JSLHR

Tutorial

An Open Conversation on Using Eye-Gaze Methods in Studies of Neurodevelopmental Disorders

Courtney E. Venker[a] and Sara T. Kover[b]

Purpose: Eye-gaze methods have the potential to advance the study of neurodevelopmental disorders. Despite their increasing use, challenges arise in using these methods with individuals with neurodevelopmental disorders and in reporting sufficient methodological detail such that the resulting research is replicable and interpretable.
Method: This tutorial presents key considerations involved in designing and conducting eye-gaze studies for individuals with neurodevelopmental disorders and proposes conventions for reporting the results of such studies.
Results: Methodological decisions (e.g., whether to use automated eye tracking or manual coding, implementing strategies to scaffold children's performance, defining valid trials) have cascading effects on the conclusions drawn from eye-gaze data. Research reports that include specific information about procedures, missing data, and selection of participants will facilitate interpretation and replication.
Conclusions: Eye-gaze methods provide exciting opportunities for studying neurodevelopmental disorders. Open discussion of the issues presented in this tutorial will improve the pace of productivity and the impact of advances in research on neurodevelopmental disorders.

Eye-gaze methods (i.e., automatic eye-tracking systems or manual off-line coding of eye movements) are appealing for studying neurodevelopmental disorders (NDD) because they have the potential to assess real-time processing with precision, have limited behavioral task demands, and can address clinically relevant questions about developmental mechanisms and individual differences. By providing a window into underlying cognitive processes, these methods allow researchers to ask innovative questions that have energized the field and accelerated the progress of research on NDD. Despite these advantages, several challenges limit the progress of this work. First, concerns with study design and data collection that are unique to children with NDD have not been thoroughly addressed. These issues center upon selecting an eye-gaze procedure, designing a task, and collecting data in a manner that supports the involvement of children from multiple participant groups, which may include a wide range of ages or developmental levels. Second, published research reports often omit key methodological details, making it challenging for scientists and consumers to interpret and evaluate results, compare findings across studies, and replicate prior work (Kylliäinen, Jones, Gomot, Warreyn, & Falck-Ytter, 2014; Oakes, 2010). Given the small samples and substantial variability that characterize many studies of individuals with NDD, it is critical to discuss issues around data processing and cleaning, analysis, and interpretation. As with any specialized research methodology, using eye-gaze tasks for the study of NDD incorporates many considerations that directly affect the conclusions drawn from research—considerations that we believe deserve increased attention.

Our goals in this tutorial are to (a) highlight the considerations unique to designing and conducting eye-gaze studies in NDD and (b) propose conventions for reporting research using eye-gaze methods in NDD. We believe that open discussion of these issues will improve the quality and integrity of research using eye-gaze methods in the field of NDD while increasing the pace of productivity and the impact of advances in this domain. Although some points may be relevant to other methodologies, the current discussion focuses on research that uses automatic eye tracking or manual off-line coding to measure the location of children's eye gaze on a screen. The majority of issues raised in this tutorial pertain to both automatic eye tracking and

[a]Waisman Center, University of Wisconsin–Madison
[b]University of Washington, Seattle
Correspondence to Courtney E. Venker: cgerickson@wisc.edu

Editor: Rhea Paul
Associate Editor: Linda Watson
Received October 28, 2014
Revision received March 19, 2015
Accepted July 27, 2015
DOI: 10.1044/2015_JSLHR-L-14-0304

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

manual coding; however, the three sections that pertain to only one method state so explicitly (i.e., "Calibration for Automatic Eye Tracking," "Manual Eye-Gaze Coding," and "Processing Data From Automatic Eye Trackers"). Because our own work has focused on children with NDD, we refer to participants as children; however, this tutorial may also be applicable to studies of adolescents and adults. We have provided brief definitions of eye-gaze terminology when appropriate; see Holmqvist et al. (2011) and Wass, Forssman, and Leppänen (2014) for more extensive definitions.

What Eye-Gaze Methods Offer to the Study of NDD

A Sensitive Measure of Individual Differences in Real Time

Eye-gaze methods are beneficial for studying NDD because they have the potential to provide information about complex cognitive processes simply by measuring where an individual looks at specified moments in time. For example, eye-gaze methods are capable of revealing subtle differences in speed and accuracy of lexical processing (Fernald & Marchman, 2012; Weisleder & Fernald, 2013). Such sensitivity is valuable because studies of NDD often examine characteristics at the level of individual children in order to uncover broader patterns within heterogeneous populations (e.g., extent of variability, associations among skills; Eigsti, de Marchena, Schuh, & Kelley, 2011). Further, eye-gaze methods provide information about visual and/or auditory attention in real time as processing occurs. Measures that capture responses after processing has occurred (e.g., verbal response, point, button press) do not offer this same advantage. Investigating real-time processing may help us to better understand processes of development (i.e., developmental mechanisms) in individuals with NDD, eventually informing how best to intervene for optimal outcomes. For example, eye-gaze methods may reveal effects of specific intervention strategies that are not immediately apparent in a treatment context; they may also capture subtle changes in processing and learning that underlie more global changes, which is important for optimizing treatments.

