Journal of Experimental Child Psychology 156 (2017) 168–178

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Online recruitment and testing of infants with Mechanical Turk

Michelle Tran, Laura Cabral, Ronak Patel, Rhodri Cusack *

Brain and Mind Institute, Western University, London, Ontario N6A 5B7, Canada

ARTICLE INFO

Article history:
Received 13 June 2016
Revised 5 December 2016
Available online 11 January 2017

Keywords:
Crowdsourcing
Online research
Amazon Mechanical Turk
Infant behavior

ABSTRACT

Testing infants in the laboratory is expensive in time and money; consequently, many studies are underpowered, reducing their reproducibility. We investigated whether the online platform, Amazon Mechanical Turk (MTurk), could be used as a resource to more easily recruit and measure the behavior of infant populations. Using a looking time paradigm, with users' webcams we recorded how long infants aged 5 to 8 months attended while viewing children's television programs. We found that infants ($N = 57$) were more reliably engaged by some movies than by others and that the most engaging movies could maintain attention for approximately 70% of a 10- to 13-min period. We then identified the cinematic features within the movies. Faces, singing-and-rhyming, and camera zooms were found to increase infant attention. Together, we established that MTurk can be used as a rapid tool for effectively recruiting and testing infants.

Introduction

Infants are difficult to recruit and test. Recruiting their busy caregivers requires broad advertising, collaboration with day-care facilities or maternity hospitals, and labor-intensive relationship building. As a consequence, it often takes long periods of time to recruit a sufficient number of participants. Once recruited, the schedules of the caregivers, infants, research staff members, and testing facilities must be coordinated, and practicalities such as transport must be resolved. Given these complexities,

* Corresponding author.
E-mail address: rhodri@cusacklab.org (R. Cusack).

http://dx.doi.org/10.1016/j.jecp.2016.12.003
0022-0965/

infant studies are relatively slow and expensive and also require patience and perseverance. This puts pressure on investigators to minimize the number of participants recruited, and as a result studies are sometimes under-powered, reducing their reproducibility (Peterson, 2016), reflecting a broader issue in psychology (Open Science Collaboration, 2015). Thus, to make it easier to conduct high-quality infant research, it is imperative to find ways in which to reduce these pressures.

One solution may lie online. During recent years, the crowdsourcing engine Amazon Mechanical Turk (MTurk) has become a central marketplace, bringing together hundreds of thousands of workers from more than 100 countries to complete "human intelligence tasks" (HITs) through a web browser for modest remuneration (see Crump, McDonnell, & Gureckis, 2013, for a review; Buhrmester, Kwang, & Gosling, 2011; Kittur, Chi, & Suh, 2008; Mason & Suri, 2011; Pontin, 2007). Often these tasks involve image annotation, rating surveys, and demographic questionnaires using templates delivered by MTurk (Mason & Suri, 2011; Paolacci, Chandler, & Ipeirotis, 2010). However, by employing external websites, requestors can generate more complex tasks that meet the demands of their experimental needs (Buhrmester et al., 2011; Goodman, Cryder, & Cheema, 2013; Mason & Suri, 2011). In combination with MTurk's simple interface and flexibility, this lends itself well to the fast and cost-effective collection of data.

Recently, experimental psychologists have used MTurk to obtain data from adults on simple tasks (Lewis, Sugarman, & Frank, 2014; Piff, Stancato, Côté, Mendoza-Denton, & Keltner, 2012; Starmans & Bloom, 2012; Sweeny, Andrews, Nelson, & Robbins, 2015). Parental report measures of child behavior have also been documented (Schneider, Yurovsky, & Frank, 2015); however, to our knowledge, direct testing of infants has never been attempted. Therefore, the aim of the current study was to examine whether MTurk could be used to recruit and test infant populations. To address this, we implemented a task that aimed to quantify looking time to a set of video stimuli in infants aged 5 to 8 months. Specifically, infants viewed children's television programs, and attention was quantified by measuring when infants fixated on the screen using their webcams.

