearly in 12–16 trials ($M = 15.4$, $SD = 1.1$).

In the experimental task, toddlers sat on a small chair or the lap of a nursery school teacher and looked at the 21.5-in touch screen (490 × 243 mm) with a viewing distance of approximately 30 cm. Toddlers' preferential looking and pointing responses were recorded via a webcam at the center of the top of the screen (30 frames/s). Toddlers completed the forced-choice task modified from Hagihara and Sakagami (2020). This task aimed to investigate whether the meanings of object words were affected by object-specific actions. This task consisted of 16 trials in which four different conditions were included, which varied in terms of how much participants had to depend on object-specific actions to make judgments about referents of familiar object words. For example, in one condition, immediately after watching two juxtaposed videos—"putting on shoes" (the target stimulus) and "rubbing two baskets in front of one's chest" (the foil stimulus)—participants were prompted to choose one of them by pointing to answer "*Kutsu wa docchi?*" [Which are shoes?]. In another condition, they watched video stimuli located side-by-side—"rubbing shoes in front of one's chest" (the target stimulus) and "putting on two baskets as if they were shoes" (the foil stimulus)—then were asked to choose one that matched the question of which is shoes. Before performing tasks, participants were prompted to look at icons on the screen's center and each corner for later calculation of angle compensation; they also engaged in warm-up trials to understand the task rule. This research was conducted according to guidelines laid down in the Declaration of Helsinki, with written informed consent obtained from the parents of all participants before data collection. All procedures involving human subjects in this research were approved by the ethics committee for human and animal research of the Graduate School of Human and Environmental Studies at Kyoto University.

2.1.2 | Face/gaze detection and pre-processing

The video for each trial was cropped with a time window from when the question ended to 2,000 ms thereafter since, at a later time, looking behavior is no longer considered to be related to the stimulus (Delle Luche et al., 2015). We set the starting point of the potential time window not at the onset or offset of the target word but at the end of the question sentence because, in Japanese, a listener cannot determine if a sentence is a question unless they hear the sentence through to the end due to the grammatical difference in word order from English. For example, in the sentence "*Kutsu wa docchi?*" [Which are shoes?], "*kutsu*" refers to the target word "shoes," "*wa*" refers to a postpositional particle indicating that the preceding word is the subject, and "*docchi*" refers to the interrogative marker "which." Note that Japanese allows for a dropping of the verb "are" in the sentence and does not distinguish between singular and plural forms of nouns.

In line with recent vigorous studies on automatic gaze estimation using webcam-based data (Chouinard et al., 2019; Papoutsaki et al., 2016), we used an open-source library, OpenFace 2.2.0 (Baltrušaitis et al., 2018) for automatic face and gaze estimation. We adopted this toolkit because it provides rich information such as facial landmarks, gaze directions, and three-dimensional head positions and angles, because it is freely available for research purposes, and works in the local environments, which reduces ethical concerns. Although this toolkit requires only an RGB camera to estimate gaze direction using machine learning techniques, the precision is far less than eye trackers (Higuchi et al., 2018). However, facial detection is more robust than gaze detection. Hence, we used horizontal face angles as well as gaze angles as measurements indicating whether a participant looked right or left side of the screen. Additionally, although complete preferential looking coding requires annotation not only whether a child is looking left or right but also whether a child is looking at or away from the screen, we began with the simplified coding to distinguish whether a child was looking right or left relative to the center of the monitor.

After estimating face and gaze angles using OpenFace 2.2.0, we calculated angle compensation so that the indices were always zero when a participant looked at the center of the monitor regardless of changing their head position (Figure 2). First, we calculated x_{fc} which was the head position when a child looked at the center icon on the screen as follows:

$$x_{fc} = x_{cal} - z_{cal} \times \tan\left(\theta_{fcal}\right),$$

where x_{cal}, z_{cal}, and θ_{fcal} represent the average of 30 frames of each variable when a child looked at the center icon of the screen. The variable x stands for the horizontal distance between the webcam and participants' head along with the screen, z for the distance between the webcam and participants' head, and θ_f for the raw horizontal face angle, respectively. The compensated face angle $\theta'_f(n)$ in frame n was then calculated as follows:

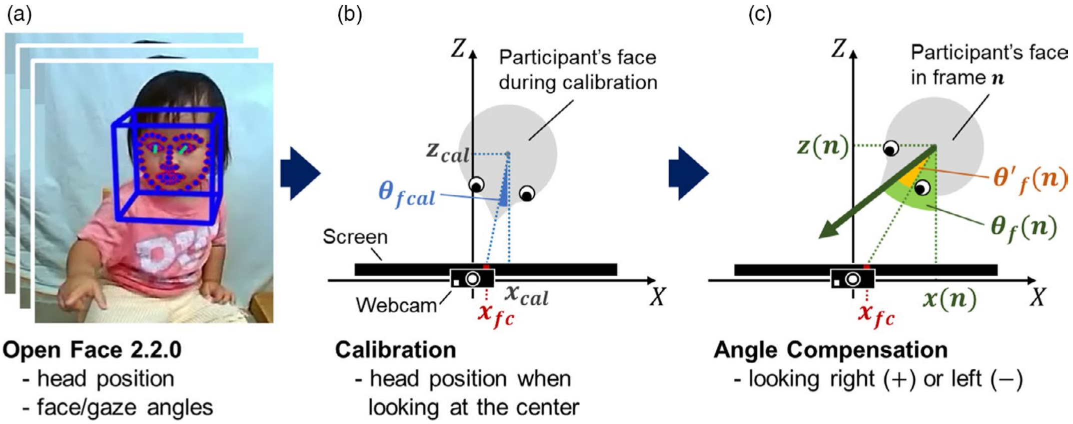


FIGURE 2 The pre-processing flow for estimating preferential looking.

