研 究 終 了 報 告 書

「Developing datasets of infant behavior that are exploitable by AI」

研究期間: 2020 年 12 月～2023 年 3 月
研 究 者: 辻 晶
加速フェーズ: 2023 年 4 月～2024 年 1 月

1. 研究のねらい

For experimental researchers in developmental psychology and related fields, infant gaze is a key outcome measures with which to access their cognitive processing. While such studies have traditionally been run in specialized infant labs, where high-quality eye tracking is available, recently remote studies, allowing infants and their caregivers participate from the comfort of their home via webcam recording, have become more popular. Remote studies yield the promise of accessing larger, more diverse samples in a more naturalistic environment. However, automatic gaze coding of infant gaze is still error-prone, making necessary costly hand-coding efforts. The aim of the present project was to develop a dataset of infant looking time studies big and diverse enough to allow the improvement of automatic gaze coding quality; at the same time creating an accessible infrastructure for infant researchers to collect and automatically analyze such datasets, including ethical and practical aspects. The acceleration phase further served to wrap up these aims, specifically tackling problems of timing precision and noise, and creating an infrastructure for infant researchers.

2. 研究成果

(1)概要

Phase 1 of the present project was to establish an infrastructure for creating the desired large and diverse dataset. We achieved this by internationalizing extant experimental software, by establishing a network of participating labs, by developing two tasks, the first purely visual, the second linking visual and auditory stimuli, for standardized data collection, and by establishing ethical and practical standards for large-scale data collection. Phase 2 was to collect a large and diverse dataset, to hand-code parts of it and to compare it with existing automatic coding. We have collected and hand-coded parts of this dataset. We have not progressed to expanding and diversifying data collection yet, because we became aware of the importance of adding a third component before doing so, as follows. Phase 3 originally was to improve automatic gaze coding algorithms based on the datasets we collect. However, we realized that before targeting the algorithms themselves, it would be more efficient to target improving data quality. We therefore identified those environmental noise factors that impact automatic gaze coding the most, and improved parental instructions to precisely improve these factors (Phase 3+). This pushes back the planned improvement of algorithms, which we now are convinced should be tackled in different ways than originally proposed, into Phase 4. We hope to get the opportunity to tackle it in an

extension phase.

In the acceleration phase, we originally planned to put collected data into a challenge connected to a talk or workshop at a Computer Vision conference. However, this plan was not feasible given the PI's institutional move in early 2024. We therefore used the time and resources in alternative productive ways, finding solution to measure and reduce noise factors detected during the collection of online data, creating documentation for researchers, and building collaborations for future study.

（2）詳細

Phase 1: Building online testing infrastructure

To create an international solution for collecting remote infant looking time data via webcam, we decided to build on the Lookit platform (lookit.mit.edu; Scott & Schulz, 2017). We decided on this solution over creating our own, since it offered all the elements we needed except for internationalization beyond English-speaking countries and countries governed by non-US data protection standards. Other reasons for choosing Lookit was that it is currently the only platform customized for the specific needs of infant looking time studies, and because it is a non-commercial, but open-source platform. The Technical Implementation part of Phase 1 was meant to create the technical infrastructure for internationalizing Lookit, which we tackled in three parts: (a) Creating the architecture for internationalization by linking the various parts of the Lookit server (both html and flatpage frontend website content and ember frameplayer and javascript experimental task content) via Github to translation tools (POEditor); (b) Discussing and implementing internationally compatible standards for demographic information within the Lookit participant entry survery; for instance, it is uncommon to ask for "race" in Japan and other countries outside the US, or US education categories do not fit Japanese and other schooling systems, which we did while discussion with the wider ManyBabies network, which has experience with such aspects; (c) Creating a network of translators that translate the Lookit content to various languages, which we succeeded doing by recruiting from the ManyBabies network. We now have volunteers for translations in 20 languages, which are in progress or have completed a first pass. In principle, a Japanese version is now available on https://lookit.mit.edu/ja/ (with other languages to follow). A few more adjustments are necessary to make this Japanese language version fully functional; once this is the case, we intend to inform the Japanese developmental community via mailing lists, publications, and conference announcements.

