Beyond Screen Time: Using Head-Mounted Eye Tracking to Study Natural Behavior

John M. Franchak

Department of Psychology, University of California, Riverside

900 University Avenue, Riverside, CA

Chen Yu

Department of Psychology, University of Texas at Austin

108 E Dean Keeton St, Austin, TX 78712

Abstract

Head-mounted eye tracking is a new method that allows researchers to catch a glimpse of what infants and children see during naturalistic activities. In this chapter, we review how mobile, wearable eye trackers improve the construct validity of important developmental constructs, such as visual object experiences and social attention, in ways that would be impossible using screen-based eye tracking. Head-mounted eye tracking improves ecological validity by allowing researchers to present more realistic and complex visual scenes, create more interactive experimental situations, and examine how the body influences what infants and children see. As with any new method, there are difficulties to overcome. Accordingly, we identify what aspects of head-mounted eye tracking study design affect the measurement quality, interpretability of the results, and efficiency of gathering data. Moreover, we provide a summary of best practices aimed at allowing researchers to make well-informed decisions about whether and how to apply head-mounted eye tracking to their own research questions.

Keywords: Eye movements, head-mounted eye tracking, mobile eye tracking, ecological validity, perceptual-motor development, joint attention, language development, computer vision, social attention

MANUSCRIPT LENGTH: 9,305 words, 4 figures, 70 references

1. Introduction

Psychology is an empirical science. Technological and methodological advances offer psychologists barrier-breaking research opportunities to collect new

Email addresses: franchak@ucr.edu (John M. Franchak), chen.yu@austin.utexas.edu (Chen Yu)

types of data, improving the quality of measurements provided by older methods. Better measurement can mean improvements in accuracy and/or precision. In the 1930s, McGraw (1935) filmed infants' footfalls, allowing unrivaled accuracy and precision in describing the parameters of infant gait compared to experimenter observation. Better measurement quality can also mean improvements in construct validity—documenting behavior in a way that more closely aligns with the psychological construct. Although accurate and precise, films of infants traversing a straight path in the lab fall short of describing how infants actually walk in day-to-day life. The advent of portable camcorders allowed researchers to capture the footsteps of infants in the home, improving the construct validity and ecological validity of locomotor measurements (Adolph et al., 2012).

Eye movement recording has seen a similar progression (Figure 1). The advent of *screen eye trackers* (SETs)—devices that record where a seated observer looks on a computer display—allowed researchers to accurately and reliably record eye gaze, but in a limited context. More recently, *head-mounted eye trackers* (HMETs)—wearable eye tracking devices that can measure eye movements in a mobile observer—allow researchers to study eye gaze in more naturalistic situations, such as driving (Land & Lee, 1994), preparing food (Land & Hayhoe, 2001), playing sports (Land & Furneaux, 1997), or going on a hike (Matthis et al., 2018). The thesis of this paper is that both aspects of measurement quality—accuracy/reliability and construct validity—are important when researchers choose how to measure behavior. Our paper will focus on the benefits and challenges of head-mounted eye tracking to illustrate how researchers can strike a balance between these two aspects of quality. Some perspectives on the new technology have been unabashedly optimistic: "... newly available portable eye-tracking systems offer a cost-effective, easy to apply, and reliable measure of eye gaze and saccades in an ecological environment" (Shamay-Tsoory & Mendelsohn, 2019). Such optimism has faced opposition. Indeed, Hessels et al. (2020) harbor such strong concerns that they claim, "Without realistic expectations of what wearable eye trackers (and wearable technology in general) may deliver, research funds and the careers of young researchers are potentially wasted."

Here, we offer an alternative to these extreme views—one that is neither as carefree as Shamay-Tsoory & Mendelsohn's nor as dire as Hessels et al.'s. We believe that head-mounted eye tracking can be challenging, but the challenges are not so prohibitive to outweigh the benefits of gathering naturalistic data. We begin the paper by describing the characteristics of head-mounted eye tracking and discuss the research pay-off: Why is it important to measure eye gaze in mobile observers? Next, we review the challenges and solutions of employing head-mounted eye tracking methods. Because head-mounted eye tracking systems vary in quality, can be applied in a wide range of contexts, and can be analyzed in myriad ways, we argue that researchers' decisions about how to implement the method determine the ease of use and quality of data. Accordingly, we argue that general claims about whether head-mounted eye tracking is (or is not) cost effective, easy to apply, and reliable should be avoided. Rather, we describe how each issue depends on researchers' decisions about how to apply

the method to their question of interest.

2. What is head-mounted eye tracking?

Comparing newer head-mounted eye trackers (HMETs) to the older and more ubiquitous screen eye trackers (SETs) helps to make it clear what limitations of SETs are solved by using HMETs. As Figure 1A shows, a SET consists of a monitor for displaying photographic or video stimuli. Beneath the monitor, a specialized set of high-speed cameras are positioned to record the eyes of an observer seated nearby. Because the cameras are fixed in place, the observer's head must remain within a small trackable areas in view of the eye tracking cameras. This precludes the observer from moving (other than small shifts in body posture). The fixed cameras and fixed observer mean that eye gaze can be tracked in a small region above the eye tracking cameras. Although this is often a computer monitor, some fixed eye tracking systems allow the monitor to be removed so that a live stimulus can be presented in that space. Either way, SET systems cannot measure how an observer looks in all different directions with respect to the body; once the head turns to the side, the observer's eyes are no longer in view to track.

A HMET solves this problem by placing the cameras on the participant's head (Figure 1B). An *eye camera* on the headgear moves with every head turn, providing a constant image of the observer's eye to detect the position of the pupil and/or corneal reflection. Likewise, an outward-facing *field of view camera* (FOV camera) fixed to the observer's head captures a video of what is in view at every moment. A calibration process, in which the observer is instructed to fixate a set of known points, creates a model of how pupil/corneal reflection positions in the eye camera correspond to gaze position in the FOV camera. The right panel of Figure 1B shows an infant participant's calibration—toys are presented in different windows in a grid to elicit looks across the visual field. The resulting data are horizontal and vertical coordinates of the estimated point of gaze within the head-centered FOV camera. Typically, HMET systems superimpose a gaze cursor on the FOV video to visualize where the observer looks as the contents of the FOV shift from moment to moment due to head and body movements. This *gaze overlay video* can be used in subsequent coding, as we will discuss below.

3. Why use head-mounted eye tracking?

Visual attention when passively viewing visual stimuli on a computer screen differs from visual attention in the real world (Foulsham et al., 2011; Tatler et al., 2013; Franchak, 2020b). Many free viewing paradigms using SETs gather eye movement data in experimental tasks that lack goals and actions. As a result, the findings based on screen-viewing paradigms may not be generalized and replicated in more naturalistic contexts. This is because the most important role of vision and visual attention in the real world is to guide everyday actions

in the complex and dynamic visual world to achieve behavioral goals. Without up-to-date information about the state of the world, it is difficult for people to know what the consequences of an action are likely to be (Hayhoe, 2000). HMETs provide a window into the moment-to-moment visual information perceived by the wearers when they engage in everyday activities (Pelz et al., 2001). When a wearer moves relative to the environment, egocentric vision captured by a HMET is a product of the simultaneous movements of eye, head, hand, and body, which serves two important roles. First, the visual information gathered about the world is used to support visually guided behavior. Second, the visual information in egocentric vision is used to estimate the change in viewer location in the environment. Because HMET reveals how visual information is collected to serve ongoing actions, it is tightly linked to behavioral goals and internal processes that those behavioral goals serve. Those intrinsic properties in egocentric vision have driven increasing interests on using egocentric vision in behavioral science and computer vision research.