Method

Participants

Ethical approval was obtained from Western University's health sciences research ethics board. We recruited infants aged 5 to 8 months using MTurk (Amazon, Seattle, WA, USA). All workers of MTurk remain de-identified and are referred to only by a unique worker identity code provided by Amazon. To participate, infant caregivers agreed to the Amazon MTurk Participation Agreement (https://www.mturk.com/mturk/conditionsofuse), which included the declaration that they were at least 18 years old, provided informed consent, and were required to have a webcam, speakers, and Adobe Flash. The same experiment was administered in two independent batches differing in compensation rate. During the first batch, 63 participants were recruited over the course of a week, reimbursing each with $1.25 (U.S.). To motivate increased participation, remuneration in the second batch was raised to $5.00, leading to 84 participants being recruited within the following 6 days. Altogether, 147 participants were recruited. However, due to the quality control requirements of our study and the difficulty of infant testing in general, 90 were excluded. The causes were technical issues associated with internet connectivity bandwidth (no webcam video being obtained from the server ($n = 3$), the webcam video becoming desynchronized from the movie ($n = 43$), the quality of the recorded video (infant's eyes not being visible; $n = 22$), blurry video ($n = 5$), and the location of the webcam being changed ($n = 1$)). Participants were also excluded if they were not infants ($n = 11$), self-reported an age outside our specifications ($n = 3$), or did not fully complete the experiment ($n = 2$). Of the remaining 57 participants included in the study ($M_{age} = 6.49$ months, $SD = 0.93$), 9 were 5 months old, 19 were 6 months old, 21 were 7 months old, and 8 were 8 months old.

Stimuli and video recording

Ten movie clips between 9 and 13 min in length were used as stimuli. They were taken from popular programs designed to appeal to infants and children: *Baby Einstein, Blue's Clues, Curious George,*

Despicable Me, Dora the Explorer, The Program With the Mouse, In the Night Garden, Teletubbies, Timmy Time, and *Up*. To present the movies and record from the webcam, Flash was used (Adobe, San Jose, CA, USA). Content was provided from and recorded to a computer in the Amazon cloud (Amazon) running Wowza Media Streaming Server (Wowza, Golden, CO, USA) using real-time streaming protocols. Our software is available on request.

Procedure

The HIT was created and posted on MTurk under the title "Infant Television Viewing." Participants viewed a webpage detailing compensation rate, the allotted time to complete the HIT, the expiration date of the HIT, a short description of the task, and the required qualifications. After accepting the HIT, participants were directed to a webpage that provided an information sheet and asked for their informed consent. Consent was obtained via online checkbox and button press. This was followed by an evaluation of the suitability of participants' computers, software, webcams, speakers, and internet connectivity. To do this, we recorded a brief 5-s video of caregivers and their infants and asked them to move and make sounds. This video was then played back to caregivers, and they were asked to check a box to indicate whether or not they were able to see and hear themselves. If they indicated they could not, they were thanked and excluded from participation. Otherwise, they were directed to a new webpage instructing them to position their infants on their laps in the center of the screen and in a well-lit room. This was specified to ensure that infants' eyes were visible while recording the webcam videos. Once they were in a comfortable position, caregivers were instructed to press a "start" button to commence the experiment. Then 1 of 10 pseudorandomly selected movies was presented. Afterward, participants completed a short demographic questionnaire from which the ages and language backgrounds of their infants were obtained.

Video annotation

The time course of overt attention was measured from the recorded webcam videos. Using the free Anvil tool (Kipp, 2001), the experimenter and a second observer annotated when in the movie each infant fixated on the screen. The two observers agreed in their assessment of looking time 99.81% of the time with a corrected kappa of .41.

Results

Are some of the movies more engaging overall?

We first quantified overall looking time by calculating the proportion of attention in a movie for each infant. This proportion was then arcsine transformed to increase the normality of its distribution. To assess whether some movies were more engaging than others, we used a two-way analysis of covariance (ANCOVA) with factors of movie (10 levels) and infant age (in months). Post hoc *t* tests were then used to compare pairs of movies.

Fig. 1 and Table 1 show arcsine-transformed proportions of attention by movie. There was a main effect of movie on attention time, $F(9, 25) = 3.56$, $p < .001$, $\eta^2 = .562$ (Fig. 1). Post hoc comparisons using Tukey's HSD (honestly significant difference) revealed that *Curious George* ($M = .33$, $SE = .09$) engaged infants significantly less than *In the Night Garden* ($M = .74$, $SE = .06$, $p < .05$), *Teletubbies* ($M = .70$, $SE = .09$, $p < .05$), and *Timmy Time* ($M = .68$, $SE = .04$, $p < .05$). Age was not found to modulate looking time, $F(3, 25) = 2.74$, *ns*, $\eta^2 = .248$.

Are some parts of the movies more engaging to infants than others?