Note. (a) Using video data recorded by a webcam located at the center of the top of the screen, head positions, and raw face/gaze angles were estimated by OpenFace 2.2.0. (b) The head position when a participant looked at the center icon of the screen was calculated in order to compensate face/gaze angles so that these indices were always zero when a participant looked at the center of the screen regardless of changing head position. (c) Compensated face/gaze angles were calculated so that the indices were positive when looking right relative to the center of the screen.

$$\theta'_f(n) = \theta_f(n) - arctan\left(\frac{x(n) - x_{fc}}{z(n)}\right).$$

Thus, when $\theta'_f(n)$ is greater than 0, it indicates that the toddler is looking right. Note that the expression calculating the compensated gaze angle $\theta'_g(n)$ needs slight changes because the plus and minus of its angle are inverted from the face angle in the use of OpenFace 2.2.0 as follows:

$$\theta'_g(n) = -\theta_g(n) - arctan\left(\frac{x(n) - x_{gc}}{z(n)}\right).$$

$$x_{gc} = x_{cal} + z_{cal} \times \tan(\theta_{gcal}).$$

As in the face direction, $\theta_g(n)$ stands for the raw horizontal gaze angle and θ_{gcal} for the averaged angle of looking at the center of the screen. For each frame, OpenFace 2.2.0 provides the "Confidence" value, which indicates how precisely a face can be detected, ranging from 0 to 1 (a higher value indicates successful face detection). When the face detection failed (i.e., the Confidence value was low), the temporal linear interpolation was conducted for estimated variables.

To confirm how reliable face and gaze angle estimations in OpenFace 2.2.0 were, we verified the agreement between the estimation and frame-by-frame manual coding of preferential looking (left or right). A trained coder manually annotated toddlers' preferential looking for approximately 25% of the data in Phase 1, in which four of each participant's trials were pseudo-randomly extracted (4 trials × 24 participants = 96 trials).

2.1.3 | Creating models for predicting pointing responses from preferential looking

Using time-series data of face and gaze horizontal angles as an input, we created two types of classifiers, which predict pointing responses from preferential looking. We regarded the pre-existing human annotation of pointing responses as the correct answer.

Simple Majority Vote (SMV) model

In the IPL paradigm, researchers have conventionally compared the proportion of looks to the target stimulus to the proportion to the foil stimulus within a certain time window (Ambridge & Rowland, 2013). In line with this procedure, we converted the compensated face and gaze angles into a binary index (left or right), and regarded the dominant side of looks as a prediction for pointing. One major difference of the SMV model from the conventional approach was that the pointing prediction was calculated while changing the target time window in order to reflect the temporal features of looking responses in the optimization of the prediction accuracy of pointing responses (Figure 3). We calculated a variable $P_{i,j}$ modeled for this method, where $NR_{i,j}$ and $NL_{i,j}$ stand for the number of frames that a participant looked right and left side, respectively. The variable $P_{i,j}$ was computed while changing the starting time point i and the ending time point j of the target time window as follows:

$$P_{i,j} = 0.5 \times \frac{NR_{i,j} - NL_{i,j}}{NR_{i,j} + NL_{i,j}} + 0.5.$$

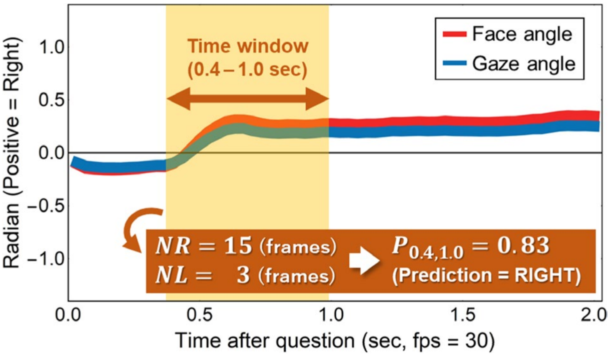


FIGURE 3 The schematic view of the SMV model.

Note. The number of frames where a participant looked right or left was calculated from the compensated face and gaze angles within a certain time window (from 0.4 to 1.0 s in this case). The dominant side was regarded as a prediction of pointing responses. The target time window was moved between 0.0 and 2.0 s after the completion of the question to optimize the prediction accuracy.

$P_{i,j}$ ranged between 0 and 1 (indicated a participant completely looked left or right, respectively) and when $P_{i,j}$ was greater than 0.5, it was predicted that the participant pointed right. If $P_{i,j}$ was equal to 0.5, we defaulted the prediction to the right side.

Machine Learning-Based (MLB) model

In addition to the SMV model, we used the decision-tree-based algorithm LightGBM (Ke et al., 2017) as a classifier since we expected that certain time ranges and/or their combinations would yield a higher prediction accuracy than simply using the proportion of looks to the target stimulus. We adopted this algorithm because it is known to be a state-of-the-art option for a relatively small number of input variables, it is one of the most popular methods in recent machine learning competitions, and the contribution of each input variable to the prediction, called "importance," can be easily quantified and visualized using LightGBM. First, we standardized the compensated face and gaze angles, respectively. We then determined the hyperparameters of the MLB model using the grid search optimization procedure, where all combinations of parameters were used to attempt to explore the optimal values of each parameter. LightGBM has several hyperparameters that are related to prediction accuracy such as the number of trees and the learning rate. For example, the number of trees refers to how many decision trees are combined; these are the sub-elements that make up the model. The learning rate refers to how much information about the learning error at a certain step is propagated to the next step. The prediction accuracy of each parameter was calculated with 10-fold cross-validation to avoid overfitting to the specific data. In 10-fold cross-validation, the given data were split into training (90%) and validation (10%), repeated training 10 times for each partition, and evaluated the model performance by averaging the obtained results. We adopted the parameters that produced the highest accuracy and the final prediction model was created using all the data. The prediction variable produced by LightGBM was continuous ranged from 0 to 1, where a value >0.5 indicated the prediction that a participant pointed right. As in the SMV model, a value that equaled to 0.5 was defaulted to pointing right.

2.1.4 | Evaluation of created models

To evaluate the prediction accuracy of the created SMV and MLB models, the predicted side of pointing responses (right or left) produced by each model was compared with previously obtained manual coding for all trials. As indices, we used the area under the curve (AUC) of receiver operating characteristic (ROC) curve, accuracy rate, and kappa coefficient. AUC is a measurement reflecting how much the model has the discriminative ability, where 1.0 indicates the model can perfectly predict the correct results whereas 0.5 indicates chance level. According to the rule of thumb (Akobeng, 2007; Swets, 1988), an AUC greater than 0.9 has high, 0.7–0.9 has moderate, and 0.5–0.7 has low accuracy. Kappa coefficient is also a well-used measurement, where the higher value indicates higher accuracy. Based on the criterion by Landis and Koch (1977), kappa of 0.81–1.00 has almost perfect, 0.61–0.80 has substantial, and 0.41–0.60 has moderate strength of accuracy.