As for the *Implementation of ethical and data protection procedures* part of Phase 1, in addition to approved ethics at the University of Tokyo, we have succeeded in getting Ethics approval at University of Nottingham covering European countries tied to GDPR. We took great care to create safe data transfer and extensive informed consent procedures to match GDPR standards. This ethics approval will serve as an umbrella approval for EU participating labs, although depending on the university participating labs might need to apply for separate ethics.

Phase 2: Paradigm & Testing

Choosing stimuli & building experiment. We have implemented the two experimental paradigms planned during the grant phase. Experiment 1 tests the feasibility of online infant looking time studies in a purely visual context. We have constructed a set of visual stimuli that allows us to assess infants' gaze direction, as well as differences in their amount of looks as a function of salience. We varied stimulus position (left, right), salience (still, moving), and complexity (low, high). We have designed a first English version of Experiment 2, which is a word recognition task in which both audio and visual stimuli are combined (Figure 1, Experiment 2). In this task, we are, first, assessing whether we can replicate toddlers' successful gaze orientation to the correct object from lab studies in online studies. We are, second, looking at whether the way words are presented affect accuracy (thus, whether presenting words in a context like "this is a key" versus "a key" versus just "key" matters). This is an important question towards internationalization, since on the one hand presenting stimuli within sentence frames makes word recognition easier for young children, but on the other hand sentence frames across languages are very different (for instance, in Japanese the target word would often occur at the start of a sentence ("Kagi da yo"), which gives the participant less time to prepare a gaze response. Insights from this experiment will thus help to create rules for different language versions, for which we will need to choose stimuli that are comparable in characteristics such as familiarity and salience. For both experiments, we have been working using collaborative documents to crowd source expertise from interested researchers (Experiment 1; Experiment 2). With regards to *Recruiting participating labs*, early in the project we successfully applied for membership of our project in ManyBabies (Frank et al., 2017), an international consortium to improve replicability in infant development research by cross-lab replication. Being known as ManyBabies-AtHome gives us visibility and interest from many international infant development labs, which are ready to invite participants from their databases once our internationalization is successful, at least for those labs where their university will allow participation under the umbrella ethics approval. As to *Running tests*, we have currently tested our two experiments in English. This is on the one hand due to the delay in ready-to-use internationalization, however, on the other hand Phase 3+ below showed us the importance of improving data quality before starting large-scale data collection. Therefore, starting large scale testing after Phase 3+ turns out to be ideal.

Phase 3: Annotation and Analysis

Establishing a standardized *hand annotation* pipeline is critical in order to compare results from different sites. We have created detailed video instructions of each step of the hand coding procedure, including installation of software, how to interpret infant gaze, and how to annotate. For all steps, we took care to use open source tools that are accessible everywhere (R, R Core Team, 2022, and Elan, Sloetjes & Wittenburg, 2008). We have created a dedicated Wiki page on the Open Science Framework which is still set to private to tie up some loose ends, but which we plan to make public at the end of this project. We have so far hand coded most of the results of the English version of E1; hand-coded results so far show us that the task can successfully detect preference for dynamic over static stimuli, but no difference regarding stimulus. As to Experiment

2, we have finished hand-coding the data (Nederlanden, Holzen, & Tsuji, in preparation). We found that toddlers can successfully recognize words in an online setting, at that the sentence frame matters. The second part of Phase 3 was originally intended to be an optimization of existing automatic gaze coding algorithms based on our new input datasets. We decided to change this last part, since we realized that before targeting the algorithms themselves, it would be more efficient to target improving data quality, as follows.