To determine whether the looking times were consistent within a movie, we conducted a correlation analysis that assessed the time course of attention. Our hypothesis was that if particular parts of a given movie were more engaging than others, greater similarity would be seen for the time courses of

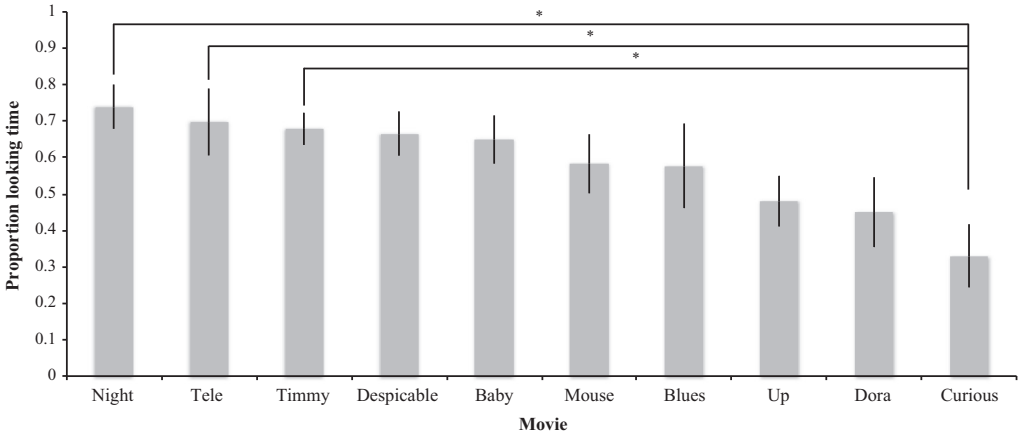

Fig. 1. Mean proportion looking times across movies ($N = 57$). Error bars represent ±1 standard error. Key to movies: Night, *In the Night Garden;* Tele, *Teletubbies*; Timmy, *Timmy Time*; Despicable, *Despicable Me*; Baby, *Baby Einstein*; Mouse, *The Program With the Mouse*; Blues, *Blue's Clues*; Up, *Up*; Dora, *Dora the Explorer*; Curious, *Curious George*.

Table 1
Mean proportion looking time for each movie, collapsed across age.

	Baby Einstein	Blue's Clues	Curious George	Despicable Me	Dora the Explorer	The Program With the Mouse	In the Night Garden	Teletubbies	Timmy Time	Up
Proportion looking time	.65	.58	.33	.66	.45	.58	.74	.70	.68	.48
SE	.07	.12	.09	.06	.10	.08	.06	.09	.04	.07
n	5	5	6	5	7	6	5	6	7	5
Movie run time (s)	604.94	652.02	767.02	680.64	654.02	637.03	735.36	552.92	580.00	736.01

attention from infants viewing the same movie than from those viewing different movies. A "binning" technique was applied in which each time series was divided into 0.5-s intervals and assigned a 0 (not looking) or 1 (looking). Due to the movies having different run times, we restricted the analysis of the time-series data to the duration of the shortest movie (552.50 s or 1105 bins).

To quantify the similarity, the Pearson correlation between the time course of each participant and each other participant was computed. A correlation of 1 would reflect identical time courses, and a correlation of 0 would represent no correspondence. We tested the hypothesis that infants watching the same movie should have a more similar time course than infants watching different movies. To do this, the mean of within-movie correlations was calculated. This was then tested against a null distribution calculated by bootstrapping, randomly shuffling the matrix of pairwise Pearson correlations of participants and taking the mean of positions in the matrix that previously held within-movie comparisons. The process was repeated 100,000 times to build the null distribution. The proportion of null values that were greater than the true value was taken as the p statistic.

Fig. 2 shows the binned time courses of attention for each of the infants grouped by movie. By eye, there do appear to be some places of some movies where infants start or stop paying attention together. However, not surprisingly, there are also large individual differences given that infants may have varying stimulus preferences or be spontaneously thinking of different things.

Therefore, we evaluated statistically whether infant attention was modulated by the content of the movies. The correlation in the time course of attention across participants watching the same movie was positive but weak ($r = .07$). However, this was highly significant when compared with the null