2.2 | Results

2.2.1 | The agreement of preferential looking between the estimation in OpenFace 2.2.0 and manual coding

For 96 trials randomly extracted from Phase 1 data, preferential looking estimated using OpenFace 2.2.0 demonstrated almost perfect reliability with frame-by-frame human coding both in the face (kappa = 0.85, 92.7% of agreement) and gaze (kappa = 0.82, 91.2% of agreement) directions. Therefore, we continued to use the automated face and gaze angle estimations produced by OpenFace 2.2.0.

2.2.2 | To what extent preferential-looking reflected pointing responses?

The indices reflecting prediction accuracy for the best models of the SMV and the MLB were shown in Table 1. In the SMV model, we explored the optimal time window that predicted toddlers' pointing

TABLE 1 Evaluation of the created models of predicting pointing responses with highest accuracy and their robustness

	Phase 1. Model construction (369 trials)		Phase 2. Model validation (183 trials)		Phase 3. Model application to no-pointing trials (176 trials)	
	SMV model	MLB model	SMV model	MLB model	SMV model	MLB model
Face angles as input						
AUC	0.945	0.941	0.934	0.938	0.908	0.877
Accuracy rate	0.892	0.897	0.858	0.869	0.818	0.795
kappa	0.781	0.793	0.703	0.727	0.637	0.590
Gaze angles as input						
AUC	0.941	0.942	0.933	0.952	0.897	0.864
Accuracy rate	0.892	0.892	0.858	0.869	0.818	0.778
kappa	0.781	0.781	0.703	0.726	0.637	0.557

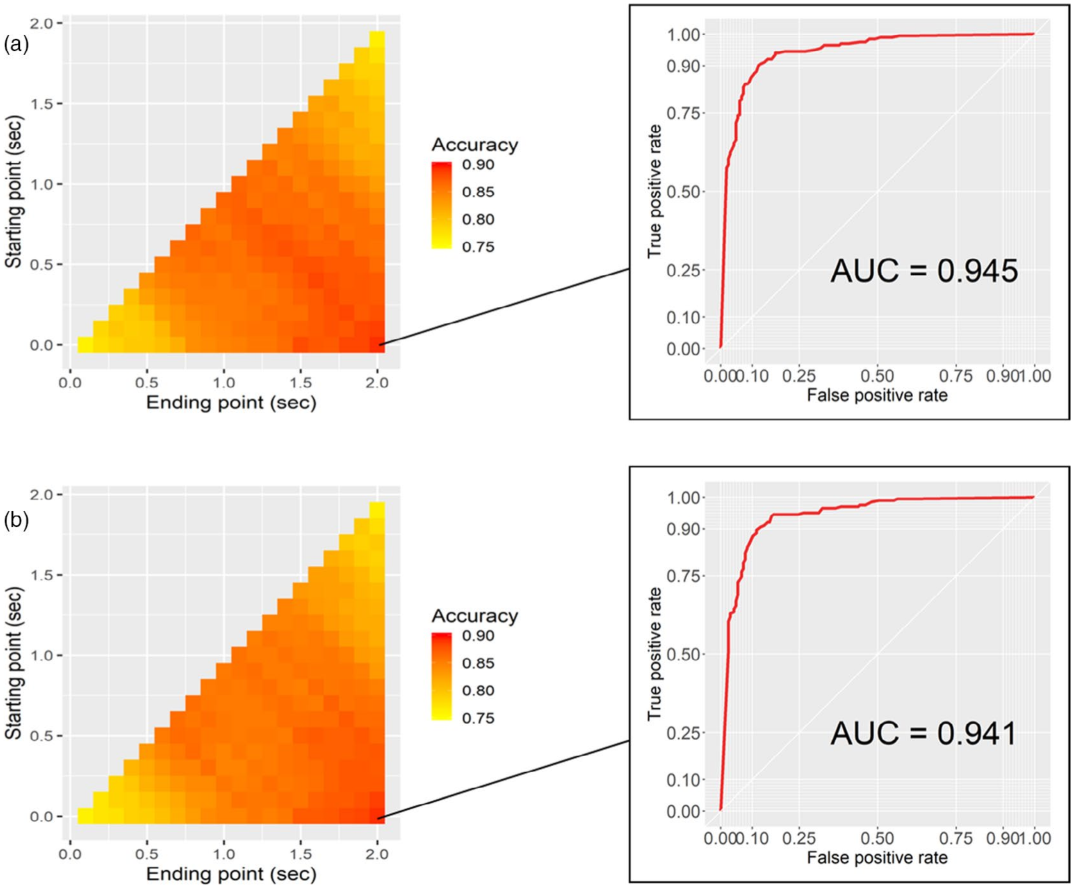
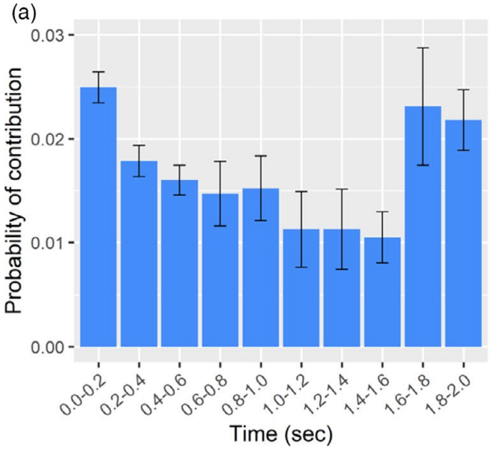


FIGURE 4 Prediction accuracy of the SMV model in each time window and ROC curves of models in the best time window.

Note. For the heatmaps, the y-axis indicates the starting point of the target time window, whereas the x-axis the ending time point. The redder the color, the more accurate the prediction was. (a) The heatmap of the accuracy rate in models using face angle trajectories. The highest accuracy was seen when the time window was set from 0.0 to 2.0 s. The ROC curve with this time window was shown on the right side. (b) The heatmap of the accuracy rate in models using gaze angle trajectories. The best time window was the same as in the model using face direction. The ROC curve with this time window was shown on the right side.

responses using all data in Phase 1. The best time window was found to be from 0.0 to 2.0 s immediately after the question ended in both models using face and gaze angle trajectories, while the narrow time windows including the very beginning or the end of all the potential time range showed lower prediction accuracy (Figure 4). The SMV model with the best time window demonstrated the highest accuracy in both models using face and gaze direction (89.2%), with high AUC and substantial kappa. The MLB model also showed high prediction rates in AUC, accuracy rate, and kappa coefficient, which were equivalent to the best SMV models. The temporal feature reflecting the prediction of pointing responses were visualized using "importance," the contribution probability of each input frame to the prediction (Figure 5). Approximately, among all 2.0 s of the potential time range, the first 1.0 s and the last 0.4 s were relatively critical to the classification for both face and gaze angle trajectories.