While SETs are exclusively used in the screen viewing setting, HMETs can be used in much broader contexts with the wearer moving in the environment. Further, SETs track where a participant looks on a 2D screen while HMETs track where a participant looks in a 3D world. For HMET used in free moving contexts, because the head is the upper portion of the body and the eyes are on the front of the head, gaze direction is a product of eye movement in the head-centered coordinate system, head orientation and body pose. Two out of the three body movements, head orientation and body pose, are captured by egocentric vision itself. Further, because the precise timing relations between eye and head are also tightly coordinated (Pelz et al., 2001; Bambach et al., 2016b), the wearer's gaze is at the center of the egocentric view in a majority of the time. Taken together, this center bias in egocentric vision allows us to approximate visual attention even without explicit eye tracking as it is reasonable to assume that the center of egocentric view often reveals gaze direction of a wearer (Bambach et al., 2016b; Franchak et al., under review). Thus, this eye-head alignment makes it possible to track gaze direction from the egocentric view captured by a head-mounted FOV camera. The addition of the eye camera to track eye position, and a calibration process to map eye position to gaze direction in the FOV, lend further precision to measuring egocentric vision. In the following, we discuss two critical characteristics of head-mounted eye tracking research: precision in space and time, and ecological validity.

3.1. "In-the-moment" precision in space and time

Humans sample visual information from the outside world. Egocentric vision captures in-the-moment visual context when an observer uses a combination of eye, head, and body movements to shift attention and sample necessary information from the environment to support ongoing actions. Eye, head, and body movements are precisely coordinated to orient the visual system in space to collect task-relevant information (Land, 2004; McGee et al., in press). Because the eye tracking cameras move with the observer, HMETs capture the culmination of eye-in-head-in-body movements to orient the visual system. Moreover,

head-mounted eye tracking provides "just-in-time" temporal accuracy. People don't need to store all the information in the surrounding environment but instead a 'just-in-time' strategy is employed to acquire the specific information they need for movement programming exactly at the point it is required in the ongoing task (Ballard et al., 1995; Hayhoe et al., 2003). Noisy sensory data and short-term memory representations point to the need for the actor to update the information about the surrounding world and take that into the decision of where to look and what action to take as a consequence. Given limited cognitive resources and the fact that the information held in the working memory decays over time, the actor needs to carefully decide on where to shift attention to update certain parts of the visual environment to serve ongoing goals and activities. Because those behaviors serve sensorimotor and cognitive processes, moment-to-moment gaze data from HMET reveal how those processes unfold in real time.

3.2. Ecological validity

Franchak (2020b) identified three aspects of ecological validity provided by HMET research that are of theoretical consequence. Improving all three aspects of ecological validity can help researchers create measurements that better represent the construct of interest. First, HMET have the potential to present scenes with visual features that more closely approximate the statistics of everyday scenes. Screen-based stimuli lack the spatial regularities and feature distributions that are characteristic of what people see in day to day life. For example, photographs—even those of outdoor scenes—are biased by photographers to place a subject centered in view, resulting in features clustered towards the center (Tseng et al., 2009). However, recordings of participants navigating a variety of naturalistic environments find that visual features tend to be clustered towards the top of the view (Schumann et al., 2008), which could change how eye movements are used to explore the scene. Second, HMET allows for interactivity in a way that SET does not. Objects on screens cannot be picked up, and people in photographs cannot be conversational partners. Since observers look towards task-relevant locations (Ballard et al., 1995; Hayhoe et al., 2003; Tatler et al., 2013), the lack of interactive tasks will change how observers distribute visual attention. For example, Foulsham et al. (2011) found that observers in a SET task who watched videos from the perspective of a person walking along a campus path looked at the faces of other pedestrians more frequently compared to observers in a HMET who were actually walking on the path. In other words, the possibility of interaction in the real-world task decreased the likelihood that adults would look at strangers' faces. Third, the ability to track eye movements in a mobile observer means that HMET better captures how embodied factors influence eye gaze. Orienting the eyes to look depends on how the head and body are positioned; studying eye movements in a seated observer who can only gaze at a computer monitor removes how the head and body contribute to visual exploration (Bambach et al., 2016b). For example, infants crawling on the ground have a more restricted view of their surroundings (Kretch et al., 2014), which limits their ability to gaze at faces (Franchak et al., 2018) or distant

objects (Luo & Franchak, 2020) compared with when they are in an upright position.

The wearable headgear paired with smartphones or other lightweight devices to record the eye and FOV videos mean that HMETs can be used in applications that SETs' lack of mobility would not allow. Taking HMETs out of the lab and into the field means that HMETs contribute to the recent push for research that prioritizes ecological validity (Dahl, 2017; Pérez-Edgar et al., 2020). But ecological validity is not merely an issue of what type of eye tracker is used and whether the study was conducted in the lab or in the field. For example, using a HMET to study children's scanning of illusory contour shapes presented on a computer monitor is certainly not a "real-world" study (Nayar et al., 2015). We agree with Hessels et al. (2020), who suggested that researchers identify which aspects of the real world have theoretical impact in the research question (although we disagree with their claim that such reasons are often lacking). If there is no theoretical benefit to testing behavior in a more complex setting, experimental control should be prioritized. However, the ultimate goal of psychological research is to discover generalizable theories of human psychology that work not only in the lab but also in the real world. Toward this goal, we believe that HMET allows researchers to improve ecological validity on several dimensions that are important for understanding perception, motor control, and social interaction (for extended discussions, see Franchak, 2020a,b).

4. Two case studies of using head-mounted eye tracking in developmental science

In developmental science, researchers are taking advantage of novel head-mounted eye tracking methods to analyze visual experiences of early life (Smith et al., 2018; Bambach et al., 2018). This work has already resulted in important findings on how young children actively explore the world to create experiences and collect visual data to facilitate early learning of words and visual objects (Suanda et al., 2018). Here we provide an overview of two studies to illustrate that new findings gathered from HMETs have significant theoretical implications and practical utilities that justify the challenges faced in using the new technology.

4.1. Visual object learning and word learning

Recently, evidence derived from HMETs in child-parent free-flowing play provide a new perspective of early word learning. Yu & Smith (2012) showed that at the moment when a parent says a novel word, it is often the case that children have one object in view. Why? Because children have short arms. When children hold objects, the objects are close. Conveniently, parents also tend to name objects that children hold. Thus, the problem of determining the referent of a novel word is not solved by constraints in the head; it is solved by embodied constraints. The idea that active toddlers use their own actions to determine their dynamic visual experiences is very different from previous

theoretical views and video data captured from third-person cameras. HMETs capture children's visual experiences as a function of their body movements and actions, which is key to understanding early cognitive development.

Studying visual experiences from the view of the infant provides a unique window on early object recognition. In human visual science and in computer vision, object recognition is viewed as a hard problem. This inherent challenge is well known because the same 3D object can be projected as different 2D images onto the retina depending on how the object is posed and positioned, how the object is lit, and how it is seen against different visual backgrounds—all of all of which can vary moment to moment in the observer's egocentric view (Pinto et al., 2008). A second challenge is that objects belonging to the same object category (e.g. different types of mug for the "mug" category) may also vary in their appearance. Despite these challenges, toddlers can recognize roughly 300 object categories, and they can generalize a newly learned label to visual instances that they have never seen before. How do young children solve this problem?

Despite decades of research on this topic that have used experimental psychology approaches to examining how young children process well-controlled visual stimuli created by experimenters, recent studies built on the advances in egocentric vision and machine learning reveal that the answer may lie in measuring the visual input that young children actually perceive in everyday learning contexts. In many everyday contexts such as toy play, toddlers don't passively perceive visual input created by others. Instead, they actively use manual actions to create many different views of the same object (Soska et al., 2010). To test the hypothesis that size and similarity properties of the visual data created by toddler facilitate visual object recognition, computational simulation studies were conducted and the results showed that the same model trained based on the child data outperformed the model trained based on the parent data (Bambach et al., 2016a, 2018). Further, a follow-up study showed that infants who generated more variable visual object images through active manipulation are more advanced learners of early vocabulary in the real world (Slone et al., 2019). Thus, active viewing creates high-quality statistical data for visual object recognition. Taken together, those results suggest that the relevant data for early learning are not the statistics of the physical and social world but only the samples that emerge within the learners' own experiences, which are interactive and embodied. The sampling process is implemented through infants' actions that create their own data with unique properties and distributions (Smith et al., 2018). Critically, this new frontier in developmental science built on egocentric vision highlights the importance of examining the learning input that make contact with the child's sensory system (Yu et al., 2009).