Limited Behavioral Response Demands

Whereas the cognitive demands of a task depend on the experimental design and the research question(s) being addressed, behavioral response demands relate primarily to the assessment method. Because eye-gaze methods require only passive engagement—namely, sitting and looking at a screen—they have limited behavioral demands relative to assessment techniques that require a purposeful response (Abbeduto, Kover, & McDuffie, 2012; Falck-Ytter, Bölte, & Gredebäck, 2013; Karatekin, 2007). This means that eye-gaze methods may be well suited for children with NDD who are unable to follow an examiner's overt requests or are distressed by them (Brady, Anderson, Hahn, Obermeier, & Kapa, 2014; Charman, Drew, Baird, & Baird, 2003). Factors such as social engagement or motor planning can affect the content validity of some examiner-directed tasks, but, depending on the experimental design, eye-gaze methods may lessen these effects and/or quantify their impact on processing and learning.

The Opportunity to Ask New Questions

Perhaps the greatest advantage of eye-gaze methods is that they offer researchers the opportunity to address new questions about developmental mechanisms in NDD (Falck-Ytter et al., 2013; Sasson & Elison, 2012). Coupled with thoughtful experimental design, studies using eye-gaze methods can investigate a wide array of developmental constructs, including attention, memory, face processing, social communication, reading, and language processing and learning—domains of ability that are central to the study of NDD (Guillon, Hadjikhani, Baduel, & Rogé, 2014; Papagiannopoulou, Chitty, Hermens, Hickie, & Lagopoulos, 2014; Swensen, Kelley, Fein, & Naigles, 2007). Findings from this type of work have not only theoretical implications for understanding NDD but also clinical implications that may help improve the lives of individuals with NDD and their families. Much research on NDD is conducted with the ultimate goal of developing improved prognostic methods, identifying moderators and mediators of treatment effects, and optimizing functional outcomes. Research advances made possible by eye-gaze methods may bring the field closer to these objectives.

Considerations for Using Eye-Gaze Methods in Studies of NDD

Utilizing eye-gaze methods to study NDD has clear advantages, but carrying out such studies requires a wide variety of skills and expertise. Although a number of published empirical articles (Nyström, Andersson, Holmqvist, & van de Weijer, 2013; Wass et al., 2014; Wass, Smith, & Johnson, 2013), methodological reviews (Gredebäck, Johnson, & von Hofsten, 2010; Karatekin, 2007; Oakes, 2012), publication guidelines (e.g., Oakes, 2010), and conferences (e.g., EyeTracKids) have addressed the use of eye-gaze measures to study typical development, methodological issues specific to children with NDD have received less attention (but see Kylliäinen et al., 2014; Sasson & Elison, 2012). Discussing this topic is worthwhile because the issues that arise in studies of individuals with NDD may differ from those that arise in studies of typically developing individuals. For example, children with NDD may demonstrate atypical visual behaviors (e.g., peering at visual stimuli; Kim & Lord, 2010) that complicate the measurement of gaze location and produce high levels of missing data. In addition, eye-gaze tasks often take place in a small, enclosed area, such as a darkened, soundproof booth—an environment that some children with NDD may find distressing. Many issues are difficult to anticipate a priori because they emerge gradually as a study progresses—sometimes despite extensive

pilot testing—and are often resolved only through direct experience and trial and error. Last, research on individuals with NDD often involves a comparison group, meaning that a single eye-gaze task needs to be optimized for multiple populations of children who may differ widely in age or language ability. The challenge is balancing standardization of the task procedures with individualization to allow the best possible experience for all participants (Kylliäinen et al., 2014). Here, we present a discussion of key considerations involved in designing and conducting eye-gaze studies for individuals with NDD.

Types of Eye-Gaze Procedures

One of the most basic considerations in any eye-gaze study is the type of procedure used. Whether conducting or interpreting an eye-gaze study, it is important to consider the numerous factors dictated by this choice. In this tutorial, we distinguish between two primary eye-gaze methods: automatic eye tracking and manual eye-gaze coding. Although some automatic eye trackers incorporate head-mounted equipment and/or head restraints, this tutorial focuses on automatic eye-tracking systems that use remote cameras to record corneal reflections and locate the pupil; these systems do not require physical contact between the equipment and the participant and are thus more suitable for many individuals with NDD (Falck-Ytter et al., 2013). Remote automatic eye trackers detect gaze by creating reflections off of the cornea of the eye using a light source, typically a near-infrared light panel; this information is recorded by an internal camera, and processing algorithms are applied to determine the position of the eye(s) and the point of gaze location.

Manual coding, in contrast, involves videotaping participants' faces during a task and later coding the direction of the participant's gaze for each frame of that video; such procedures may be referred to as looking while listening (Fernald, Zangl, Portillo, & Marchman, 2008; Swingley, 2012) or intermodal preferential looking (Piotroski & Naigles, 2012). Manual eye-gaze coding studies typically present visual stimuli on a large screen (e.g., 55 in.) and also involve no physical contact with the participant. The "high-tech" nature of eye tracking may initially be more attractive than the slower-paced nature of coding data by hand; however, a choice made entirely on the basis of this novelty is unlikely to be satisfactory.