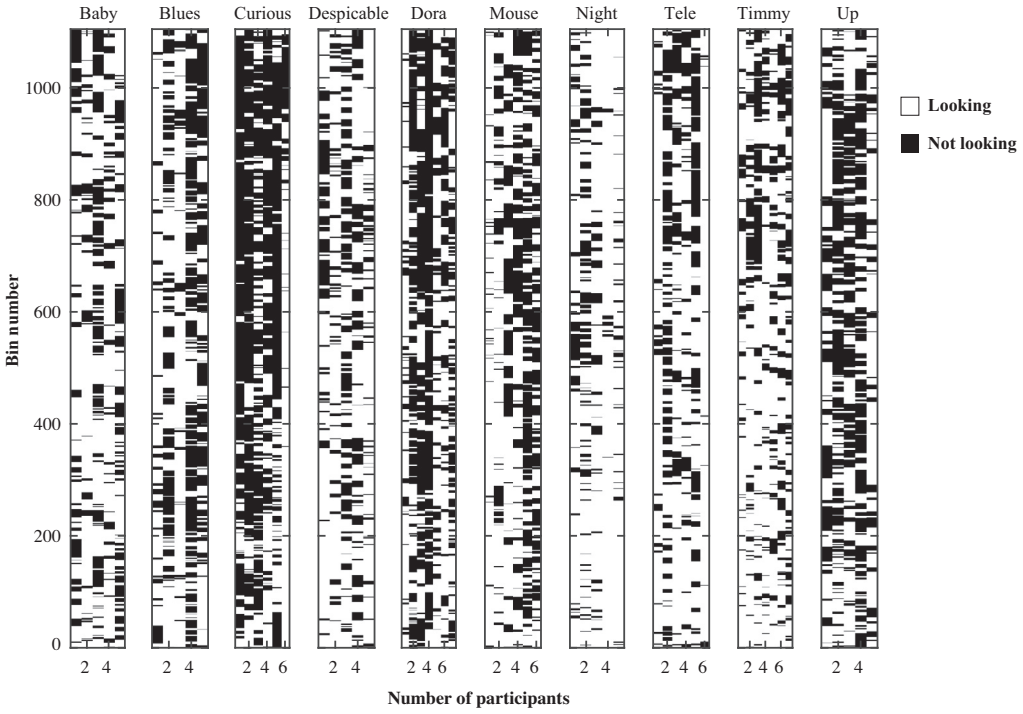

Fig. 2. Binned time courses of attention for infants grouped by movie. See Fig. 1 for key to movies.

distribution calculated through bootstrapping (Fig. 3; $p < .001$), confirming that the movie content modulated infant attention.

What features make the movies more engaging?

Furthermore, we assessed what cinematic features of the movies were driving the infants' attention. To address this, we annotated the top five movies that showed the highest proportion of looking time for production elements prominently found in infant television programs (Goodrich, Pempek, & Calvert, 2009). We annotated for 10 production features: faces, action (high and low), camera techniques (camera cut, zoom, and scene change), and auditory (background music, singing-and-rhyming, sound effects, and vocalizations) elements of the movie (Table 2). Linear regression was then used to test which cinematic features of a movie predicted the time course of infants' fixation. Specifically, from the time course of attention, $A_p(t)$ for participant $p$was modeled as the sum of the 10 production features $F_n(t)$ weighted by the participant-specific coefficients (β_{pn}):

$$A_p(t) = \sum_{n=1}^{10} \beta_{pn} F_n(t) + \varepsilon_p(t). \quad (1)$$

The weighting coefficients (β_{pn}) were derived using least-squares minimization of the residual error [$\varepsilon_p(t)$]. If there was no consistent effect of a production feature on attention, then the expected value of this coefficient would be zero. Thus, to test a feature n for significance, we performed a one-sample t test of the betas for that coefficient across participants.

For the five movies for which cinematic annotations were available (Fig. 4), we used linear regression to investigate what drove infant attention. Because no effect of age was seen in the overall looking time, we collapsed across ages and grouped infants by the movie they viewed ($n = 28$). From the 10