4.2. Joint attention in parent-child social interaction

Our second case study uses the example of infant joint attention (JA), which is a widely studied topic in early social development. Infants' ability to follow the gaze of a social partner has a prominent place in the joint attention literature (Moore et al., 2014). Many experimental studies have been designed to

elicit and measure how well infants follow gaze to establish joint attention with their social partners. The evidence for gaze following in experimental contexts is unambiguous: Infants are able to follow a social partner's gaze direction to attend to a target object gazed by the partner. Usually, such studies are conducted in discrete experimental trials with clear and repeated signals from the partner to enhance the temporal availability of gaze signals. The assumption is that infants' abilities to follow the looking behavior of others—which are reliably demonstrated in the laboratory—scale up to real-world contexts.

Studies using dual HMETs (one worn by the infant and one worn by the parent), paired with synchronized frame-by-frame coding of manual actions, have objectively measured joint attention and the sensory-motor behaviors that underlie it (Yu & Smith, 2013, 2017b,a). By tracking the momentary gaze behavior and hand actions of each participant, we precisely determined just how often they looked at the same object at the same time, the visual behaviors that preceded joint attention, and manual behaviors that preceded and co-occurred with joint attention. We found that infants rarely looked at the parent's face during toy play and therefore gaze following was rarely employed by infants to establish joint attention bouts with their parents. Instead, both toddlers and parents lead and follow the other to create joint attention bouts. In other words, each member of the dyad sends behavioral signals to their partner and adjusts their own looking behavior in response to their partner's behavior. Further, both child-led and parent-led JA bouts may be created through the gaze following and hand following pathways. However, the frequency of these pathways differ as a function of who is leading and who is following, with hand following being much more likely by the toddler than the parent overall. In addition, there are developmental changes in this multi-pathway system evidenced as variations in strength among multiple routes (Yu & Smith, 2017b). A dual head-mounted eye tracking study of joint attention in freely moving infants and caregivers found high rates of JA despite infrequent looks by the infant to parents' faces (Franchak et al., 2018), providing converging evidence in a different task and motor context.

A followup study used HMETs to investigate the effects of children's prelingual hearing loss on how they achieve coordinated attention with their hearing parents during free-flowing object play (Chen et al., 2020). They found that 24- to 37-month-olds with hearing loss had similar overall gaze patterns (e.g., gaze length and proportion of face looking) as their normal hearing peers. In addition, children's hearing status did not affect how likely parents and children attended to the same object at the same time during play. However, when following parents' attention, children with hearing loss used both parents' gaze directions and hand actions as cues, whereas children with normal hearing mainly relied on parents' hand actions. The diversity of pathways leading to coordinated attention suggests the flexibility and robustness of developing systems in using multiple pathways to achieve the same functional end. These recent studies on parent-child social interaction suggest that coordinated visual attention between parents and toddlers is primarily a sensory-motor behavior. Skill in achieving coordinated visual attention in social settings—like skills in other sensory-motor

domains—emerges from multiple pathways to the same functional outcome (Yu & Smith, 2017b).

In additional to this theoretical implication, those findings also provide insights on interventions for atypically-developing populations. For example, previous eye tracking studies have shown that children with autism spectrum disorder (ASD) differ from that of typically developing (TD) children, with reduced face looking, reduced eye contact, and reduced joint attention. As such, these behaviors have become primary targets of research and intervention for children with ASD. However, using dual head-mounted eye tracking, a recent study examined these behaviors of toddlers with and without ASD during free-flowing toy play with a parent (Yurkovic et al., 2021). Contrary to expectations based on studies that use explicitly-cued attention prompts and screen-based stimuli, researchers found that children with ASD rarely looked at the parents' faces during toy play (though parents frequently look to their children's faces), and that mutual eye contact was an extremely rare phenomenon. Nevertheless, both ASD group and the age-matched TD group achieved similarly high levels of joint attention, suggesting that typical levels of joint attention in ASD arise during naturalistic play and are not primarily dependent on attention to faces and eyes.

5. Challenges and solutions for studying real-world eye gaze

In the preceding sections, we demonstrated how head-mounted eye tracking can provide new insights about important processes in developmental psychology by capturing in-the-moment data with high ecological validity. Now, we turn back to the issue of whether "...portable eye-tracking systems offer a cost-effective, easy to apply, and reliable measure of eye gaze and saccades in an ecological environment" (Shamay-Tsoory & Mendelsohn, 2019). As we discuss the key challenges in gathering high-quality HMET data, the key takeaway is that the research question, population, and procedures dictate how cost-effective, how easy to apply, and how reliable measurements will be with any type of eye tracker (SET or HMET). By identifying the degrees of freedom in how researchers might apply eye tracking methods that affect data quality, and by offering strategies for improving data quality, we hope that researchers can better decide whether HMET might be appropriate for their particular use case.

5.1. Is head-mounted eye tracking an accurate and precise way to measure eye gaze?

It depends on the application. Although SETs offer superior accuracy and reliability to track gaze on a 2D screen, HMETs can be sufficiently accurate and precise to track gaze in a 3D environment to answer a variety of questions. The needed accuracy and precision depend on the research question. At one extreme, researchers who study reading need to distinguish which letter in a word an observer is looking at, so using a high resolution screen eye tracker with the observer's head fixed in a chin rest may be necessary to minimize error. In other situations, such as when targets are large and sparsely distributed

in the environment, eye gaze data with larger errors may be acceptable. Eye-tracking *accuracy* refers to the size of the error between the true gaze location (where the observer is really looking) and the estimated gaze location (Nyström et al., 2013). *Precision* refers to the absence of noise in the eye tracking signal that leads to spatial variability over successive samples. Figure 2 illustrates accuracy and precision by showing an example calibration stimulus (the true location is the fingertip of the experimenter) surrounded by five estimated gaze points that are both accurate and precise (A), precise but inaccurate (B), accurate but imprecise (C), or both inaccurate and imprecise (D). Developmental researchers must be vigilant using any type of eye tracker because infant and toddler compliance has the potential to affect both aspects data quality.

5.1.1. Accuracy

Accuracy is measured by reporting the spatial error in degrees of visual angle, with smaller errors indicating better accuracy. The dotted line in Figure 2D shows the spatial offset between one gaze point and the true location. It is important to separate the manufacturer-reported accuracy of a system versus the accuracy of any single participant's recording. The former value, what the manufacturer lists in the specification sheet, expresses the ideal performance of the system. Generally, the manufacturer-specified accuracy of SETs is better (.02-0.5 degrees) compared with HMETs (0.5-2 degrees). In practice, the size of error for any single participant may exceed these ideal values. Many factors affect accuracy: how well the eye tracker detects the participant's eyes, how accurately the participant looked at the calibration points, and how much movement interfered with the recording.

How does one determine accuracy? After calibration, most SET systems can be configured to run a validation procedure. Validation means eliciting gaze to a set of targets on the screen and then measuring the spatial offset between each point and the estimated point of gaze (as in Figure 2D). It is customary to report this value in papers and to set an accuracy criterion for inclusion in the study. This is particularly important in developmental studies, because comparisons across age could be confounded by differences in eye tracking accuracy. Infants may not gaze exactly at the location of a calibration target, leading to worse accuracy (Frank et al., 2009; Wass et al., 2013). In a recent SET study of 6-month-olds to 12-year-olds and adults, we found a modest decrease in error with age of r = -.2 (Kadooka & Franchak, 2020). However, the difference in accuracy between infants ($M = 0.66^{o}$) and adults ($M = 0.43^{o}$) was only 0.2^{o}. Without measuring and reporting the accuracy, age-related differences in accuracy could be a confound lurking within a developmental study.