It is critical that researchers choose the procedure that is best suited for their participants and their research questions. In some cases, research questions could be addressed using either method (e.g., studies that measure looking to one half of the screen vs. the other). When there is a choice, selection of an eye-gaze procedure would ideally be based entirely on substantive factors—the research questions, the intended outcome variables, and the behavioral characteristics of the participants. In reality, this decision is also likely to incorporate logistical factors, such as equipment cost and availability. In some cases, the decision is relatively straightforward because one method lends itself more naturally to the research questions and intended outcome variables. For example, studies that require distinguishing gaze location with high spatial precision (e.g., identifying individual fixations within each image on the screen; Yu & Smith, 2011) will likely use automatic eye tracking. However, researchers should be aware that features such as accuracy and precision are not inherent to an eye-gaze method or a given eye-tracking system. In the case of automatic eye trackers, specifications provided by a vendor are likely based on behaviors of typical adults and could be quite different for participants with NDD (although recent advances have made it possible for researchers using automatic eye tracking to calculate aspects of data quality for their participant groups; see "Missing Data"). The discussion below presents several factors that can help guide researchers in selecting an eye-gaze method.

Determining Gaze Location

Because eye trackers record gaze location automatically, they lend objectivity and limit the potential for human error in determining gaze location (see "Processing Data From Automatic Eye Trackers"). Manual coding is neither automatic nor free of the risk of bias because it relies on human coders to make decisions about gaze location (Oakes, 2012; Wass et al., 2013). In addition, manual coding takes time (e.g., 1–2 hr for a 4-min video) as does training coders to lab standards and coordinating lab-wide agreement checks (see "Manual Eye-Gaze Coding").

Eye tracking also has potential drawbacks. Eye trackers produce missing data when children move outside of the tracking range, and manual coding may capture data during these portions of the trial (see "Missing Data"). Processing eye-tracking data can be quite complex and time-consuming (Oakes, 2012) and may require an experienced computer programmer to develop customized data-processing scripts in programming languages, such as MATLAB (MathWorks, Natick, MA) or R: A Language and Environment for Statistical Computing (Vienna, Austria). Default processing algorithms may make unknown or undesirable assumptions, and changing aspects of eye-tracking processing algorithms can produce drastically different results (see "Processing Data From Automatic Eye Trackers"). Automatic eye tracking also requires calibration, which may be a disadvantage for collecting data from some individuals with NDD.

Calibration for Automatic Eye Tracking

Automatic eye trackers require calibration at the start of a task to maximize accurate measurement of gaze location (manual coding does not require calibration). During calibration, a participant's gaze is recorded while viewing a sequence of images on the screen; depending on the type of calibration used, the examiner may need to press a button when the participant is looking at each respective image to indicate when the eye tracker should record calibration data. On the basis of the difference between the location of the calibration stimuli and the recorded gaze location, the eye tracker applies a correction to the algorithms for

measuring gaze location during the remainder of the session. Researchers have several choices to make about calibration, including what stimuli to use and how many calibration points to present. Using silent red dots may be effective in studies of typical adults, but individuals with NDD may show better performance with a calibration option that uses dynamic, musical animations to attract attention. Although some studies use a 9-point calibration, using fewer calibration points (e.g., five) will allow participants to move past the calibration phase of an experiment more quickly (Sasson & Elison, 2012).

Although calibration may seem straightforward, there are several issues of which to be aware. Some participants may not be able to complete calibration and therefore may need to be excluded from the analyses (e.g., Tenenbaum, Amso, Abar, & Sheinkopf, 2014). Calibration errors may necessitate off-line corrections with an adjustment algorithm (Frank, Vul, & Saxe, 2012), and the quality of calibration can have implications for data quality (Nyström et al., 2013). Recalibrating at multiple points during and/or after the experiment (e.g., by recording gaze to a fixation cross between trials) allows researchers to quantify and report calibration error (see Oakes, 2012), which may differ across participant groups.

Missing Data

Limiting data loss is desirable in any study, but this issue is particularly important in studies of individuals with NDD for several reasons. First, participants with NDD are more challenging to recruit than those with typical development because NDD affect only a portion of the population. Small sample sizes have little tolerance for missing data, especially when the anticipated effects are small and there is considerable variability among participants. Second, behavioral phenotypes associated with NDD may increase the likelihood of missing data. For example, individuals with fragile X syndrome may accrue more missing data due to task-related gaze avoidance (Kover, 2012; Murphy, Abbeduto, Schroeder, & Serlin, 2007). Third, many studies of individuals with NDD use a group-matching design, which can complicate issues surrounding missing data (see "Group Matching"). Data should be included from as many participants as possible to increase the generalizability of findings (see "Interpretation"), particularly when participants' travel to a testing site requires considerable time and resources.