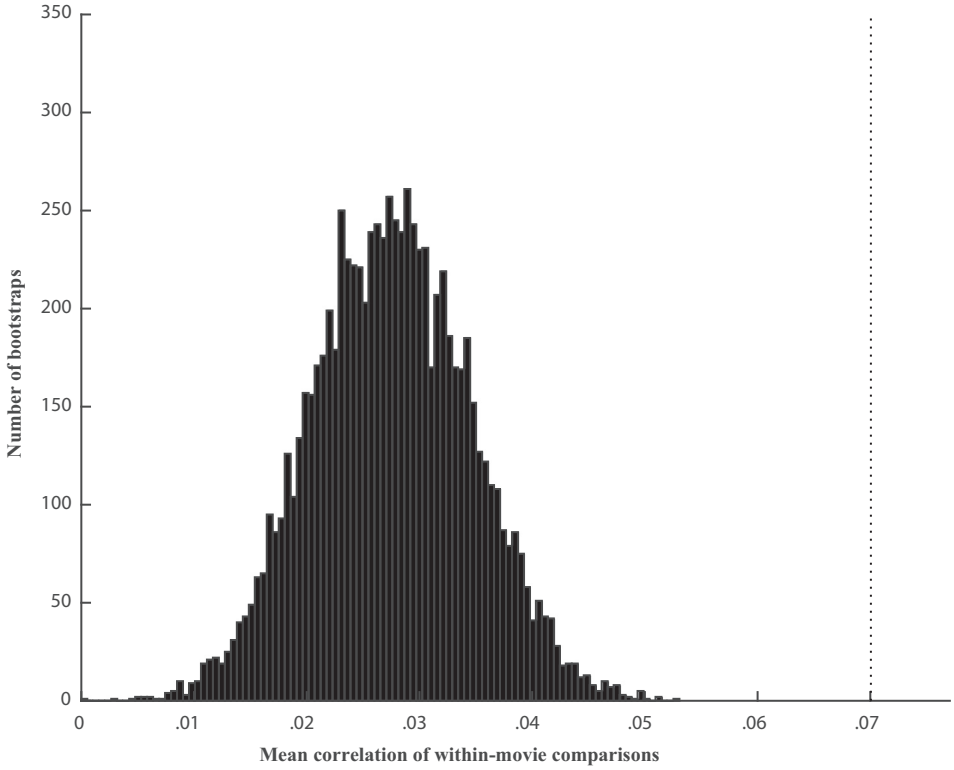

Fig. 3. Null distribution created from shuffling the mean of within-movie correlations. Dotted line represents the true measured correlation of the time course of attention across participants watching the same movie, which was well outside the null distribution.

annotated features (Table 2), it was found that singing-and-rhyming, $t(27) = 2.60$, $p < .05$, camera zooms, $t(27) = 2.16$, $p < .05$, and faces, $t(27) = 2.98$, $p < .01$, significantly increased movie engagement (Fig. 5).

We conducted an initial analysis to assess the validity of these conclusions. Linear regression (Eq. (1)) cannot separate effects when regressors $[F_n(t)]$ are highly correlated, and under such circumstances the effect of one feature might erroneously be interpreted as the effect of another feature. Therefore, we investigated the degree to which the regressors were correlated with one another. Low correlations were observed between faces and singing-and-rhyming ($r = .08 \pm .03$, mean ± standard deviation across movies), between faces and camera zooms ($r = -.15 \pm .08$), and between singing-and-rhyming and camera zooms ($r = -.03 \pm .03$). These findings demonstrate the power of the linear regression to separate the effect of production features on looking time and do not raise a concern for erroneous interpretation.

Discussion

Our exploratory study supports the principle that MTurk can be used as an efficient tool to recruit infants. In a looking time paradigm, we showed that infants aged 5 to 8 months were engaged by different child-directed movies more so than others, that some parts of the movies were more engaging than others, and that the cinematic features of faces, singing-and-rhyming, and camera zooms within the movie increased attention. No age differences in attention were found. Most important, we report for the first time an online experiment capable of capturing and quantifying infant behavior directly.

Table 2
Taxonomy and descriptions of the 10 production features.

Feature	Description
Low action	Low levels of character and/or object movement on-screen
High action	Rapid character and/or object movement on-screen
Camera cut	Abrupt change from one camera shot to another within the same scene
Camera zoom	Camera continuously moves away from or toward an object or character within a scene
Scene change	A shift from one scene to another
Background music	Music presented with dialogue and/or other sounds
Vocalization	On-screen non-language sound made by a character
Sound effect	Sound other than dialogue or music that is edited into the auditory element of a given scene (e.g., drum roll, whistle)
Singing-and-rhyming	Presenting words melodically and/or in a rhyming scheme
Faces	Visual presentation of facial features (eyes, nose, and lips) holistically in the form of faces

Source. The table is a modified version of that created by Goodrich and colleagues (2009).