Phase 3+: Improving data quality

Infant data are noisy. This is true in lab studies, but the problem is exuberated in at-home settings, where no experimenter is present to prepare the environment or instruct caregivers. Indeed, while coding 61 videos from infant at-home studies we discovered a wide range of noise factors such as bad lighting, positioning, or infant movement, and that the majority of videos contained several of them. Such environmental noise likely worsens performance of existent automatic gaze coding algorithms, but it is currently unknown to what extent this is the case and which of these factors matter most. To answer this question, based on the most commonly occurring factors we identified from the 61 infant videos, we designed an adult experiment to simulate these factors. We chose an adult setting since this enabled us to manipulate these factors systematically experimentally. We chose lighting, position relative to screen, head tilt, and a range of visual angles to account for different possible screen sizes. We have tested a cross-cultural sample of 30 participants in Tokyo (19 female, mean age = 28.3 years), and 31 participants in Dublin (13 female, mean age = 24.7 years). These data have been analyzed with the state-of-the-art automatic gaze detector iCatcher+ (Erel et al., 2022) to assess which factors affect model performance the most. We assessed the percentage of successful face detection and correct prediction of gaze direction using linear mixed effects modeling. As to gaze detection, it had almost 100% accuracy under ideal conditions, but all potential noise factors significantly affected detection with distance from screen having an especially strong influence. As to correct prediction of gaze direction, maximum accuracy dropped to around 60%, and noise factors additionally affected outcomes. The results of this study are currently in the minor revision stage at the journal *Behavior Research Methods*. Based on this analysis, we have created parental instructions focusing on the most important noise factors, and we will run Experiment 1 with and without these instructions in the future.

The acceleration phase was intended to focus on collecting a big data set and submit it to a challenge for improving automatic gaze coding algorithms. Because of the PI's institutional move and the impossibility to keep the grant until end of March 2024, we were not able to put this plan into action (the challenge submission would have been associated with a conference presentation, but possible conference dates either conflicted with my moving dates, or would have taken place after the grant period); however, we have used part of the resources to collect and annotate relevant data, and we will realize this plan at a later stage. Instead, we decided to use the remaining funds in productive ways to tackle further noise factors detected during the collection of online data, to create documentation for researchers, and to build collaborations for future study.

Tackling noise factors

In addition to the noise factors associated with infant home environmental factors identified above, we found our online studies to suffer from severe timing issues, due to delays during transmission. We initially noticed this problem because of inconsistencies between the trial onset relative to video start reported by the stimulus presentation software and the timing of the sounds we were able to hear in the final videos. We identified that such delays could affect the final videos in four ways. Of these, the audio–video asynchrony that might have occurred during presentation (AVP) is not directly measurable for us, but the other three might be.

Audio–audio asynchrony (AA) was easy to detect by, as mentioned above, comparing the difference between reported and actual stimulus sound onset, if a sound was present. We successfully tackled the AA asynchrony by implementing a routine where the experimental stimulus is always preceded at a constant timing by a sound that is easy to detect automatically over environmental noise. We wrote a script to automatically detected the actual trial onset from the sound in the videos.

In theory, video–video asynchrony (VV) could be tackled in similar ways, namely by comparing the difference between reported and actual visual stimulus onset. However, different from stimulus sounds that are available in recorded videos, these videos only show a recording of the infant's face, but not of the stimuli they see on the screen. We experimented with flashing images with highly contrastive luminance on the screen to see whether we were able to automatically detect these differences from the reflections in the recorded video, however, this was only possible in relatively dark room settings, which are not realistic for infant studies. The anecdotal trials we performed suggest that the delays were comparable to the AA ones, which led us to not pursue this avenue further for the moment. More rigorous tests should be conducted in the future.

As to audio–video asynchrony in the recorded videos, we had manual coders go through part of our videos and tag instances where there was a salient movement with associated noise (e.g. child vocalization, tapping on an object) and code whether or not the video and audio were synchronous. This analysis showed that there was virtually no AV asynchrony such that this factor can be considered as not problematic.

Sensitivity of automatic gaze coding to different stimulus conditions

Another factor important for researchers is to know for what kind of remote study is recommendable to use automatic gaze coding. While our current study protocol offers a somewhat limited window into the range of possible studies, we were able to assess the effect of two factors: The effect of globally more attention–catching stimuli (presence of absence of background music in a trial) and the effect of selectively more attention–catching stimuli (moving versus static stimuli). If some of these factors lead to more unambiguous gaze behavior, for instance because they catch infants' attention to a greater extent, this could lead to a better correspondence between manual and automatic coding. We did not find evidence for such an effect, with no significant differences in the correspondence between manual and automatic gaze coding based on these factors.

Publishing best practices for online infant studies

We are currently summarizing our findings, solutions, as well as other resources on our OSF page (https://osf.io/axtqr/wiki/home/), with the purpose of making users aware of possible sources of noise, how they can deal with them, and to provide scripts to help with this process. We have also written up an instructional paper in Japanese for the Japanese developmental research community, for the Special issue "Towards infant and preschooler research in post-COVID era" (ポストコロナ時代の乳幼児研究に向けて) in Japanese Psychological Review (心理学評論) (currently in revision).