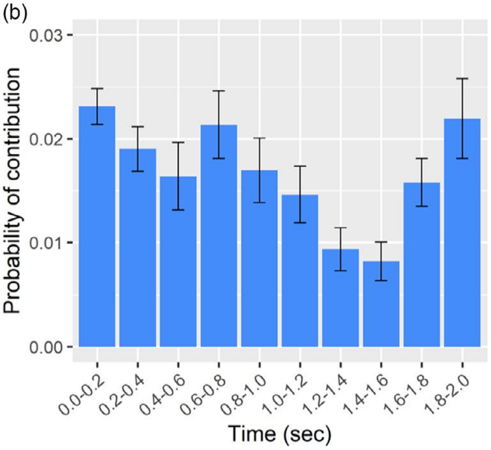


FIGURE 5 Contribution value of each time window to the prediction in the MLB model.
Note. The y-axis indicates the probability of contribution of each input to the classification prediction, called "importance," produced by LightGBM. The x-axis indicates the time after the completion of the question summarized every 0.2 seconds (i.e., 6 frames). The error bar indicates standard error. (a) “Importance” using face angles, (b) “Importance” using gaze angles.

3 | PHASE 2: MODEL VALIDATION

3.1 | Methods

3.1.1 | Video recordings

To evaluate the robustness of the created best models, another dataset that was not used in Phase 1 was utilized to conduct the model validation. As described previously, we used pre-existing webcam-based data (Hagihara et al., 2020). For Phase 2, we used videos of 183 trials with 12 monolingual Japanese toddlers aged 18–23 months (6 girls; *M* = 20.7, *SD* = 1.8). Nine trials (5 toddlers) were excluded for lack of explicit pointing responses. Each remaining participant responded clearly in 13–16 trials (*M* = 15.3, *SD* = 1.1). The experimental task was the same as in Phase 1.

3.1.2 | Face/gaze detection, pre-processing, and model validation

All video recordings were pre-processed as in Phase 1. Namely, we estimated face and gaze angle using OpenFace 2.2.0, identified the head position when a participant looked at the center of the screen, and compensated angles so that the values were always zero when a participant looked at the center of the screen regardless of changing their head position. To predict pointing responses, the pre-processed data were then put into the SMV and the MLB models, which showed the highest accuracy in Phase 1 with fixed parameters. We evaluated these models' prediction accuracy using AUC, accuracy rate, and kappa coefficient.

3.2 | Results

3.2.1 | How much robust the created models were?

Using a new dataset that was not used in Phase 1 as input, both the SMV and the MLB models still demonstrated high performance (Table 1). In all models, the prediction accuracy of pointing responses was around 86%. AUC showed high accuracy and kappa was substantial.

4 | PHASE 3: MODEL APPLICATION TO DATA WITH NO-POINTING RESPONSES

In this phase, we exploratorily applied the created models to data lacking overt pointing responses in order to preliminarily investigate whether features of preferential looking with clear pointing were equivalent to those without pointing. Indeed, there were no absolute correct answers; however, by calculating the agreement of pointing estimations from looking behavior between human coders and the created models, and comparing this agreement to the prediction accuracy obtained in Phases 1 and 2, we tried to see if there were qualitative differences between looking responses with and without manual ones. If the agreement in Phase 3 were equivalently high to the prediction accuracy in Phases 1 and 2, it would indicate that looking responses were similar to some extent regardless of pointing execution. However, if the agreement dropped, it would suggest that looking responses without arm movement had some different qualitative features compared to those with clear arm movement.

4.1 | Methods

4.1.1 | Video recordings

For Phase 3, we used videos of 176 trials with 12 monolingual Japanese toddlers aged 18–22 months (6 girls; $M = 19.25$, $SD = 1.5$) from the same pre-existing data as in Phases 1 and 2 (Hagihara et al., 2020). This sample was extracted from the participants who lacked clear pointing responses for two thirds or more in all 16 trials so that age and gender were not biased. Another 16 trials with five participants were excluded because they showed explicit pointing behavior. The final data used for analysis included between 11 and 16 trials with each participant ($M = 14.7$, $SD = 1.8$). The experimental task was the same as in Phase 1.

Since there were no absolute correct answers in the videos in Phase 3 (i.e., absent pointing responses), we preliminarily adopted manual annotation for estimating toddlers' volitional choice from looking behavior as a pseudo-correct index. A trained and a naive rater independently evaluated which side the participant seemed to choose for all videos based on toddlers' preferential looking behavior. For discrepancies in raters' coding, a third rater annotated participants' choosing; the annotations with agreement by two of the three raters were used. To verify how accurately the human raters can estimate pointing responses based only on preferential looking without seeing the exact pointing, the raters also manually annotated approximately 25% of the data in Phase 1 while participants' pointing responses were masked in the video. Four of each participant's trials from data in Phase 1 were pseudo-randomly extracted (4 trials × 24 participants = 96 trials) and annotated as in Phase 3. For all manual annotations in Phases 1 and 3, raters also evaluated their degree of confidence in their estimation of pointing using a five-point Likert scale, where 5 indicated highest confidence. These confidence values were calculated by averaging their ratings.

4.1.2 | Face/gaze detection, pre-processing, and model application

All video recordings were pre-processed as in Phase 1 using OpenFace 2.2.0. The pre-processed data were then put into the SMV and the MLB models to predict pointing responses. Note that the parameters in these models were the same as those in Phase 2, which means that the models used in Phase 3 were identical to the best models constructed in Phase 1. The agreement of pointing predictions (left or right) between the created models and human rater was evaluated using AUC, accuracy rate, and kappa coefficient. In addition, we defined the confidence value of pointing predictions for each model. For the SMV and the MLB models, the final variable for the prediction ranged from 0 to 1, where 0 indicated left while 1 indicated right, the distance of the value to 0.5 (e.g., $|P_{i,j} - 0.5|$ for the SMV model) was treated as the confidence value. The correlation of confidence values between each model and manual annotation was analyzed using Spearman's rank correlation coefficient.