Missing data aligns with a measure of data quality known as *robustness*, defined as "…the relative proportion of periods of data presence vs. absence during recording" (Wass et al., 2014, p. 440). It is unfortunate, although not surprising, that data quality can differ between and within individuals and can have a direct impact on the results of a study. In a study by Wass et al. (2014), infants had lower data quality than adults; data quality was higher at the beginning than the end of a session; and less robust data was associated with increased head movement. Even more concerning was the finding that lower data quality was associated with specific outcome variables—longer latencies and less looking to the eyes of a face. On the basis of these findings, Wass et al. concluded that it is critically important for researchers to assess data quality, compare quality between groups, test the impact of quality on the study results, and consider including quality as a covariate when appropriate. They have developed algorithms in MATLAB (MathWorks) for analyzing data quality, available for download (https://www.mrc-cbu.cam.ac.uk/people/sam.wass/).

Given the importance of limiting data loss, it is valuable to consider how missing data is handled in automatic eye tracking versus manual coding. Automatic eye-tracking systems typically allow for some degree of head movement, but blinks, gaze away from the screen, excessive head movement, and eye-tracker errors can nonetheless lead to considerable amounts of missing data (Brock, Norbury, Einav, & Nation, 2008; Oakes, 2012; Wass et al., 2014; Yu & Smith, 2011). In addition, there is a period of time required for the eye tracker to "recover" after losing a participant's gaze, so there may be instances when participants are fixating the visual stimuli but the eye tracker is not recording gaze (Oakes, 2012). Using a track status box during an experiment helps to address these issues; this box monitors the quality of gaze tracking in real time and allows the examiner to change the participant's positioning in cases of drastic data loss.

One aspect of automatic eye tracking that has the potential to limit missing data is gaze contingency, a design feature that allows trials to proceed only after the child's gaze is registered on the screen for a specified amount of time. Gaze contingencies can also be incorporated into manually coded experiments (e.g., through a button press), but these contingencies are temporally and spatially less precise because they require the examiner to judge gaze location in real time. In theory, gaze contingencies should limit data loss because they increase the likelihood that the child will be looking at the screen immediately before the presentation of key information. However, using a gaze contingency may have undesirable side effects that limit its value, such as increasing the length of an experiment for a child who frequently looks away from the screen or frustrating a child who is looking at the screen but is outside the tracking range. Researchers using a gaze contingency should report descriptive data about the time required for the contingency criterion to be met for each participant group. Additional research is needed to determine the impact of gaze contingencies on children with NDD.

Although empirical data are required to confirm this hypothesis, manual coding may limit data loss because determining gaze location from video in some cases offers more flexibility for participants who move around a great deal during the experiment. For example, an automatic eye tracker cannot track gaze location when a child leans outside of the tracking range; however, it may be possible to reliably identify gaze location through manual coding if the child's eyes are visible on the video. In addition, manual coding may be advantageous for studying the amount of attention allocated to certain stimuli because coders can

differentiate the exact source of missing data (e.g., blinking, looking away, video error). In automatic eye-tracking output, missing data attributable to movement outside the tracking range or to eye-tracker errors cannot be easily distinguished from missing data due to looking away from the screen.

When the research questions and planned outcome variables make this possible, it may be advantageous to use both automatic eye tracking and manual coding methods in combination. Video recording participants' faces during an automatic eye-tracking task is an adaption that allows for manual coding in addition to the gaze data collected by the eye tracker; in these cases, however, manual coding may be more difficult because of the restricted visual angle of a smaller screen. An automatic eye-tracking component can also be added to manual coding studies by placing a mobile eye tracker (e.g., Tobii X2-60) in front of a large screen along with the video camera. Using both methods may offer at least two advantages. First, this practice can limit data loss and maximize data-processing efficiency in individual studies. Second, it will allow researchers to conduct methodological comparisons that have broader implications for the field.

Temporal Resolution

Temporal resolution refers to the frequency with which gaze location is measured (i.e., sampling rate). Automatic eye trackers typically offer higher sampling rates (e.g., 60 Hz or 120 Hz, producing time bins of approximately 16.7 ms or 8.3 ms) than manual eye-gaze procedures (e.g., 30 Hz, producing time bins of approximately 33.3 ms, although high-speed cameras are available). Although higher sampling rates may seem advantageous, the temporal resolution required depends entirely on the research questions and proposed outcome variables. Increased temporal resolution may be valuable if the planned analyses will take advantage of the number of samples provided. However, gaze patterns are often averaged and analyzed over longer spans of time, in which case the potential advantage of high temporal resolution is lost (Brady et al., 2014; Naigles, Kelty, Jaffery, & Fein, 2011; Venker, Eernisse, Saffran, & Ellis Weismer, 2013). Growth curve analyses that examine gaze patterns over time may take advantage of higher temporal resolution (see "Time Course Analysis"), but even researchers using this approach may aggregate eye-tracking data over longer time windows (e.g., 50 ms) to account for eye movement-based dependencies (Barr, 2008) and to reduce the sheer volume of data.