Because HMETs are not tied to a screen interface that can present gaze targets and automatically calculate the spatial error, users may need to get a bit more creative to determine the accuracy if the system they are using does not provide a validation procedure. One method is to use a standard-sized target to assess accuracy. Infant HMET studies often calibrate the eye tracker by showing a board with small windows and eliciting looks to each window by showing an attractive toy (Kretch et al., 2014; Franchak et al., 2018), as seen

in Figure 1. If the calibration board is presented at the same distance, it is straightforward to calculate the visual angle of each window to use that as a standard. For example, the windows in Figure 1 are approximately 4° wide. If the estimated point of gaze falls within the window during a set of validation check trials, the researcher can infer that the spatial error is within 2° (a version of this method was used in Franchak et al., 2011). Although this method does not provide an absolute error value for each participant, it can help assure that data collected are within an acceptable range.

With a bit more effort, researchers can approximate the error with any HMET system. By eliciting looks to validation targets, the resulting gaze overlay video can be used to measure the offset between the gaze estimate and target. Based on the field of view (in degrees) of the eye tracker camera and the resolution (in pixels) of the video file, the number of pixels of spatial offset between the target and the point of gaze can be converted to degrees of visual angle. Figure 2D illustrates the spatial offset for one gaze point and shows how the pixel-to-degree ratio can be used to approximate error. Using this method, we reported accuracy of 1.55° in infants and 0.85° in caregivers who both wore eye trackers during a laboratory play session (Franchak et al., 2018). However, because eye tracking field of view cameras often use wide-angle lenses that distort space, converting pixels to degrees is only an approximation. If more precise measures of error are needed, researchers can employ tools, such as the Matlab Camera Calibration Toolbox or OpenCV libraries, that can measure and model lens distortion and then map gaze data into an undistorted space (e.g., McGee et al., in press; Tomasi et al., 2016).

Movement can make it more difficult for eye tracking cameras to detect eye position. SET systems that allow for head movement (without a chin rest) can have errors introduced as a participant moves in depth (Niehorster et al., 2018), especially in infants who may wiggle from front to back. A unique threat to HMET accuracy is headgear "slippage" (Niehorster et al., 2020). Once the eye tracker is calibrated, estimates of gaze direction are based on the location of the pupil and corneal reflection in the eye camera view. Systematically changing that location by moving the location of the headgear will introduce error. For this reason, we recommend that researchers collect calibration and validation data at the beginning and end of the session to ensure that accuracy has not worsened due to headgear movement. In longer sessions, it may be advisable to include multiple, brief (30-s) calibration checks. If a deviation in accuracy is detected, the researcher can often find the moment of slippage by scanning through the eye camera video to find when it moved. If multiple sets of calibration data are collected over time, the session can be salvaged by using a different calibration after the headgear moved.

5.1.2. Precision

Sources of noise include difficulty in detecting the pupil or corneal reflection from one sample to the next, which can result in the gaze estimate "jumping" from place to place. Although noisy data can be accurate on average, noise in eye tracking data makes it particularly challenging to understand the temporal

aspects of eye gaze (Wass et al., 2014; Nyström et al., 2013). A common practice in eye tracking research is to parse eye movements into fixations (moments during which the eyes remain stable on a target) and saccades (high velocity shifts of eye position from one location to another). Since noisy data erroneously indicate that the eye is moving when in fact it may be relatively stable, noise can lead to underestimating the lengths of fixations (e.g., a long fixation may be interrupted by a jump in eye position that is incorrectly parsed as a saccade). Poor data precision may also impact coding whether participants looked at one area of interest or another (Wass et al., 2014).

Even in SET studies, infant data precision may be worse compared with older participants, since infants do not sit still and have eye features that may are more difficult to track (Wass et al., 2014). However, this may not always be the case: Using Wass et al.'s precision metric to describe the sample-to-sample noise in SET data, we failed to find a difference across our sample of 6-month-olds to 12-year-olds and adults (Kadooka & Franchak, 2020). Little empirical work has characterized data reliability in head-mounted eye trackers. Niehorster et al. (2018) found, unsurprisingly, that headgear slippage introduced noise from the beginning to the end of a session. However, our strategy of collecting multiple calibrations/validations would help mitigate this issue by recalibrating the eye tracker after the headgear has moved.

Some HMETs, such as the Positive Science system, require that the researcher manually orients the eye camera to obtain the best possible image of the participant's eye. In our experience, this orienting step is crucial for ensuring low-noise data. If the eye is large in the eye camera view, centered so that it does not overlap the edge of the camera's view, and brightly illuminated, the computer vision algorithms do well at detecting the pupil (Figure 3A-B). Failure to adjust the camera to get a good eye image will result in noisy data, because the algorithm struggles from one sample to the next to detect the pupil. Figure 3C-D show poor eye camera images that will result in noisy data as the detection algorithm struggles to detect the location of the pupil. In addition, periodically checking the eye camera image during a study can help ensure that the quality standard is maintained throughout a session.

Without ensuring that the data are precise (or without ensuring that precision is consistent across participants/age groups/conditions), researchers should be cautious about applying event-based metrics to eye tracking data. As we will discuss in the next section, many research questions do not rely on categorizing eye gaze data into events such as fixations and saccades. Researchers may instead opt to report aggregate measures of looking behavior, such as the proportion of eye gaze samples directed to one area of interest versus another, which could be more robust to noise compared with claims about average fixation duration or fixation sequence. However, if those measures are of primary interest, researchers should set inclusion criteria based on empirical measurements of eye tracking precision. In the past we have informally excluded participants if the eye image seemed poor and resulted in a noisy track; however, a more objective option would be to calculate precision using Wass et al.'s precision metric (after converting scene camera pixels to degrees) and setting a threshold for what level

of precision counts as usable data.

5.2. Is head-mounted eye tracking easy to apply?

Sometimes. HMETs can be easy to apply in some situations and populations, and difficult to apply in others. In their critique, Hessels et al. (2020) wrote, "It is true that for some wearable eye trackers, it is easy to put one on and start a measurement. But 'easy to apply' suggests that the wearable eye tracker is easy to use for drawing conclusions on perception, cognition, and so forth. We think that this is not the case because (a) clear theoretical reasons for conducting wearable eye-tracking research are often lacking, (b) important concepts in eye-tracking research are often not or ill-defined, and (c) signal processing and analysing where someone looked in the world is problematic." For us, as developmental researchers, this statement seems backwards. Placing eye-tracking equipment on infants can sometimes be the biggest challenge in the study. Because we have described techniques for placing the equipment (and keeping it) on infants elsewhere (Franchak et al., 2011; Franchak, 2017; Slone et al., 2018), we do not to discuss that issue here. Earlier in this paper, we discussed the theoretical issues for conducting head-mounted eye tracking research (criticism 'a'). Here, we address the second and third criticisms about ease of use (criticisms 'b' and 'c').

5.2.1. Defining fixations and saccades

The "important concepts in eye-tracking research" that Hessels et al. referred to are categorizing eye gaze as fixations and saccades. Humans can make a variety of eye movement types (Land & Tatler, 2009): Besides fixations and saccades (which we previously defined), the eyes can smoothly pursue moving targets and can counter-rotate to compensate for head movement (the vestibular-ocular reflex). Moving objects and moving observers make it impossible for HMETs to automatically define fixations and saccades from the position of the eyes within the head in a way that is commensurate with SET systems. Assuming that the concept of a fixation is in fact well-defined and used consistently in screen-based eye tracking (Hessels et al., 2018, suggests it may not), does the problem of defining fixations disadvantage HMET research?

In some instances, measuring fixations that can be directly compared to SET work may be vital. Researchers who study eye movement control need to categorize eye movement types, and researchers who study executive function may need to measure fixation durations as an index of attentional control (e.g., Wass & Smith, 2014). Yet, in many other use cases the concept of a fixation is not critical. For characterizing joint attention and face looking as reviewed in the previous section, it may be sufficient to simply measure how frequently an infant looks at the caregiver's face in a trial or a session (Franchak et al., 2018; Yu & Smith, 2013). To understand how frequently infants look at objects when hearing an object name, measuring the proportion of frames in which gaze lands on each of the available objects in view would be sufficient to answer the research question. Even though infants may generate multiple fixations

toward the same object, the number of fixations on the that object may not be the most meaningful measure for building a word-object mapping. Likewise, the definition of a fixation is not important for determining how frequently infants and children look at obstacles before traversing them (Franchak et al., 2011; Kretch & Adolph, 2017; Franchak & Adolph, 2010). In all of these cases, researchers may place a minimum duration of what counts as looking (e.g, 2 or 3 successive video frames at 30 Hz) to reduce the odds of capturing incidental gaze (such as when the eyes momentarily fall on a target during a saccade from one location to another). Regardless, in all of these examples (and many others), the research goal is to measure what the observer saw to draw conclusions about cognitive, motor, or social development, not to detect whether the eye was in one state versus another. With no or minimal head movement (by using a chin rest), it is certainly easier to detect fixations and saccades using SET during screen viewing. However, because the eyes and the head naturally move together to serve attention switching in a 3D world, measuring eye gaze amidst head and body movement enhances, rather than diminishes, our understanding of how infants and children visually explore in naturalistic contexts.