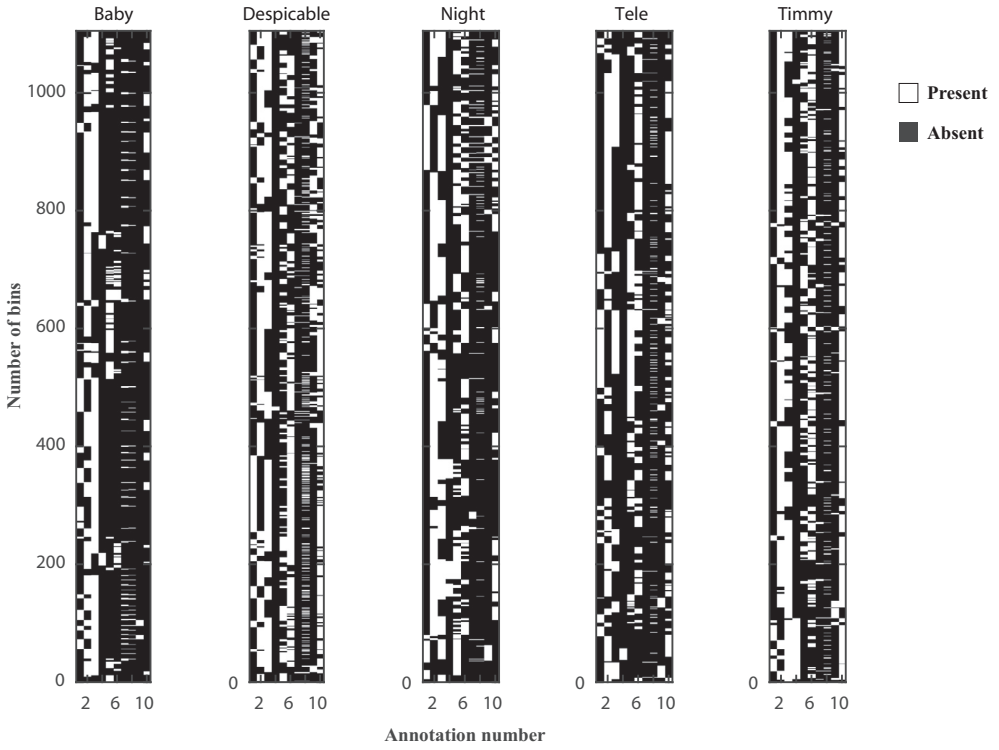

Fig. 4. Binned time courses of movies with annotated cinematic features. Annotated elements were as follows: (1) high action; (2) low action; (3) background sounds; (4) singing-and-rhyming; (5) sound effects; (6) vocalizations; (7) scene change; (8) camera cut; (9) camera zoom; (10) faces. See Fig. 1 for key to movies.

Our finding that infants demonstrate preference for faces and face-like stimuli even in the midst of distracters and dynamic visual scenes is supported by a substantial existing literature (Di Giorgio, Turati, Altoè, & Simion, 2012; Farroni et al., 2005; Franchak, Heeger, Hasson, & Adolph, 2015; Frank, Vul, & Johnson, 2009). Similarly, our finding that singing-and-rhyming attracts attention concurs with previous reports indicating preferences for engaging melodies over dissonant sounds (Costa-Giomi &

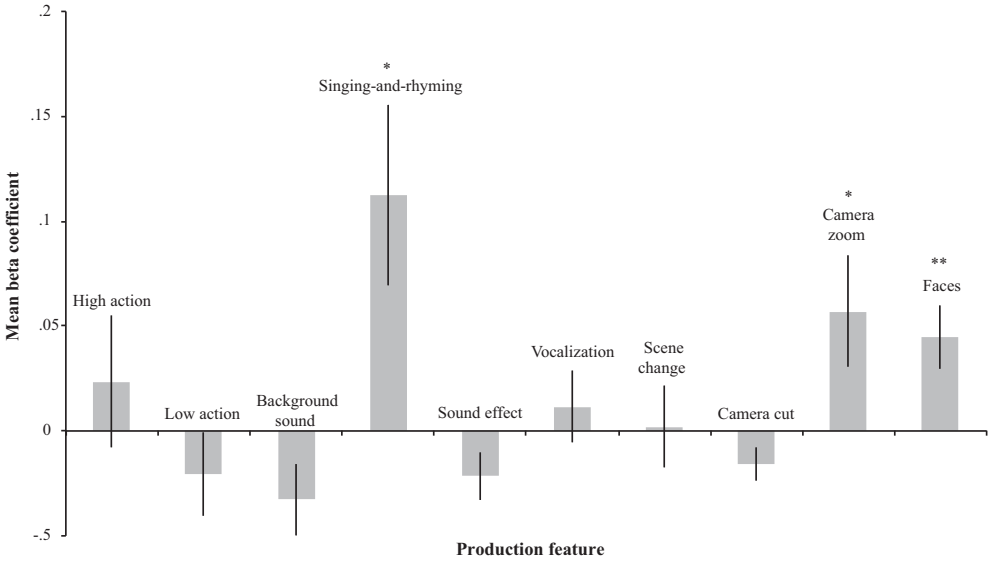

Fig. 5. Mean beta coefficients for production features across movies ($n = 28$). Error bars represent ±1 standard error. Asterisks denote $^{*}p < 0.05$ and $^{**}p < 0.01$.

Ilari, 2014; Nakata & Trehub, 2004; Trainor, 1996; Trainor, Tsang, & Cheung, 2002). High engagement toward both of these cinematic elements has been associated with internal biases reflective of maternal behavior and infants' keen interest to attend to stimuli rich in social information (Bushnell, Sai, & Mullin, 1989; Nakata & Trehub, 2004).