Understanding temporal resolution is also important when combining data gathered using different eye-gaze methods, either at a single site or across sites. Research on NDD often involves collaboration across sites to increase recruitment. When testing at multiple sites is necessary, monitor resolution for automatic eye trackers might need to be adjusted prior to collecting data to ensure that data can be aggregated without errors (Benjamin et al., 2014). It is also possible to adjust the sampling rate after data have been collected, during data processing. For example, data from automatic eye tracking and manual coding can be combined or compared by "down-sampling" the temporal resolution to the lower rate of the two methods.

Spatial Resolution

Spatial resolution refers to the sensitivity with which gaze location is measured. Manual coding has relatively limited spatial resolution because coders judge where children are looking on the basis of children's eye movements. Coders can be trained to reliably identify gaze to the left, right, and center of the screen as well as shifts or gaze away from the screen (Fernald et al., 2008; Naigles & Tovar, 2012). However, it is more difficult to reliably differentiate gaze to smaller areas. Automatic eye tracking, on the other hand, can provide a much more precise measure of gaze location both within and between researcher-defined areas of interest (AOIs, e.g., the eye region of a face situated in a social scene; see "Areas of Interest").

The increased spatial precision offered by automatic eye tracking can allow gaze location within relatively small areas to be parsed into fixations (i.e., periods of focused gaze), saccades (i.e., ballistic eye movements that attempt to bring visual stimuli into view), and smooth pursuit eye movements (i.e., movements required to track objects moving slowly and smoothly; Karatekin, 2007). Some high-speed eye trackers are capable of tracking microsaccades, small saccades during a fixation that occur one to two times per second (Martinez-Conde, Otero-Millan, & Macknik, 2013). However, parsing gaze data into separate components is complex and can require a great deal of expertise (see "Processing Data From Automatic Eye Trackers").

Physical Equipment

Automatic eye tracking and manual coding involve different physical equipment with accompanying benefits and drawbacks. The physical equipment required for manual eye-gaze coding tends to be less expensive than automatic eye trackers, but the price of automatic eye trackers has decreased considerably in recent years. Most eye-tracking screens are the size of a standard or widescreen desktop monitor, although some eye trackers can be used with a larger screen. Optimal tracking of gaze location requires that participants sit a specific distance away from the screen (e.g., 60 cm). In contrast, manual coding studies often involve a larger screen and require only that participants sit in a position where their eyes can be recorded clearly by a video camera for later off-line coding. Pilot testing can determine the utility of the equipment for specific participant populations.

Both automatic eye tracking (e.g., Tobii X2-60) and manual coding procedures (e.g., mobile intermodal preferential looking; Naigles & Tovar, 2012) offer portable equipment. Portable equipment allows a child to be assessed in off-site locations, such as the home or school setting, which allows researchers to examine behavior in more naturalistic conditions and may optimize a child's performance relative to an unfamiliar testing site (Naigles & Tovar, 2012). In addition, this equipment may be more cost-effective than

stationary testing booths and permits collection of data from participants who find it difficult to travel to a testing site. However, off-site procedures also require more care in assuring that testing conditions are similar across participants and that environmental distractions are limited.

Experimental Design

The primary advantage of eye-gaze methods lies in their ability to address interesting research questions when utilized with a thoughtful experimental design—that is, the specifics of the experimental task are paramount. Although the differences between automatic eye tracking and manual coding are numerous, there is overlap in the considerations for designing and conducting the studies that use these methods. Here, we highlight two task design factors.

Length of Task

As with any experimental task, an effective eye-gaze task provides a valid measure of the construct of interest by balancing the number of trials (more is better) with the length of the task (shorter is better). Pilot testing helps determine when the task is long enough to provide an accurate estimate but short enough that children do not become bored (Kylliäinen et al., 2014). Because of attentional limitations, tasks tend to be shorter in studies of younger children or children with NDD; however, this may reduce measurement reliability (Karatekin, 2007), which is unfortunate because reliability is critical for tracking change in individual children and making accurate longitudinal predictions. One option for increasing the total number of trials is to have participants complete blocks of a task on multiple visits; if there are no significant differences in performance across visits, it may be possible to combine the trials for each participant. Performance may also differ across the first and second half of a task or become more variable as the task progresses; this is also true of data quality (Wass et al., 2014). These effects may be undesirable, difficult to control, or different across groups and should be assessed during the piloting process and reported in full if they occur. Additional studies are needed to investigate this issue (but see Farris-Trimble & McMurray, 2013; Farzin, Scaggs, Hervey, Berry-Kravis, & Hessl, 2011, for examples of reliability studies).

Stimuli

Effective eye-gaze tasks are comprised of stimuli that are appropriately salient for the given research questions and participant characteristics. Auditory stimuli may be engaging (e.g., child-directed speech) or neutral (e.g., adult-directed speech). Visual stimuli may be dynamic (e.g., videos of social scenes) or static (e.g., images of common objects). Static stimuli are often placed on a gray square to provide a background contrast appropriate for both lighter and darker colored objects. In most cases, visual salience should be balanced across trials and conditions to avoid bias; for example, young children may prefer to look at animate rather than inanimate objects (Swingley, 2012), which can lead to an unintended confound if salience is not balanced. In addition, some eye-gaze tasks contain both familiar and novel stimuli to help maintain participants' attention by allowing them to take breaks from a challenging task.