5.2.2. Signal processing and analysis

Another criticism that Hessels et al. levy is that HMETs make it difficult to interpret exactly where people looked, since the eye tracker measures the point of gaze within a head-centered field of view camera (in two dimensions), not where gaze falls in the three-dimensional world. They note that the problem of determining what people looked at becomes more difficult as the environment becomes more complex. Whereas it may be easier in a controlled environment to decide that an infant is looking at one of three brightly colored objects (e.g, Yu & Smith, 2013), it is more difficult to categorize what object a person is looking at as they walk along a city street ('t Hart & Einhauser, 2012) or college campus (Foulsham et al., 2011; McGee et al., in press) when there are many potential targets of interest of varying shapes, sizes, and distances. This is certainly true, however, it is not a problem unique to HMET research, but is a result of the complexity of real-world visual scenes. The trend in SET research has been to move towards measuring how people view so-called "natural" images or videos that contain densely clustered targets that are more like everyday scenes. Even with a highly accurate and highly precise SET, there is always some spatial error and noise. As the distance between areas of interest (AOIs) decreases, there will be a point at which eye tracker error and/or noise will result in mis-classification of looking. This is true for a photograph or video on an SET or a real-world scene viewed with a HMET. Figure 4A shows an example of rectangular AOIs drawn to classify looks to the experimenter's eyes versus mouth, with an example gaze point falling in the mouth region. Error bars to the right of the gaze point show how vertical errors of 6° versus 4° versus 2° would lead to varying confidence in whether that gaze estimate was truly a look to the mouth versus the eyes.

If researchers are taking an AOI approach (which, as we will discuss below, is not the only option for analyzing data), there are different options for how to implement AOI analyses. The first is to draw areas of interest on the stimulus

or gaze overlay video, and calculate whether the point of gaze lands within the boundaries. The example image in Figure 4A shows rectangular AOIs drawn on a video frame. Most SET systems are bundled with analysis software that allow researchers to draw areas of interest on the stimulus (typically ovals or rectangles, but sometimes free-form contours). Although this can be done quickly if the stimulus is a photograph, it can be time-consuming to do so for every frame of a video stimulus as targets move within the image. We have made free software available to allow researchers to make similar annotations based on the gaze recording videos from HMETs (Franchak, 2019), but this is even more time-consuming for HMET studies because it must be done for each individual participant. Because eye gaze estimates can be noisy and will always contain some amount of error, researchers often choose to make AOIs larger than the target so that "borderline" looks get counted (as in Figure 4A). Regardless of the type of eye tracking system used, researchers should report the size of the AOIs and the degree to which they may be padded, since these decisions can make the analyses count AOI looks more liberally versus more conservatively. For more details about reporting results from AOI analyses, see Oakes (2010).

A second way to score AOI looking is through manual coding of the gaze overlay video. A coder goes through the frames of a video at reduced speed and annotates a look to a particular target when the gaze cursor lands on the target. For example, a definition of face looking could be "count a face look as any time the gaze crosshair lands on any skin region of a face (excluding the hair, neck, and ears) for at least 2 successive video frames". Because such coding is performed on the gaze overlay video, it must be repeated for every participant's session, which can be time-consuming. However, for HMET studies either type of analysis (drawing AOIs versus coding looks) would need to be performed for every session, and making a judgment about which of a set of potential targets the gaze crosshair falls on is often quicker compared to drawing multiple AOIs on each video frame. In addition, a well-trained coder can visually segment the borders of an object that is not circular or rectangular with ease, but drawing a precise AOI around a complex object can be laborious. Padding to account for error in eye gaze estimates can be accomplished by creating a gaze cursor that visualizes exactly how close to a target the gaze cursor needs to be to count as a look. Figure 4B shows a gaze cursor with 2° (solid circle) and 4° (dotted circle) radii. Defining a look based on the circle overlapping any code-able region of a target accomplishes the goal of padding, and does so more precisely compared with a rectangular or circular AOI that inconsistently increases the size of the target. As with any type of behavioral coding, computerized coding software such as Datavyu (www.datavyu.org) or ELAN (https://archive.mpi.nl/tla/elan) can make the process more efficient. HMET manufacturers may offer specialized annotation software as part of the system (sometimes at an additional cost). Some researchers have shared open-source coding software specific to eye tracking analyses (Benjamins et al., 2018). Finally, reliability coding can help ensure that codes are accurate and reproducible, and blinding coders to experimental conditions can reduce bias.

Regardless of the method, AOI analyses should be designed with the ex-

pected accuracy/precision in mind. Pilot testing with the HMET system can help researchers determine what level of accuracy and precision to expect from their population. Recording videos ahead of time from the HMET system in the testing environment with objects of interest allows researchers to see how objects' size, distance, and proximity to other objects makes them appear in the FOV camera image. Note that a participant does not even need to wear the eye tracker to gain these insights—a live view of the FOV camera in the testing environment can let the experimenter move targets in different locations to see the visual consequences from the participant's point of view. In situations where the experimenter can control the size and locations of objects, making objects large in the field of view (either by increasing the size or decreasing the distance of objects) and spacing them far apart can make later coding easier. Of course, this may not be possible in applications that seek to test participants in completely unconstrained environments. In those situations, coding pilot data can help researchers determine which objects can be reliably discriminated. For example, coding looks to the eyes versus mouth may be reliable if the face is always close to the observer (making the target regions large) and if the data collected are sufficiently accurate and precise. Coding eyes versus mouth looks would be feasible in the example in Figure 4A when the error is 2° but not 6°; however, if the experimenter moves farther away from the participant and their face becomes visually smaller, coding eyes versus mouth might not be reliable even with 2° accuracy. Pilot testing and attempting to code the desired regions can quickly make it apparent whether such coding can be done reliably. Researchers may need to settle on more basic coding schemes (e.g., coding looks to faces rather than looks to eyes versus mouth) depending on the parameters of their testing environment and data quality.

5.3. Is head-mounted eye tracking cost effective?

Maybe! We hope that theme by now is clear—it depends on how method is applied. Hessels et al.'s concern is that even though the price tag to acquire a HMET system might be cost effective, the labor of manual coding done by staff researchers could be a hidden cost. As we will review in this section, researchers have a lot of options available to analyze HMET data. Some options are more time-consuming than others. Manual coding of any sort of data can be slow, but this is already a common challenge for developmental psychologists. Transcribing utterances or counting every footstep a baby takes time to do accurately and reliably. For eye movement data—in which eye position might move 2-3 times per second—coding every single place a person looks for a few minutes can be a large undertaking. Furthermore, learning how to develop codes, use coding software, and implement quality control procedures (e.g., reliability coding) is an investment for a lab that is not experienced in behavioral coding.

First of all, researchers should ask whether AOI coding is necessary to answer the research question. For some applications, it may be sufficient to describe the pattern of eye movements based on the eye gaze coordinates without scoring what exact things were looked at. Some researchers used HMETs to record eye movements in adults walking in outdoor environments to describe the horizontal

versus vertical distribution of eye-in-head movements ('t Hart & Einhauser, 2012; Foulsham et al., 2011; Tomasi et al., 2016). The horizontal spread of eye movements depends on the task that observers are performing—spread increases to visually explore around the body when searching for targets compared to simply walking along a path (McGee et al., in press). Bambach et al. (2013) used eye position data to show that both infants and adults tend to keep their eyes centered within the head during a play session, showing that head movements are frequently used to shift gaze between targets. In all of these examples, the horizontal and vertical timeseries of eye position within the head-centered field of view were sufficient to answer questions without the need for AOIs or manual coding of the gaze overlay video.