4.2 | Results

4.2.1 | Manual estimation of pointing responses based only on preferential looking

For 96 trials randomly extracted from Phase 1 data, inter-rater reliability was substantial (kappa = 0.77, 88.5% of agreement). Eleven of the trials were coded differently between raters, requiring the third rater to produce a majority opinion. The adopted manual annotation demonstrated almost perfect reliability of correct responses (kappa = 0.92, 95.8% of agreement). Therefore, it indicated that human raters could reliably estimate pointing responses based only on preferential looking behavior, not on the exact pointing responses. For all 176 trials in Phase 3, inter-rater reliability was also substantial (kappa = 0.62, 81.2% of agreement).

4.2.2 | Agreement between the created models and manual annotation for no-pointing trials

The SMV and the MLB models were applied to the data lacking clear pointing responses (176 trials in Phase 3). When focusing on AUC, the agreement between model and manual-based estimation demonstrated high accuracy only in the SMV model using face directions as input, whereas other models showed moderate accuracy despite close values (Table 1). The reliability between the SMV model and manual annotation was substantial both when using face and gaze directions, while the MLB model was moderate (kappa < 0.6).

4.2.3 | Correlation of confidence between created models and manual annotation

For 96 trials with clear pointing responses randomly extracted from Phase 1 data, confidence value in the SMV and the MLB models showed a significant positive correlation with the manually annotated

TABLE 2 Correlation of confidence value between created models and manual annotation

	For trials of clear pointing responses (96 trials randomly extracted from data in Phase 1)		For trials of no-pointing responses (All 176 trials from data in Phase 3)	
	SMV model	MLB model	SMV model	MLB model
Face angles as input				
rs	0.445	0.348	0.202	0.228
p-value	<0.0001	0.0005	0.0072	0.0023
Gaze angles as input				
rs	0.433	0.361	0.261	0.213
p-value	<0.0001	0.0003	0.0005	0.0045

one, and the correlation in the SMV model was slightly higher than in the MLB model (Table 2). For 176 trials with no-pointing responses, confidence value in all created models were still significantly positive despite being relatively weaker than the value from Phase 1, which ranged from 0.202 to 0.261.

5 | DISCUSSION

To address the question of how well looking measurements can be interpreted as equivalent to manual ones in 2AFC tasks, this study investigated how accurately pointing responses (i.e., left or right) could be predicted from concurrent preferential looking. Using pre-existing webcam-based data, we created the SVM and the MLB models and tested their prediction abilities. Results showed that toddlers' preferential looking substantially predicted pointing responses, even though looking was only roughly quantified by face or gaze using a webcam.

From Phases 1 and 2 using data with clear pointing responses, we found that both the SMV and MLB models showed equivalently high prediction accuracy. This indicates that preferential looking can be interpreted as equivalent to concurrent pointing responses at least to some extent, which is compatible with previous findings that toddlers looked longer at the stimulus that matched their subsequent reaching (Hendrickson & Friend, 2013). Regarding the temporal features of looking responses, we explored the most appropriate time window for the prediction and found that it was from 0.0 to 2.0 s after the completion of the question through the SMV model construction. The result that it was necessary to investigate at least 2 s as a potential time range is compatible with the conventional and empirical method where, in the IPL paradigm, the time window of approximately 2-s duration has been utilized for analysis (Delle Luche et al., 2015). The fact that the best time window of our results and conventional one were consistent provides supporting evidence that traditional naïve setting of time window was a reasonable way of delimiting and summarizing time-series data to reflect toddlers' volitional response to a question. However, the more appropriate time window might be found outside of the 2-s duration since the best time window was explored only within this time interval in this study. Future studies will reveal this by applying the models to other experimental tasks.

According to the "importance" visualization produced through the MLB construction, the contribution rate of each input for the prediction showed roughly an inversed U shape in accordance with elapsed time, which indicates there was no critical narrow time window reflecting toddlers' volitional choice in general. This importance visualization showed that the best time-bin contributed to the prediction was

0–200 ms in the MLB model, especially when using face angles. This might be attributed to some methodological errors (e.g., overfitting), or the adjusted starting point of the potential time window to match Japanese grammar. We set the starting point of the potential time window at the end of the question because, as mentioned above, in Japanese, a listener cannot find that the sentence is a question unless hearing the sentence through the end. The grammatical characteristics of Japanese, in which the target word is positioned at the beginning of the sentence and the interrogative marker comes at the end of the sentence, might have resulted in the highest probability of the contribution of the 0–200-ms time-bin, such that participants were able to start looking at the stimulus during the period of utterances between the target word and the interrogative marker. Besides, in this study using LightGBM (Ke et al., 2017), a similar prediction accuracy was obtained from both the SMV and the MLB models. Paradoxically, this fact showed the effectiveness of the traditional method for analysis in which the proportion of total looks to the target stimuli within a certain time window was regarded as a dependent variable. If devising input data, the accuracy of the MLB model might be improved such as adding angular velocity or facial expression variables, or converting raw time-series data to different ones.

The results from no-pointing trials (Phase 3) revealed that only the SMV model showed high accuracy and substantial agreement to manual estimations of volitional choices although each prediction index of the SMV model was close to one of the MLB model. Note that in Phase 3, manual estimation was regarded as a pseudo-correct index of pointing; however, some may be skeptical of the validity of the index itself. Indeed, there were no absolute correct answers, but considering human raters could estimate toddlers' choices from data in Phase 1 as almost perfect accuracy without seeing the exact arm movement, manual annotation seemed practically reasonable to use as a pseudo-correct index at a certain level. Overall, the substantial agreement of pointing between manual and model-based estimations indicated that toddlers possibly demonstrated their volitional choices by looking responses even when manual responses were absent. Hence, it would be practically reasonable, to some extent, to use preferential looking as alternative measures of pointing in order to avoid wasting collected data due to missing manual responses, at least for children aged 18–23 months. However, considering the prediction indices in Phase 3 dropped compared to ones in Phases 1 and 2, preferential looking without pointing would be qualitatively different from that with overt pointing. Hendrickson and Friend (2013) claimed that looking and manual responses reflect different levels of word understanding, based on connectionist studies (Munakata, 2001; Munakata & McClelland, 2003). Hendrickson et al. (2017) further revealed that, in the familiar word comprehension task, words for which toddlers did not execute reaching at 16 months were still unknown at 22 months, whereas words for which 16-month-olds reached to distracter rather than target stimulus were more likely to be known six months later. These findings indicate that the absence of manual responses itself has insightful information on early word knowledge (e.g., fragility or uncertainty).