Another strategy used to maintain children's attention is the incorporation of "attention-getter" stimuli among the experimental trials. Attention-getters can take a variety of formats, including encouraging sayings (e.g., "You're doing great!") presented with pictures that move across the screen or short, animated video clips with musical backgrounds (e.g., nature scenes accompanied by classical music). Simple visual centering stimuli between trials, such as flashing lights, may be useful for some children, including those with autism spectrum disorders (ASDs; Piotroski & Naigles, 2012). Participants' ages, developmental levels, and language comprehension abilities should inform stimulus creation, and adequate task piloting will allow researchers to optimize task design. Short of having standardized, norm-referenced eye-gaze measures designed for children with NDD, researchers must take these factors into account as they create stimuli.

Implementing an Eye-Gaze Task

There is some skill required to successfully implement an eye-gaze task with any child, and the behavioral phenotypes of some children with NDD underscore the importance of carefully designing eye-gaze testing protocols. Involvement of an examiner with considerable clinical experience with individuals with NDD will optimize participants' performance by limiting behavioral interference and minimizing data loss (Kylliäinen et al., 2014).

Preparing the Participant

Some children with NDD may have difficulty participating in eye-gaze tasks because they involve new activities in unfamiliar environments (Sasson & Elison, 2012). Reading a story about the procedures ahead of time, either at home or at the research site, may decrease anxiety. For example, an illustrated booklet read before the task could inform children that it is now time to watch a short movie, that they will go into a room that is a little bit dark, that they can sit quietly and watch and listen, and that they will get a sticker (Benjamin et al., 2014; Kover, 2012).

Transitioning Between Activities

Many children with NDD have difficulty transitioning between activities. One strategy used to ease transitions is a visual schedule—a printed schedule, with words and/or pictures, that shows what activities have been completed and what comes next. Visual schedules can increase children's tolerance by showing that an eye-gaze task is part of a larger testing protocol or that an eye-gaze task contains multiple blocks with breaks in between. In this way, children know when the task will occur and what will happen when it is finished (e.g., a play break). They can check off or place a sticker next to each activity after it occurs.

Scaffolding the Environment

To limit distraction, eye-gaze studies are often conducted in a soundproof booth or a small area within a larger room defined by modular walls. Wall coverings that obscure wires and other physical equipment may help participants focus on the screen, particularly if the room is brightly illuminated before the task. Although it is typically detrimental for children to hold anything that might distract them (Swingley, 2012), some children with ASD may be more likely to attend to the task while holding an object (e.g., a toy, food; Piotroski & Naigles, 2012). Implementing a standard procedure for offering a calming strategy or object to participants in all groups maintains consistency and may allow for easier interpretation of performance across groups (Kylliäinen et al., 2014).

The seating arrangement should provide a clear expectation of where the participant should be situated. Having infants and toddlers sit on the lap of a parent or caregiver often places children at an appropriate height to view the screen, and the presence of a familiar adult provides reassurance. Other studies use booster seats with straps to minimize movement and adapt for differences in body size (e.g., Brady et al., 2014). Because automatic eye trackers require participants to be seated within a relatively restricted three-dimensional space, it is important to allow adjustment for participants of variable heights (Sasson & Elison, 2012). This flexibility can be achieved by changing the height of the eye tracker (e.g., with a movable wall mount) and/or the height of the chair (e.g., with a hydraulic chair; Brady et al., 2014). When a child sits independently in a chair to complete an eye-gaze task, an informed adult should be present in the testing booth for safety. Making a knowledgeable decision about seating on the basis of the age and characteristics of the participants will maximize the likelihood of obtaining valid data.

Instructing the Adult in the Room

Regardless of whether the adult in the testing room is a caregiver or a member of the research team, this person must receive some explicit instruction on what he or she should and should not do during the task. Parents may be requested not to talk or gesture to avoid unintentionally influencing children's performance. A trained and experienced research assistant may be instructed to sit or stand behind or next to the child with gaze down, disengaged from the child and the task. To prevent accidental influence on the child's behavior, the adult may also be blinded to the auditory or visual information being presented in the form of sunglasses blacked out with tape or headphones that play masking music. Children with NDD may be distracted by such items, particularly when used by a caregiver; it is helpful to have a backup plan wherein the adult simply closes his or her eyes. However, the adult will minimally need to either see or hear the child to determine if the task should be stopped if the child becomes distressed. In addition to explaining the overall purpose of a study at the start of the session, it may also be helpful to provide a full debriefing to a parent or older child after the experiment, so he or she can better understand what the researchers are trying to learn and how such a task helps answer the research questions.