If AOI analyses are required, researchers should determine which targets need to be coded and how frequently they need to be coded to answer the question. Although it may sound attractive to have a full sequence of every object that was every gazed at, rarely is this necessary. Instead, researchers should only code what they need to code based on their hypotheses. For example, in two investigations we reported the proportion of time that infants looked at caregivers' faces was 4% and 11% (Franchak et al., 2018; Yu & Smith, 2013). Thus, if the main variable of interest is face looking, coders can efficiently go through much of the session and only annotate the relatively rare moments that infants looked at caregivers' faces. In a study that asked whether participants fixated obstacles in advance of traversing them, frame-by-frame coding of the entire video was avoided by first reducing the data by identifying all of the obstacle events (Franchak & Adolph, 2010). After finding the times when participants stepped up, down, and over obstacles, a second pass of coding identified whether the obstacle was looked at in the preceding 5 s. Depending on the length of the recording, it may not be necessary to code every frame. Jayaraman et al. (2015) recorded infants' egocentric perspective using head-mounted cameras to determine how often faces and hands were in view. Rather than coding every video frame of a corpus exceeding 100 hours, they reduced the data to code video frames every 5 s to increase efficiency. By coding only what is necessary, researchers can minimize the time needed to answer their question to make HMET research more cost effective.

Finally, there have been promising developments over the past decade to show that manual coding might eventually be replaced by automatic computer vision coding for some applications. The emerging field of egocentric computer vision aims to automatically understand the images and video captured from first-person cameras. Besides the scientific applications mentioned above, wearable cameras will likely create a variety of consumer applications, especially as low-cost devices (like Google Glass, Narrative Clip, etc.) become mainstream. For instance, first-person cameras are now worn by police officers to document their interactions with the public, have been tested as a means for helping people with dementia record and recall forgotten memories, and can potentially assist people with visual impairments to navigate physical and social situations through real-time feedback provided by computer vision analysis. All of these applications collect and record such vast amounts of image data that computer

vision is needed to automatically organize and analyze images, potentially in real time. Researchers have explored egocentric vision for activity and object recognition, temporal segmentation, scene understanding, interaction analysis, hand detection, gaze estimation, gaze analysis, visual saliency, social saliency, and motion capture. Many of those algorithms—including pre-trained deep learning models—are open source software that are free for academic research and can be used in an off-the-shelf way. All this make it feasible for automatic annotation of attended objects, faces and hands from FOV videos. For example, we developed a computer vision annotation algorithm to automatically detect and identify infants' versus parents' hands from HMET videos (Bambach et al., 2014, 2015). Others have used off-the-shelf algorithms, such as the OpenPose library to recognize hands and faces, and successfully replicated findings that were conducted by manual coding (Long et al., 2021). We have also successfully used YOLO and Resnet to automatically detect visual objects in the FOV camera view (Bambach et al., 2018). Those computational tools not only obviate the time needed for manual coding, but also provide more objective and precise measures of visual information in egocentric vision compared with human coding. As the developmental research field moves toward "big data" approaches, those computer vision tools become an indispensable component of a whole system of using HMET in the real world.

5.4. Summary of recommendations

Head-mounted eye tracking research is challenging, but not so prohibitively difficult that researchers should avoid applying it. Sometimes, difficult things are worth doing. Researchers have always been at the front end of paving new ways to advance science. Going beyond the screen, HMET has allowed us to answer new questions that would be impossible to address using SETs or any other extant method. Moreover, there are solutions available to mitigate many of the difficulties faced by using HMET with infant and child populations. We end the paper with a summary of our recommendations of best practices, with the hope that researchers can use these guidelines to decide if and how to apply HMET to their own research questions.

1. Start with a clearly defined research question. As we emphasized throughout this chapter, the demands on data quality, signal processing, manual coding, etc. all stem from the particular research question. There is no need for researchers to expend resources to recruit additional subjects to replace those with good enough (but not stellar) data quality, nor to engage in laborious data coding if those improvements will not affect the outcome of the study. Likewise, detailing the needs of the research question might rule out HMET altogether. For example, a study that needed to identify which letter of different words on a billboard were fixated by an observer in a crowded train station might not be feasible given the current limits of HMET.

2. Assess which aspects of ecological validity matter for measuring the constructs of interest. Set up the study to prioritize realistic visual features, interactivity,

and/or embodiment as needed, but control other aspects of the study to simplify later coding and analysis. If a moving observer is not critical to the research question, having the observer sit in place can help constrain the geometry of the scene to facilitate AOI coding and can reduce headgear slippage. Explicitly reporting which aspects of ecological validity are central to the question can help other researchers decide whether those need to be accounted for in related work.

3. Decide prospectively what aspects of gaze behavior need to be measured. Is area of interest coding necessary to answer the research question? Can the question be asked with a simpler coding scheme or by using raw gaze coordinates? Do fixations need to be defined, or would aggregate measures of looking time be sufficient? Each of these decisions place different demands on the data quality required and the amount of labor needed to analyze the data. Considering these options in advance can help the researcher plan a more direct route to answering the research question.

4. Measure the accuracy and precision of eye gaze data. Reporting the accuracy and precision from validation checks is a best practice in eye tracking research (Oakes, 2010; Wass et al., 2014). This allows the researcher, reviewers, and readers to determine whether the data are of sufficient quality for event detection and/or AOI coding. Ensuring that data are of comparable quality between different subject groups/conditions can rule out a potentially major confound. Developmental researchers studying infants and children should pilot test to estimate the accuracy and precision they can expect with different age groups using the eye tracking system, as it will rarely approach the levels listed in the manufacturer's specifications. Matching participants or setting inclusion criteria based on quality metrics may be necessary. Finally, researchers can use built-in software, rule of thumb metrics, or convert spatial offsets into degrees using camera parameters to accomplish these goals with any type of eye tracking system.

5. Plan AOI coding with respect to eye tracking data quality. Setting up the testing environment and measuring the degrees of visual angle subtended by different targets at their expected locations/distances can reveal what level of eye tracking accuracy is needed to confidently score looks to different AOIs. Researchers should report how they defined AOIs (e.g., drawing regions over an image, manual coding based on the gaze cursor, etc.), whether and how they were padded, and, if in a controlled situation, the size and location of AOIs. Knowing that AOIs were large and spread apart relative to the eye tracking data precision will increase the reader's confidence in the results.

6. Prioritize efficient rather than exhaustive data coding. It is rarely necessary to code every gaze event to every possible location in the testing environment to answer the question of interest. Coding only what is necessary for the hypothesized result saves time. Using computerized coding software can make coding

more efficient, and allows researchers to easily return to a data source to execute additional coding in the future, if warranted. For example, if the research question only requires coding face-looking events, a follow-up analysis could go back to those events to categorize them as looks to eye versus mouth regions (assuming the data are sufficiently precise). Exploring options for computer vision analyses can facilitate automatic data coding, particularly for categories that have open source, off-the-shelf algorithms available to apply.

References

Adolph, K. E., Cole, W. G., Komati, M., Garciaguirre, J. S., Badaly, D., Lingeman, J. M., Chan, G., & Sotsky, R. B. (2012). How do you learn to walk? Thousands of steps and dozens of falls per day. *Psychological Science*, *23*, 1387–1394. doi:10.1177/0956797612446346.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*, 66–80. doi:10.1162/jocn.1995.7.1.66.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2016a). Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2018). Toddler-inspired visual object learning. In *Advances in Neural Information Processing Systems*. volume 31.

Bambach, S., Crandall, D. J., & Yu, C. (2013). Understanding embodied visual attention in child-parent interaction. In *Proceedings of the IEEE Conference on Development and Learning*.

Bambach, S., Franchak, J. M., Crandall, D. J., & Yu, C. (2014). Detecting hands in children's egocentric views to understand embodied attention during social interaction. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1949–1957).