For both the SMV and the MLB models, the correlation of confidence value between the created models and human raters remained moderate or weak, which indicated that human raters likely conducted confidence evaluation based not only on mere time-series data of face and gaze angles but also other richer information that could be obtained from videos. Extracting other variables such as facial expression might be beneficial to a more precise prediction of confidence for pointing estimation. Reliable confidence prediction can be utilized practically when estimating children's choices based only on preferential looking behavior automatically. For example, data with high confidence values may be used for subsequent analysis, whereas data with low confidence values may need manual inspection or elimination. Such techniques might be useful in a situation where arm responses are accidentally lost in webcam-based videos during the FCP paradigm due to the narrow viewing angle of a camera. Further research on confidence quantification is needed.

Although this was not the main objective of this study, we could confirm the reliability and the usefulness of OpenFace 2.2.0 (Baltrušaitis et al., 2018), which was quite important and necessary to judge if we could rely on time-series data of face and gaze angles produced by this tool. As far as a webcam is positioned at the center of the top of the screen, the screen size is relatively large, and very rough data of preferential looking is enough for analysis, OpenFace 2.2.0 may be a useful and powerful tool for the automatic coding. At this point, it is not clear whether this remains powerful when a webcam is located in a different position, but generally, it seems robust in more challenging conditions (Higuchi et al., 2018). To confirm the availability of these automated coding algorithms is quite beneficial for researchers who have limited resources or are in restricted situations to reduce the burden of data collection. Although eye trackers can collect and annotate looking responses easily (Ambridge & Rowland, 2013; Delle Luche et al., 2015) with high temporal and spatial resolution, situations where eye trackers are available are still restricted because these devices are still expensive and many of them are not handy enough to conduct experiments outside a laboratory setting (e.g., nursery schools or online experiments). In contrast, frame-by-frame manual coding is labor-intensive and time-consuming (Friend & Keplinger, 2008). High-cost data collection regarding time and money can be an obstacle to the promotion of open science, such that researchers who have fewer resources cannot collect each data sample or participate in larger international projects (Frank, 2019). The usefulness of OpenFace 2.2.0 shown in this study would be helpful to overcome such obstacles just like recent vigorous studies on webcam-based data collection (Chouinard et al., 2019; Papoutsaki et al., 2016; Scott & Schulz, 2017; Semmelmann et al., 2017; Tran et al., 2017).

Taken together, this study revealed that looking and manual measurements could be interpreted analogously. Hence, it would be practically reasonable to use preferential looking in the FCP tasks as a dependent variable instead of manual ones when the unpredicted or unignorable amount of trials lacking pointing responses is observed. However, looking measurements with and without manual ones may have different features that reflect different language abilities. Therefore, it may be recommended and beneficial that both results from pointing and looking measures, and the proportion of absent pointing responses should be reported to interpret obtained data in more detail. In fact, some researchers conducted a similar task using both modalities as dependent variables and compared differences between them (Abbot-Smith et al., 2017; Gurteen et al., 2011; Hendrickson & Friend, 2013; Hendrickson et al., 2015, 2017). Future research is needed to investigate a more nuanced relationship between looking and manual measurements. For adults or older children, this relationship has been scrutinized using more sophisticated models such as drift-diffusion models (Ratcliff et al., 2012; Thomas et al., 2019). However, since model fitting using drift-diffusion models requires numerous trials per participant (Lerche et al., 2017), there is only one study, to our knowledge, that applied them to toddlers (Leckey et al., 2020). A methodological improvement would be needed to apply these models to young children.

Another practical strategy would be to estimate and interpolate missing pointing responses (i.e., left or right) from preferential looking, leveraging the findings that both manual and looking measurements have overlapping underlying information on toddlers' volitional choice and the former measures are substantially predicted from the latter ones regardless of the execution of pointing or reaching. This strategy will help researchers prevent from wasting or excluding the collected data. However, researchers adopting this option should confirm that the results from manual measures with and without their interpolation from looking measures are compatible with each other. Researchers may have to consider what is implied by absent manual responses since such absences themselves may reflect toddlers' knowledge states about word comprehension. For future study, it would be necessary to verify directly which interpolation method reflects the true result between the created models in this study and usual statistic techniques estimating missing data, by making artificial missing data in order to evaluate how the models proposed here are useful for missing data interpolation.

At this point, it is not obvious whether the created models in this study can be applied intact to other studies using the IPL or the FCP paradigm due to a single experiment used for analysis. Additionally, although we found that preferential looking while executing pointing responses were closely related to pointing itself, it remained unclear whether this relationship could be generalized to preferential looking irrelevant to the execution of arm movements. To directly address it, an experimental design that allows separating looking and manual responses would be needed such as asking toddlers to just look at the target stimulus (e.g., "Look! There are shoes!") then asking them to point at it at least two seconds later (e.g., "Which are shoes? Point them!"). Also, it will be necessary to explore if there is a change in prediction accuracy or visualized importance tendency as a result of extending the range of the potential time window. For instance, comparing the differences between setting a starting point at the onset of the target word and at the end of the question would be beneficial to determine which is more important in choosing action: listening to the target word while understanding that the sentence is a question or merely listening to the target word. Another limitation is there were some prediction errors despite the substantial accuracy of created models. Roughly speaking, the trials in which the predictions were incorrect had any of the following features: face and gaze direction mostly remained around the center of the monitor for the target time window; pointing responses occurred within first 1 s and then the participant looked at the other stimulus; or, right up to pointing, participants looked at the side of a stimulus opposite the one chosen (some examples were described in Figure S1). Since some of these features can be dealt with by predicting the timing of pointing responses, we plan to explore the possibility of estimating reaction times for each trial. Future study is also necessary to develop software that can easily create models from other datasets using the method proposed in this study as in previous work (Kominsky, 2019).