In-the-Moment Scaffolding

It is necessary to consider ahead of time if and how in-the-moment scaffolding will be used to support children's participation. Children with limited attention may benefit from nonverbal cues (e.g., placing hands on the child's shoulders) or verbal cues (e.g., quietly reminding the child to watch the movie) provided between trials or blocks. Integrating an optional pause between trials (e.g., as part of the attention-getter) allows children to take a short break when necessary. An ideal testing session would require no additional scaffolding, but designating a standardized hierarchy of increasingly supportive cues may allow a child to complete the task in such a way that his or her data contribute to the planned analyses. Such a protocol might, for example, instruct a research assistant to place his or her hands on the child's shoulders only after he or she has shown considerable inattention for several trials. Prompts can also be built into the experiment and thereby automatically logged. Examiners should use a data sheet to take written notes on the type and extent of scaffolding used as well as any other notable changes in protocol (Kylliäinen et al., 2014).

Conventions for Reporting Research Using Eye-Gaze Methods in NDD

Each of the factors described above can ultimately affect the conclusions drawn from an eye-gaze study. However, there are other aspects of conducting research using eye-gaze methods that can be even more influential but less likely to be reported in published studies: namely, decisions about data coding and processing, data cleaning, and analysis. By providing a set of conventions, our hope is to improve the quality of research on NDD using eye-gaze procedures.

Data Processing

Areas of Interest

Regardless of the type of stimuli or the eye-gaze method used, researchers must designate AOIs—researcher-defined locations on the screen that will be used to differentiate gaze locations during manual coding or during processing of automatic eye-tracker data (e.g., left half of the screen vs. right half of the screen, eyes vs. mouth, named object vs. unnamed object). It is useful to think about the size and placement of AOIs in terms of visual angle; in general, AOIs that are farther apart have a larger visual angle of separation, which helps coders more accurately differentiate gaze location and reduces the chance of error in automatic eye-tracking studies. Although it is possible for AOIs to be adjacent and narrowly defined, "designing experiments with AOIs defined in this way is not in accordance with the practices recommended by the eye-tracker manufacturers, who

recommend defining AOI sizes on the basis of the expected accuracy performance of the eye-tracker and the study design" (Wass et al., 2014, p. 446). Researchers may designate AOIs on the basis of the expected accuracy and precision of their automatic eye tracker as reported by the vendor. As noted, however, researchers should consider quantifying the accuracy and precision for their specific participant groups because these factors are not inherent to a given eye-tracking system.

Manual Eye-Gaze Coding

Unlike automatic eye tracking, manual off-line eye-gaze coding is completed frame by frame, by hand. Learning to reliably code eye gaze is a nontrivial task, and studies of individuals with NDD that use manual coding require special attention to training, coding, and intercoder agreement and reliability. Coding training typically involves reviewing a manual of coding guidelines, observing an experienced coder, and coding a series of training videos. Depending on the populations being studied, it may be beneficial to develop separate training videos of typically developing children and children with NDD because coders will need extensive experience coding atypical behaviors (e.g., peering) to ensure consistency. Decisions also need to be made about situations that may arise for older children with NDD who participate in adaptations of eye-gaze tasks originally designed for younger infants and children. For example, children are more likely than infants to point at the screen, potentially blocking their eyes, and coding manuals should present consistent guidelines for how to handle such behaviors.

It is beneficial to conduct lab-wide agreement checks at periodic intervals (e.g., twice per year), in which each coder independently codes the same video and any discrepancies are discussed at a group meeting. These checks prevent drift from the original coding guidelines and ensure that coders maintain consistency within and across studies. It may also be useful to compare independently coded videos for a specific study more often (e.g., every fifth video; Yoder & Symons, 2010) to limit drift between coders and from the original coding guidelines over the course of a study. This practice is particularly important in studies of individuals with NDD because recruitment and data collection may take place over a long period of time, resulting in relatively high rates of coder turnover and extended time lapses since initial training.

To maximize the accuracy of manual coding, it is necessary to establish standards for intercoder agreement. Because these standards depend on study-specific factors (e.g., task, participant populations), we provide only general guidelines for researchers to consider (also see Yoder & Symons, 2010). First, coding agreement should be reported separately for each participant group. It may be systematically more challenging to code one group than another, and reporting only a study-wide average could unintentionally obscure lower rates of agreement in a particular group. Second, it is beneficial to report agreement measures that correspond to the outcome variables of interest. For example, one coding system examines two types of intercoder agreement: one that pertains to accuracy and is calculated across the full task and one that pertains to processing speed and is calculated only during gaze shifts (Fernald et al., 2008). This is a valuable distinction because coding focused gaze is much simpler than coding shifts in gaze, and failing to separately report agreement for gaze shifts could conceal this difference.

Third, intercoder agreement calculations should encompass all relevant parameters and reflect all coded trials. By default, some coding software may omit from agreement calculations those trials with a different number of gaze shifts (i.e., noncomparable trials). This practice artificially inflates agreement by excluding trials with the highest levels of disagreement. In such cases, the percentage of noncomparable trials should be reported for each group, as well as how these trials were handled for calculating intercoder agreement. For example, noncomparable trials may be consensus coded: Two coders could recode such trials on the basis of a joint decision, and the agreement percentages could be recalculated. Although published manuscripts do not typically report information about noncomparable trials, it is critical for establishing standards for intercoder agreement of manual coding and should be presented in published reports.