Bambach, S., Smith, L. B., Crandall, D. J., & Yu, C. (2016b). Objects in the center: How the infant's body constrains infant scenes. In *IEEE 6th Joint International Conference on Development and Learning and Epigenetic Robotics*.

Benjamins, J. S., Hessels, R. S., & Hooge, I. T. (2018). Gazecode: Open-source software for manual mapping of mobile eye-tracking data. In *Proceedings of*

the 2018 ACM Symposium on Eye Tracking Research & Applications (pp. 1–4).

Chen, C.-h., Castellanos, I., Yu, C., & Houston, D. M. (2020). What leads to coordinated attention in parent–toddler interactions? Children's hearing status matters. *Developmental Science*, *23*, e12919.

Dahl, A. (2017). Ecological commitments: Why developmental science needs naturalistic methods. *Child Development Perspectives*, *11*, 79–84.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*, 1920–1931. doi:10.1016/j.visres.2011.07.002.

Franchak, J. M. (2017). Using head-mounted eye tracking to study development. In B. Hopkins, E. Geangu, & S. Linkenauger (Eds.), *The Cambridge Encyclopedia of Child Development* (pp. 113–116). Cambridge: Cambridge University Press. (2nd ed.).

Franchak, J. M. (2019). Dynamic ROI Coder for Matlab. doi:10.17605/OSF.IO/G3UT5.

Franchak, J. M. (2020a). Looking with the eyes and head. In J. B. Wagman, & J. J. C. Blau (Eds.), *Perception as Information Detection: Reflections on Gibson's Ecological Approach to Visual Perception* (pp. 205–221). Routledge.

Franchak, J. M. (2020b). Visual exploratory behavior and its development. In K. D. Federmeier, & E. R. Schotter (Eds.), *Psychology of Learning and Motivation* (pp. 59–94). Elsevier. doi:10.1016/bs.plm.2020.07.001.

Franchak, J. M., & Adolph, K. E. (2010). Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults. *Vision Research*, *50*, 2766–2774. doi:10.1016/j.visres.2010.09.024.

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2018). See and be seen: Infant–caregiver social looking during locomotor free play. *Developmental Science*, *21*, e12626. doi:10.1111/desc.12626f.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development*, *82*, 1738–1750. doi:10.1111/j.1467-8624.2011.01670.x.

Franchak, J. M., Smith, L. B., & Yu, C. (under review). Developmental changes in how head orientation structures infants' visual experiences. Manuscript submitted for publication.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*, 160–170. doi:10.1016/j.cognition.2008.11.010.

't Hart, B. M., & Einhauser, W. (2012). Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, *223*, 233–249.

Hayhoe, M. M. (2000). Vision using routines: A functional account of vision. *Visual Cognition*, *7*, 43–64.

Hayhoe, M. M., Shrivastava, A., Mruczek, R. E. B., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*, 49–63.

Hessels, R. S., Niehorster, D. C., Holleman, G. A., Benjamins, J. S., & Hooge, I. T. C. (2020). Wearable technology for "real-world research": Realistic or not? *Perception*, *49*, 611–615. doi:10.1177/0301006620928324.

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*, 180502.

Jayaraman, W., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PloS One*, *10*, e0123780. doi:10.1371/journal.pone.0123780.

Kadooka, K., & Franchak, J. M. (2020). Developmental changes in infants' and children's attention to faces and salient regions vary across and within video stimuli. *Developmental Psychology*, *56*, 2065–2079.

Kretch, K. S., & Adolph, K. E. (2017). The organization of exploratory behaviors in infant locomotor planning. *Developmental Science*, *20*, e12421.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, *85*, 1503–1518. doi:10.1111/cdev.12206.

Land, M., & Tatler, B. (2009). *Looking and acting: Vision and eye movements in natural behaviour*. Oxford University Press.

Land, M. F. (2004). The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations. *Experimental Brain Research*, *159*, 151–160. doi:10.1007/s00221-004-1951-9.

Land, M. F., & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B*, *352*, 1231–1239.

Land, M. F., & Hayhoe, M. M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565. doi:10.1016/S0042-6989(01)00102-X.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*, 742–744.

Long, B., Sanchez, A., Kraus, A. M., Agrawal, K., & Frank, M. C. (2021). Automated detections reveal the social information in the changing infant view. doi:10.31234/osf.io/cmj65.

Luo, C., & Franchak, J. M. (2020). Head and body structure infants' visual experiences during mobile, naturalistic play. *PLoS ONE*, *15*, e0242009.

Matthis, J. S., Yates, J. L., & Hayhoe, M. M. (2018). Gaze and the control of foot placement when walking in natural terrain. *Current Biology*, *28*, 1224–1233.

McGee, B., Blanch, G., & Franchak, J. M. (in press). Adapting the coordination of eyes and head to differences in task and environment during fully-mobile visual exploration. *PLoS ONE*, .

McGraw, M. B. (1935). *Growth: A study of Johnny and Jimmy*. New York, NY: Appleton-Century Crofts.

Moore, C., Dunham, P. J., & Dunham, P. (2014). *Joint attention: Its origins and role in development*. Psychology Press.

Nayar, K., Franchak, J. M., Adolph, K. E., & Kiorpes, L. (2015). From local to global processing: The development of illusory contour perception. *Journal of Experimental Child Psychology*, *131*, 38–55.

Niehorster, D. C., Cornelissen, T. H., Holmqvist, K., Hooge, I. T., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, *50*, 213–227.

Niehorster, D. C., Santini, T., Hessels, R. S., Hooge, I. T., Kasneci, E., & Nyström, M. (2020). The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, *52*, 1140–1160.

Nyström, M., Andersson, R., Holmqvist, K., & Van De Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, *45*, 272–288.

Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy*, *15*, 1–5.

Pelz, J. B., Hayhoe, M. M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, *139*, 266–277. doi:10.1007/s002210100745.

Pérez-Edgar, K., MacNeill, L. A., & Fu, X. (2020). Navigating through the experienced environment: Insights from mobile eye tracking. *Current Directions in Psychological Science*, *29*, 286–292.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*, e27.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & Koenig, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision*, *8*, 12–12.

Shamay-Tsoory, S. G., & Mendelsohn, A. (2019). Real-life neuroscience: an ecological approach to brain and behavior research. *Perspectives on Psychological Science*, *14*, 841–859.

Slone, L. K., Abney, D. H., Borjon, J. I., Chen, C.-h., Franchak, J. M., Pearcy, D., Suarez-Rivera, C., Xu, T. L., Zhang, Y., Smith, L. B. et al. (2018). Gaze in action: Head-mounted eye tracking of children's dynamic visual attention during naturalistic behavior. *Journal of Visualized Experiments*, *141*, e58496.

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science*, *22*, e12816.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, *22*, 325–336.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: Motor skill acquisition facilitates three-dimensional object completion. *Developmental Psychology*, *46*, 129–138. doi:10.1037/a0014618.

Suanda, S. H., Barnhart, M., Smith, L. B., & Yu, C. (2018). The signal in the noise: The visual ecology of parents' object naming. *Infancy*, *24*, 455–476.

Tatler, B. W., Hirose, Y., Finnegan, S. K., Pievilainen, R., Kirtley, C., & Kennedy, A. (2013). Priorities for selection and representation in natural tasks. *Philosophical Transactions of the Royal Society B*, *368*, 20130066.

Tomasi, M., Pundlik, S., Bowers, A. R., Peli, E., & Luo, G. (2016). Mobile gaze tracking system for outdoor walking behavioral studies. *Journal of Vision*, *16*, 27–27.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*, 1–16. doi:10.1167/12.13.3.

Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, *19*, 427–460. doi:10.1111/infa.12055.

Wass, S. V., & Smith, T. J. (2014). Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy*, *19*, 352–384. doi:10.1111/infa.12049.

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*, 229–250. doi:10.3758/s13428-012-0245-6.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*, 244–262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE*, *8*, e79659. doi:10.1371/journal.pone.0079659.

Yu, C., & Smith, L. B. (2017a). Hand-eye coordination predicts joint attention. *Child Development*, *88*, 2060–2078.