References

Erel, Y., Shannon, K. A., Scott, K., Cao, P., Tan, X., Hart, P. K., ... & Liu, S. (2022). iCatcher+: Robust and automated annotation of infant gaze from videos collected in the lab and online.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Infancy, 22(4), 421-435.

Nederlanden, der S., Holzen, van K., & Tsuji, S. (in preparation). Testing Infants' Word Comprehension Online Using a Looking-While-Listening Procedure: Age and Carrier Phrase Matter.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. Open Mind, 1(1), 4-14.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)

3. 今後の展開

At the end of the project phase, this project will have allowed us to create an infrastructure for international, diverse, assessment of infant cognition from their home, including best practices to gather high quality data. On the short term, it would be desirable to complement this set of achievements with possibilities to share these data widely to enable other researchers to build on these data and to improve automatic gaze coding. This is a topic we address in our application for the extension phase.

For my own research, which has so far mainly been centered on in-lab experiments, having gained this infrastructure is an invaluable and immediate addition to the toolset I have available. Remote infant studies are a rapid and low-cost tool to pilot factors such as stimuli, different conditions, or age groups. On the mid-term during the next 3 years, I am also looking forward to using such remote studies for collecting more diverse samples – for instance accessing rural or low socio-economic-status families in Japan, or to access international samples. I would like to interface with experts I have encountered in the Act-X network such as Prof. Yukino Baba to set up citizen science and crowd sourcing experiments using remote

infant testing. I can imagine a participative three-step process, for instance in collaboration with science museums such as Miraikan, where I have previously run experiments (Phillipsen, Nagai, & Tsuji, 2022, Frontiers in Neurorobotics), where we would (1) crowdsource parents' pressing questions from parents, (2) find researchers that are experts on these topics and design adequate online experiments (3) use a citizen science approach to gather data. Such an effort could take place within the next 5 years and could have a transformative effect on the interface between science and society. I have adequate experience to set up such an endeavor, since I have already set up a successful multi-lingual science communication blog based on topics crowdsourced from parents (https://www.kotoboo.org/), thus, this idea could build on this.

Finally, I am excited about contributing to further technological developments, for instance using automatic recognition technology to code for behavior beyond gaze, or to work towards online gaze recognition in infants.

4. 自己評価

I will critically reflect upon the following topics: 1) Project delays, 2) Relationship to AI, and 3) Network formation.

1) Project delays

A major holdup that impacted the whole project was the internationalization of the experimental software, which was largely due to external factors out of my hands, and due to going with low-resourced open-source solutions. Even if progress can be slower when one creates open-access, generalizable solutions, I am still convinced we chose the right tools and partners for this project. In hindsight, with more experience I could have set more explicit timelines and deliverables for all partners to prevent the delays encounted to a certain degree, which is a valuable lesson for the future. Apart from this, figuring out ethics and creating generalizable instructions as to study, coding, and analysis protocols, which is all "invisible" work that does not directly produce results, also took a large chunk of time. I am, however, convinced that this focus on a sustainable solution was the right one. In the end, as mentioned above, the delay also allowed us to focus on data quality aspects in the meantime, which we can now incorporate into the large scale data collection effort and will therefore make it better than it would have been at an earlier point in time.

This project puts a spotlight on a globally collaborative, crowdsourced, inclusive way to conduct science, and I am thankful this grant gave me the opportunity to build this framework with sufficient time and resources.

2) Relationship to AI

Another issue I would like to address is the relationship to AI, since we did not work directly on machine learning algorithms and the like. Although we did propose this as a last step in the original proposal, the focus of this project was always to create a dataset that is exploitable by AI later. Indeed, gathering baby datasets is my expertise, and my experience

collaborating with computationalists allows me to understand what kind of datasets can be useful for improving AI. Therefore, the best way I can contribute to AI is indeed by focusing on creating data, and to propose how they can interface with AI – both by using AI methods both to support data collection (by enabling automatic gaze coding), and to deliver data that can improve AI (by creating otherwise hard-to-access unique infant datasets).