Despite these limitations, we believe that this study plays an important role to bridge two different measurements (i.e., looking and manual) in 2AFC tasks practically and theoretically. Since 2AFC tasks have been widely used in research not only on early language development but also on other cognitive domains, the findings obtained in this study would help researchers avoid wasting collected data and interpret results from different modalities more comprehensively. Additionally, in terms of webcam-based data collection, this study will contribute to conducting online experiments as in related studies (Papoutsaki et al., 2016; Chouinard et al., 2019; Semmelmann et al., 2017; Tran et al., 2017), or further promotion of open science movement (Frank, 2019).

ACKNOWLEDGEMENTS

We thank Kei Shimonishi and Shingo Sawa for helping develop a system that can simultaneously show stimuli and record participants' responses by a webcam. We are also grateful to Sho Tsuji, Monica Barbir, Irena Lovcevic, and Akihiko Nikkuni for insightful comments and to Mana Yada, Tomohiko Kuzuba, and Hiroki Yamamoto for helping analyze the data. We deeply thank all those who participated and cooperated in our study.

CONFLICT OF INTEREST

The authors declare no conflicts of interest with regard to the funding source for this study.

DATA AVAILABILITY STATEMENT

The datasets used in this study are available in the GitHub repository (https://github.com/hagi-hara/looking-represents-choosing).

ORCID

Hiromichi Hagihara https://orcid.org/0000-0003-3316-600X
Naoto Ienaga https://orcid.org/0000-0003-4420-6361
Kei Terayama https://orcid.org/0000-0003-3914-248X
Yusuke Moriguchi https://orcid.org/0000-0002-9002-7834

REFERENCES

Abbot-Smith, K., Chang, F., Rowland, C., Ferguson, H., & Pine, J. (2017). Do two and three-year-old children use an incremental first-NP-as-agent bias to process active transitive and passive sentences? A Permutation Analysis. *Plos One*, *12*(10), e0186129. https://doi.org/10.1371/journal.pone.0186129

Ahmed, A., & Ruffman, T. (1998). Why do infants make A not B errors in a search task, yet show memory for the location of hidden objects in a nonsearch task? *Developmental Psychology*, *34*(3), 441–453. https://doi.org/10.1037/0012-1649.34.3.441

Akobeng, A. K. (2007). Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatrica*, *96*(5), 644–647. https://doi.org/10.1111/j.1651-2227.2006.00178.x

Ambridge, B., & Rowland, C. F. (2013). Experimental methods in studying child language acquisition. *Wires Cognitive Science Interdisciplinary Reviews*, *4*(2), 149–168. https://doi.org/10.1002/wcs.1215

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, *17*(2), 1265–1282. https://doi.org/10.1016/S0885-2014(02)00116-8

Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 59–66). IEEE.

Chan, A., Meints, K., Lieven, E., & Tomasello, M. (2010). Young children's comprehension of English SVO word order revisited: Testing the same children in act-out and intermodal preferential looking tasks. *Cognitive Development*, *25*(1), 30–45. https://doi.org/10.1016/j.cogdev.2009.10.002

Charles, E. P., & Rivera, S. M. (2009). Object permanence and method of disappearance: looking measures further contradict reaching measures. *Developmental Science*, *12*(6), 991–1006. https://doi.org/10.1111/j.1467-7687.2009.00844.x

Chouinard, B., Scott, K., & Cusack, R. (2019). Using automatic face analysis to score infant behavior from video collected online. *Infant Behavior and Development*, *54*, 1–12. https://doi.org/10.1016/j.infbeh.2018.11.004

Delle Luche, C., Durrant, S., Poltrock, S., & Floccia, C. (2015). A methodological investigation of the Intermodal Preferential Looking paradigm: Methods of analyses, picture selection and data rejection criteria. *Infant Behavior and Development*, *40*, 151–172. https://doi.org/10.1016/j.infbeh.2018.11.004

Durrant, S., Delle Luche, C., Cattani, A., & Floccia, C. (2015). Monodialectal and multidialectal infants' representation of familiar words. *Journal of Child Language*, *42*(2), 447–465. https://doi.org/10.1017/S0305000914000063

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670. https://doi.org/10.1126/science.146.3644.668

Fernald, A., Pinto, J. P., Swingley, D., Weinbergy, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231. https://doi.org/10.1111/1467-9280.00044

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). John Benjamins.

Fernandes, K. J., Marcus, G. F., Di Nubila, J. A., & Vouloumanos, A. (2006). From semantics to syntax and back again: Argument structure in the third year of life. *Cognition*, *100*(2), B10–B20. https://doi.org/10.1016/j.cognition.2005.08.003

Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, *31*(1), 41–81. https://doi.org/10.1006/cogp.1996.0012

Frank, M. C. (2019). Towards a more robust and replicable science of infant development. *Infant Behavior and Development*, *57*, 101349. https://doi.org/10.1006/cogp.1996.0012

Friend, M., & Keplinger, M. (2003). An infant-based assessment of early lexicon acquisition. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 302–309. https://doi.org/10.3758/BF03202556

Friend, M., & Keplinger, M. (2008). Reliability and validity of the Computerized Comprehension Task (CCT): data from American English and Mexican Spanish infants. *Journal of Child Language*, *35*(1), 77–98. https://doi.org/10.1017/S0305000907008264

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*(1), 23–45. https://doi.org/10.1017/s030500090001271x

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316–339. https://doi.org/10.1177/1745691613484936

Gurteen, P. M., Horne, P. J., & Erjavec, M. (2011). Rapid word learning in 13-and 17-month-olds in a naturalistic two-word procedure: Looking versus reaching measures. *Journal of Experimental Child Psychology*, *109*(2), 201–217. https://doi.org/10.1016/j.jecp.2010.12.001

Hagihara, H., & Sakagami, M. (2020). Initial noun meanings do not differentiate into object categories: An experimental approach to Werner and Kaplan's hypothesis. *Journal of Experimental Child Psychology*, *190*, 104710. https://doi.org/10.1016/j.jecp.2019.104710

Hagihara, H., Yamamoto, H., Moriguchi, Y., & Sakagami, M. (2020). *When "shoe" becomes free from "putting on": Initial meanings of object words are intertwined with object-specific actions*. Manuscript in progress