In addition, camera zooms also recruited increased visual attention. This agrees with recent evidence that optic flow—structured patterns of motion across the visual field—is processed in the brains of children aged 4 to 8 years (Gilmore, Thomas, & Fesi, 2016). Furthermore, even neonates have mechanisms that can quantify the degree of optic flow (Jouen, Lepecq, Gapenne, & Bertenthal, 2000). Regarding the more specific effect of optic flow on attention, an existing report found that preschool-aged children and toddlers demonstrated unchanged or reduced visual attention when camera zooms were present compared with when they were absent (Levin & Anderson, 1976; Susman, 1978). These authors suggested that camera zooms deter attention because of their tendency to disrupt the visual flow of content by taking the viewer from a whole perspective to a part perspective. The difference between these results and ours may be due to the different ages of the infants tested or to the very different stimuli. For example, the camera zooms in our stimuli served an artistic or communicative vision determined by the movie directors and, thus, may have been more congruent with the overall content in comparison with Susman's (1978) study.

MTurk recruitment was found to be easy and enabled us to collect a large infant data set in a relatively short period. Unlike in the laboratory where local participants are tested one at a time, MTurk permits workers from across the world to carry out the same HIT in parallel (Mason & Suri, 2011; Paolacci et al., 2010). The service is entirely online, which allows caregivers to recruit their children in the convenience of their own homes without needing to worry about the demands that a novel environment imposes on their children and restrictive participation time slots. Reflecting this increased convenience for participants, payments to participants were much lower than in typical studies.

Although recruitment was easy, there was a trade-off in data quality. Due to our limited control in screening participants and their equipment online, many workers (~40% of our data) needed to be excluded from our analyses. Although we implemented screening measures to constrain who could complete and view the HIT, the majority of exclusions were due to issues regarding internet connectivity rather than task performance. Therefore, although the internet is inherently involved

in using MTurk, in order to enhance data quality when bi-directional video streaming is required, internet speeds could be better prescreened in future experiments. Imposing stricter screening procedures will likely improve data quality; however, given the low cost of recruiting participants, a feasible solution could be to just accept a substantial rejection rate.

To maximize the potential of MTurk, it will be important to develop ways in which to eliminate as many interfering factors as possible. In future studies, it would be beneficial to gather additional information on the display configuration by querying information from the browser (e.g., the window size and screen resolution) and by asking users for information (e.g., the screen size, distance, and screen model). It might also establish, for example, whether a proxy for viewing distance can be obtained from infant face size, as recorded with the webcam. Ways in which to quantify lighting conditions and sound levels would also be advantageous. And further demographics, such as the sex and socioeconomic status of the infants, could be informative.

Because the study took place in participants' homes, the experimental context for each participant differed. We directed caregivers on how to seat their children for the experiment; however, we did not specify details relating to the environment in which it should be carried out. We recognized in parts of the webcam video that there were occasional distracters present in the room (e.g., toys, other people, telephone, television) that potentially could have added noise to our measures. In addition, because caregivers were aware of the movie content, they may have given unintentional cues to their infants even though caregivers were asked to remain still. Furthermore, we did not have control of the screen size or specify children's viewing distance from the computer, likely affecting the stimulus visual angle and possibly adding further noise. Furthermore, we obtained a moderate kappa value (Viera & Garrett, 2005). Although such a value could be attributed to both coders being relatively new to video annotating, the study's unconstrained viewing procedure could have made quantifying looking behavior more subjective than studies completed in the laboratory. It will be useful for future validations to include comparative measurements in a laboratory setting. These could not be conducted currently in our laboratory because it was winding down prior to a shift in location. Better understanding of the effect of online and offline contexts on infant behavior will enable us to elucidate the extent to which virtual studies produce similar outcomes to laboratory conditions. Emerging tools such as Lookit (https://lookit.mit.edu/) will be valuable in conducting these studies.

Ideally, appropriate recording conditions would involve a well-lit environment where shadowing of the face is limited and where participants' eyes are visible. In our task, we explicitly asked for infants to be in a well-lit room; however, we did not convey that infants' eyes needed to be visible. Failure to do so may have resulted in the 22 participants being omitted from analysis. Therefore, in the future it would be beneficial for researchers to provide details pertinent to the experiment so that tasks are more likely to be carried out properly. Furthermore, this could potentially reduce rejection rates. In addition, online looking time paradigms should aim to have each participant's face in the center of the screen. Our experiment asked parents to position their infants in the center of the webcam video to discern whether or not the infants were looking at the stimulus. However, what was "center" for one participant could have been definitively different for another participant. Due to the differences in screen and webcam parameters, viewing of the stimulus was likely idiosyncratic. As a solution, incorporation of facial recognition software could be implemented into the research design to standardize testing procedures and improve data quality overall. Moreover, in optimizing recording conditions, researchers should also specify that video viewing take place in an enclosed room limited in distractions.