Yu, C., & Smith, L. B. (2017b). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*, *41*, 5–31. doi:10.1111/cogs.12366.

Yu, C., Smith, L. B., Shen, H., Pereira, A. F., & Smith, T. G. (2009). Active information selection: Visual attention through the hands. In *IEEE Transactions on Autonomous Mental Development* (pp. 141–151). volume 1.

Yurkovic, J. R., Lisandrelli, G., Shaffer, R. C., Dominick, K. C., Pedapati, E. V., Erickson, C. A., Kennedy, D. P., & Yu, C. (2021). Using head-mounted eye tracking to examine visual and manual exploration during naturalistic toy play in children with and without autism spectrum disorder. *Scientific Reports*, *11*, 3578. doi:10.1038/s41598-021-81102-0.

Figure 1: A) Infant viewing a video stimulus on a screen eye tracker (SET) while seated in a high chair. B) Left image shows an experimenter placing a head-mounted eye tracker (HMET) on an infant participant; arrows indicate the location of the inward-facing eye camera and the outward-facing field of view (FOV) camera. Right image shows the gaze overlay from a participant during calibration. The FOV camera records the infants' head-centered view, and a circular gaze cursor is superimposed on the video to show the infant looking at the experimenter presenting a toy in a window. Concentric circles around the point of gaze show different error radii in degrees.

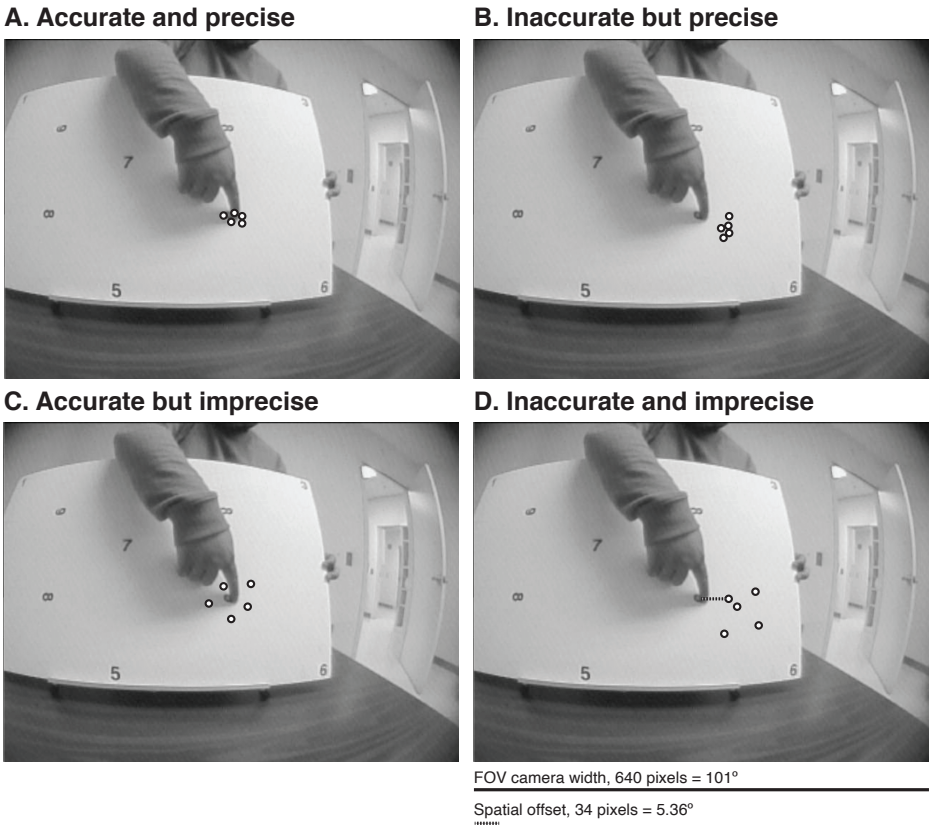

Figure 2: Example validation stimulus from a HMET FOV camera showing 5 gaze estimates (black and white circles) that should be directed to the fingertip of the experimenter. The four panels illustrate: (A) data that are accurate and precise, (B) data that are inaccurate but precise, (C) data that accurate but imprecise, and (D) data that are inaccurate and imprecise. The graphics below (D) show how the field of view in degrees of the FOV camera can be used to approximately convert a spatial offset, measured in pixels, into degrees of visual angle.

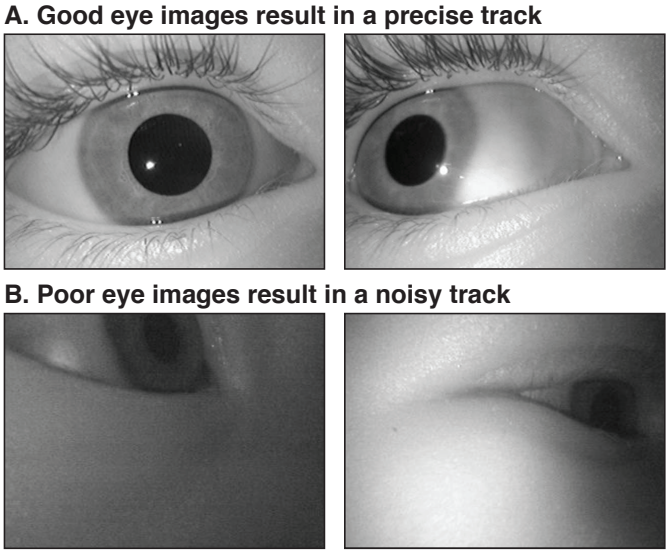

Figure 3: A) and B) show high quality eye images with the eye large and well-centered in view, with bright illumination that makes the pupil stand out from the image. Such images make it easy for a computer vision algorithm to detect the pupil, decreasing the noise in the eye gaze estimate. C) and D) show eye images with the camera too far from the eye, centered poorly, and lacking good illumination. Noise will result from the pupil going out of view, as well as detection problems resulting from the low-contrast of the pupil compared with the rest of the image.

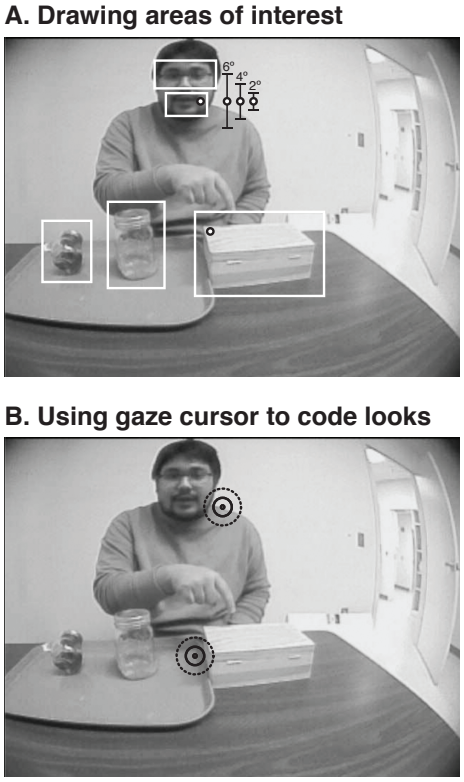

Figure 4: A) Areas of interest (AOIs) drawn on a HMET gaze overlay video frame. White rectangles show the boundaries for scoring a look to three objects or to the experimenter's eyes and mouth. Two gaze points (black and white circles) illustrate looks to the mouth and to the box because the point of gaze falls within the AOI rectangles. To the right of the mouth AOI, error bars of varying degrees show how confidence in scoring a look to an AOI decreases when eye tracking accuracy is worse. B) Using a gaze cursor to score looks to AOIs. Instead of drawing regions in the video frame, a coder manually identifies moments when the gaze cursor overlaps with a specified area, such as the face or the box. By defining looks based on the point of gaze plus a circular "error region", researchers can choose to code looking more liberally based on the error of the system. The dotted, outer circle indicates a 4° error radius and the solid, inner circle indicates a 2° radius. In this example, a look to the face would be scored using the 4° definition but not the 2° definition.