3) Network formation

I immensely profited from feedback I got during the network meetings. For instance, suggestions to expand the scope of automatic coding beyond infant gaze, to increase data quality not only by parental instructions, but also by using automatic software solutions (for instance that detect and correct facial position), and to put this research in the context of crowdsourcing and citizen science have all deeply resonated with me and are future research avenues I will pursue with some certainty, and hopefully in collaboration with the researchers that gave me these suggestions.

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：2件

1. Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., & Bergmann, C. A Global Perspective on Testing Infants Online: Introducing ManyBabies-AtHome. Frontiers in Psychology. 2021. 12. 3811.
Online testing holds great promise for infant scientists. It could increase participant diversity, improve reproducibility and collaborative possibilities, and reduce costs for researchers and participants. However, despite the rise of platforms and participant databases, little work has been done to overcome the challenges of making this approach available to researchers across the world. In this paper, we elaborate on the benefits of online infant testing from a global perspective and identify challenges for the international community that have been outside of the scope of previous literature. Furthermore, we introduce ManyBabies-AtHome, an international, multi-lab collaboration that is actively working to facilitate practical and technical aspects of online testing and address ethical concerns regarding data storage and protection, and cross-cultural variation. The ultimate goal of this collaboration is to improve the method of testing infants online and make it globally available.
2. Tsuji, S., Amso, D., Cusack, R., Kirkham, N., & Oakes, L. M. Empirical Research at a Distance: New Methods for Developmental Science. Frontiers in Psychology. 2022. 12. 3011.
The COVID-19 pandemic presented many challenges for the research community. The

collection of papers in this Research Topic illustrate how developmental scientists met those challenges and created clever and innovative methods to continue research when it was not safe to have children and families physically in the lab. Soon after labs were closed by universities and institutions, developmental scientists were scheduling video conferences with children to collect data, programming web-based procedures for participation, and considering ways to reevaluate previously collected data. The papers presented here demonstrate how the community continued to conduct research even though we were not able to work directly with our participants.

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

Von Holzen, K. & the ManyBabies-AtHome Consortium（2022, October 28）. ManyBabies -AtHome. Oral presentation at the Big Team Science conference (2022 BTSCON), Virtual Conference.

Hagihara, H., Zaadnoordijk, L., Cusack, R., & Tsuji, S. (2022, September 23). A video dataset for the exploration of factors affecting webcam-based automated gaze coding. Innovations in Online Research 2022, Online.

Sander-Montant, A., Marie-Louise, L., Barbir, M., Crimon, C., The Kotoboo Consortium, & Tsuji, S. (2022, July 7-10). Bringing developmental science to the spotlight: how to make research accessible to parents. Poster presented at the Biennial International Congress of Infant Studies 2022, Ottawa, Canada.

Zaadnoordijk, L., Bergmann, C., Buckler, H, Cusack, R., Tsuji, S. (2022, January 10 -14). Making Online Testing Accessible Across the World: Insights from ManyBabies-AtHome. Poster presented at the virtual Budapest CEU Conference for Cognitive Development.

Tsuji, S., Bergmann, C., Buckler, H., Cusack, R., & Zaadnoordijk, L. (2021, April 7-9). Toward a large-scale collaboration for infant online testing: Introducing ManyBabies-AtHome. Oral presentation at the Virtual Biennial Meeting of the Society for Research in Child Development (SRCD 2021).

Tsuji, S. (2021, Mar 29). Remote looking time studies during the new normal. Invited Presentation at 日本発達心理学会 第 32 回大会.

Tsuji, S. (2020, Nov 28). ManyBabies-AtHome: ウィズコロナ時代に適応した乳幼児遠隔視線行動測定方法の確立(ManyBabies-AtHome: Establishing remote infant looking time methods in the

with-corona era. Invited Presentation at Symposium シンポジウム2：ウィズコロナ時代の安全な研究環境 (Safe research environments in the with-corona era) at the 9th 日本発達神経科学会 (Conference of the Japanese Society for Developmental Neuroscience), Virtual Conference

Bergmann, C., Buckler, H., Cusack, R., Tsuji, S., Zaadnoordijk, L., & The Manybabies-AtHome Consortium (2020, Oct 22-23). Toward a large-scale collaboration for infant online testing: Introducing ManyBabies-AtHome. Many Paths to Language (virtual MPaL 2020), Nijmegen, The Netherlands.