The opportunity to record with a webcam made employing a looking time paradigm possible. In the laboratory, looking time paradigms have been widely used to study a variety of domains in infants such as intentionality (Hamlin, Wynn, & Bloom, 2007; Woodward, 1998), emotion (LaBarbera, Izard, Vietze, & Parisi, 1976; Montague & Walker-Andrews, 2001), and speech preference (Cooper & Aslin, 1990; Maye, Werker, & Gerken, 2002). Having shown in this study that behavior can be easily captured, reviewed, and quantified, this opens up the possibility of putting similar paradigms and laboratory-based studies that use comparable equipment on MTurk. This is not to say that all studies can be employed online; some require specific equipment not readily available and, hence, require local testing; however, we suggest that select tasks have the potential of being administered online.

Conclusions

Our study demonstrates that MTurk is a powerful new tool for recruiting infant populations. We designed an online study that was capable of capturing infant behavior directly and, in particular, that informed us of stimulus features in movies by which infants were most engaged. Online infant testing could reduce the high costs experienced in running experiments in the laboratory, and by removing barriers to larger samples, this could lead to increasing data reproducibility.

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Bushnell, I. W. R., Sai, F., & Mullin, J. T. (1989). Neonatal recognition of the mother's face. *British Journal of Developmental Psychology, 7*, 3–15.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*, 1584–1595.

Costa-Giomi, E., & Ilari, B. (2014). Infants' preferential attention to sung and spoken stimuli. *Journal of Research in Music Education, 62*, 188–194.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), e57410.

Di Giorgio, E., Turati, C., Altoè, G., & Simion, F. (2012). Face detection in complex visual displays: An eye-tracking study with 3- and 6-month-old infants and adults. *Journal of Experimental Child Psychology, 113*, 66–77.

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 17245–17250.

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2015). Free viewing gaze behavior in infants and adults. *Infancy, 21*, 262–287.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition, 110*, 160–170.

Gilmore, R. O., Thomas, A. L., & Fesi, J. D. (2016). Children's brain responses to optic flow vary by pattern type and motion speed. *PLoS ONE, 11*(6), e0157911.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*, 213–224.

Goodrich, S. A., Pempek, T. A., & Calvert, S. L. (2009). Formal production features of infant and toddler DVDs. *Archives of Pediatrics & Adolescent Medicine, 163*, 1151–1156.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*, 557–559.

Jouen, F., Lepecq, J.-C., Gapenne, O., & Bertenthal, B. I. (2000). Optic flow sensitivity in neonates. *Infant Behavior and Development, 23*, 271–284.

Kipp, M. (2001). Anvil—A generic annotation tool for multimodal dialogue. In: P. Dalsgaard, B. Lindberg, H. Benner, & Z.-H. Tan (Eds.), *Proceedings of the 7th European Conference on Speech Communication and Technology* (EUROSPEECH) (pp. 1367–1370). Aalborg, Denmark.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–456). New York: ACM.

LaBarbera, J. D., Izard, C. E., Vietze, P., & Parisi, S. A. (1976). Four- and six-month-old infants' visual responses to joy, anger, and neutral expressions. *Child Development, 47*, 535–538.

Levin, S. R., & Anderson, D. R. (1976). The development of attention. *Journal of Communication, 26*, 126–135.

Lewis, M., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavioral Research Methods, 44*, 1–23.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, 101–111.

Montague, D. P. F., & Walker-Andrews, A. S. (2001). Peekaboo: A new look at infants' perception of emotion expressions. *Developmental Psychology, 37*, 826–838.

Nakata, T., & Trehub, S. E. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behavior and Development, 27*, 455–464.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. http://dx.doi.org/10.1126/science.aac4716.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic World, 2*, 1–10. http://dx.doi.org/10.1177/2378023115625071.

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., & Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 4086–4091.

Pontin, J. (2007). *Artificial intelligence, with help from the humans*. The New York Times. Retrieved from http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html?_r=0.

Schneider, R. M., Yurovsky, D., & Frank, M. C. (2015). Large-scale investigations of variability in children's first words. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Starmans, C., & Bloom, P. (2012). Windows to the soul: Children and adults see the eyes as the location of the self. *Cognition, 123*, 313–318.

Susman, E. J. (1978). Visual and verbal attributes of television and selective attention in preschool children. *Developmental Psychology, 14*, 565–566.

Sweeny, K., Andrews, S. E., Nelson, S. K., & Robbins, M. L. (2015). Waiting for a baby: Navigating uncertainty in recollections of trying to conceive. *Social Science and Medicine, 141*, 123–132.

Trainor, L. J. (1996). Infant preferences for infant-directed versus non-infant-directed playsongs and lullabies. *Infant Behavior and Development, 19*, 83–92.

Trainor, L. J., Tsang, C. D., & Cheung, V. H. W. (2002). Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception: An Interdisciplinary Journal, 20*, 187–194.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*, 360–363.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1–34.