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development*, *21*(2), 167–179. https://doi.org/10.1016/S0163-6383(98)90001-7

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*(6), 923–929. https://doi.org/10.1111/j.1467-7687.2010.00951.x

Hendrickson, K., & Friend, M. (2013). Quantifying the relationship between infants' haptic and visual response to word-object pairings. In S. Biaz, N. Goldman, & R. Hawkes (Eds.), *Online Proceedings Supplement of BUCLD 37: The 37th Annual Boston University Conference on Language Development*. Retrieved from: http://www.bu.edu/bucld/supplementvol37

Hendrickson, K., Mitsven, S., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2015). Looking and touching: What extant approaches reveal about the structure of early word knowledge. *Developmental Science*, *18*(5), 723–735. https://doi.org/10.1111/desc.12250

Hendrickson, K., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2017). Assessing a continuum of lexical–semantic knowledge in the second year of life: A multimodal approach. *Journal of Experimental Child Psychology*, *158*, 95–111. https://doi.org/10.1016/j.jecp.2017.01.003

Higuchi, K., Matsuda, S., Kamikubo, R., Enomoto, T., Sugano, Y., Yamamoto, J., & Sato, Y. (2018). Visualizing gaze direction to support video coding of social attention for children with autism spectrum disorder. *Proceedings of 23rd International Conference on Intelligent User Interfaces, Japan* (pp. 571–582). https://doi.org/10.1145/3172944.3172960

Horváth, K., Myers, K., Foster, R., & Plunkett, K. (2015). Napping facilitates word learning in early lexical development. *Journal of Sleep Research*, *24*(5), 503–509. https://doi.org/10.1111/jsr.12306

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS One*, *10*(2), e0116494. https://doi.org/10.1371/journal.pone.0116494

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3146–3154). Long Beach, CA: Curran Associates. Retrieved from https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree

Kominsky, J. F. (2019). PyHab: Open-source real time infant gaze coding and stimulus presentation software. *Infant Behavior and Development*, *54*, 114–119. https://doi.org/10.1016/j.infbeh.2018.11.006

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Leckey, S., Selmeczy, D., Kazemi, A., Johnson, E. G., Hembacher, E., & Ghetti, S. (2020). Response latencies and eye gaze provide insight on how toddlers gather evidence under uncertainty. *Nature Human Behaviour*, *4*(9), 928–936. https://doi.org/10.1038/s41562-020-0913-y

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537. https://doi.org/10.3758/s13428-016-0740-2

Maguire, M. J., Hirsh-Pasek, K., Golinkoff, R. M., & Brandone, A. C. (2008). Focusing on the relation: Fewer exemplars facilitate children's initial verb learning and extension. *Developmental Science*, *11*(4), 628–634. https://doi.org/10.1111/j.1467-7687.2008.00707.x

Mani, N., & Plunkett, K. (2010). Twelve-month-olds know their cups from their keps and tups. *Infancy*, *15*(5), 445–470. https://doi.org/10.1111/j.1532-7078.2009.00027.x

Mani, N., & Plunkett, K. (2011). Does size matter? Subsegmental cues to vowel mispronunciation detection. *Journal of Child Language*, *38*(3), 606–627. https://doi.org/10.1017/S0305000910000243

Munakata, Y. (2001). Graded representations in behavioral dissociations. *Trends in Cognitive Sciences*, *5*(7), 309–315. https://doi.org/10.1016/S1364-6613(00)01682-X

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*(4), 413–429. https://doi.org/10.1111/1467-7687.00296

Noble, C. H., Rowland, C. F., & Pine, J. M. (2011). Comprehension of argument structure and semantic roles: Evidence from English-learning children and the forced-choice pointing paradigm. *Cognitive Science*, *35*(5), 963–982. https://doi.org/10.1111/j.1551-6709.2011.01175.x

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. In S. Kambhampati (Ed.), *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 3839–3845). New York, NY: AAAI Press / International Joint Conferences on Artificial Intelligence. Retrieved from https://www.ijcai.org/Proceedings/2016

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. E. (2012). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses. *Child Development*, *83*(1), 367–381. https://doi.org/10.1111/j.1467-8624.2011.01683.x

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology*, *80*(3), 201–224. https://doi.org/10.1006/jecp.2001.2633

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind: Discoveries in Cognitive Science*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Semmelmann, K., Hönekopp, A., & Weigelt, S. (2017). Looking tasks online: utilizing webcams to collect video data from home. *Frontiers in Psychology*, *8*, 1582. https://doi.org/10.3389/fpsyg.2017.01582

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Starkey, P., Spelke, E. S., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science*, *222*(4620), 179–181. https://doi.org/10.1126/science.6623069

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Tafreshi, D., Thompson, J. J., & Racine, T. P. (2014). An analysis of the conceptual foundations of the infant preferential looking paradigm. *Human Development*, *57*(4), 222–240. https://doi.org/10.1159/000363487

Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., & Mohr, P. N. (2019). Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, *3*(6), 625–635. https://doi.org/10.1038/s41562-019-0584-8

Tran, M., Cabral, L., Patel, R., & Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *Journal of Experimental Child Psychology*, *156*, 168–178. https://doi.org/10.1016/j.jecp.2016.12.003

Valleau, M. J., Konishi, H., Golinkoff, R. M., Hirsh-Pasek, K., & Arunachalam, S. (2018). An eye-tracking study of receptive verb knowledge in toddlers. *Journal of Speech, Language, and Hearing Research*, *61*(12), 2917–2933. https://doi.org/10.1044/2018_JSLHR-L-17-0363

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition*, *119*(1), 10–22. https://doi.org/10.1016/j.cognition.2010.11.014

Winters, S., Dubuc, C., & Higham, J. P. (2015). Perspectives: the looking time experimental paradigm in studies of animal visual perception and cognition. *Ethology*, *121*(7), 625–640. https://doi.org/10.1111/eth.12378

Yuan, S., & Fisher, C. (2009). “Really? She blicked the baby?” Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, *20*(5), 619–626. https://doi.org/10.1111/j.1467-9280.2009.02341.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Hagihara H, Ienaga N, Terayama K, Moriguchi Y, Sakagami M. Looking represents choosing in toddlers: Exploring the equivalence between multimodal measures in forced-choice tasks. *Infancy*. 2020;26:148–167. https://doi.org/10.1111/infa.12377