Last, although intercoder percentage agreement is the most commonly reported indicator of coding quality, it may be beneficial to report other metrics as well, such as the intraclass correlation coefficient (ICC). ICC is an index of intercoder reliability that takes into account the amount of variance in the reliability sample; it represents the proportion of variance accounted for by differences in participants' true score estimates of a particular variable as opposed to the differences between coders (Yoder & Symons, 2010). Said another way, ICC represents the degree of resemblance between microunits (i.e., independent coders) within the same macrounit (i.e., participant). Whereas percentage agreement compares the gaze location assigned by each coder for each time frame, ICC compares the designated outcome variables (e.g., average length of first look) derived from each coder's file. For example, average length of first look would be calculated twice for each participant in the reliability sample—once on the basis of the first coder's file and once on the basis of the second coder's file—and these values would be entered into the ICC calculation. ICC should be based on a reliability sample of at least five different videos coded by two independent coders (Yoder & Symons, 2010).

Processing Data From Automatic Eye Trackers

Some aspects of processing eye-gaze data are specific to automatic eye trackers. For example, manual coding guidelines may instruct coders to ignore short blinks, thus avoiding the corresponding segments of missing data; in contrast, blinks are represented as missing data in raw eye-tracking data and must be handled during the processing stage. For this reason, it is common to interpolate (i.e., fill in) short segments of missing eye-tracking data under a specified maximum length (e.g., 150 ms; Wass et al., 2014).

Processing automatic eye-tracking data also involves mapping gaze location onto user-defined AOIs that are designated at the outset of the study (see "Areas of Interest").

If a researcher plans to categorize and analyze specific types of gaze (e.g., fixations, saccades), these behaviors must also be defined at the data-processing stage. One option is to use default algorithms to define the temporal and spatial parameters of these behaviors; however, default processing algorithms may make assumptions of which the researcher is not aware (Oakes, 2010). Default algorithms may also produce undesirable systematic relationships between data quality and fixation length, potentially allowing poor measurement quality to obscure true gaze patterns (Wass et al., 2013). Another option is to customize algorithms. As discussed by Oakes (2012), however, there are no standards for defining the temporal and spatial boundaries of what constitutes a fixation in typically developing infants; this is also the case for individuals with NDD. It is perhaps even more concerning that changing aspects of algorithms can drastically change a study's results (Shic, Chawarska, & Scassellati, 2008).

Because of the complications involved in defining fixations and saccades and the potential impact on study results, some researchers may prefer to use raw gaze data. Using raw data requires fewer assumptions and can be simpler to explain and understand. Instead of parsing data into fixations and saccades, researchers can categorize each time frame of data as within or outside of the specified AOIs. This method produces small-scale longitudinal data that can be analyzed in a variety of ways, including averaging gaze across a large window or analyzing patterns of change over time (see "Time Course Analysis").

In any case, manual processing of data generated by automatic eye trackers should be reported (also see Oakes, 2010). Clear reporting of the parameters used in specific studies is a necessary first step in better understanding what constitutes a fixation in different populations of individuals with NDD (Gillespie-Lynch, Elias, Escudero, Hutman, & Johnson, 2013). It is particularly beneficial to the research community when researchers share new procedures they have developed to address limitations of default processing algorithms (Wass et al., 2013).

Data Cleaning

Prior to analyzing automatic or manually coded eye-gaze data, researchers may eliminate trials they consider to be invalid—a process sometimes referred to as *data cleaning*. Some trials are considered invalid due to task interference, such as a parent talking or a computer glitch that caused a trial to end prematurely. However, most invalid trials are those in which children have not looked at the experimental stimuli long enough to provide a valid measure of the variable(s) of interest (Fernald et al., 2008; Sasson & Elison, 2012). Although most researchers would agree that a 5-s trial with only 50 ms of looking time to the experimental stimuli is inadequate, it can be challenging to determine how much time is adequate, whether it must be continuous or cumulative, and which portion of the trial should be considered (see "Test Windows"). Furthermore, cleaning must occur not only at the trial level but also at the level of conditions and participants: How many trials must a participant contribute for a particular condition, and when should a participant be excluded altogether? Although one valid trial may not be enough for a participant's data to be included, it can be difficult to determine an appropriate threshold. Compounding these issues is the fact that published manuscripts often contain no explicit information about data loss (Oakes, 2010, 2012; but see Elsabbagh et al., 2013).

The purpose of data cleaning is to retain as much valid data as possible while excluding any data that provide an invalid measure of the construct in question. Requiring a large amount of looking time (i.e., eliminating more trials) increases confidence in the accuracy of the data but decreases statistical power by eliminating participants from the analysis and limiting the extent to which results can be generalized to the population represented by the original participant sample. Requiring a smaller amount of looking time (i.e., eliminating fewer trials) may increase statistical power by retaining more participants in the analysis and thus the generalizability of results, but runs the risk of including invalid data and thus producing inaccurate findings.

There are no industry standards for excluding invalid data even within the same area of research. For example, three studies investigating various aspects of language processing in children with ASD used the following cleaning criteria as minimum looking times for valid trials: 115 ms (25% of a window lasting 460 ms; Brock et al., 2008); 300 ms (Naigles et al., 2011); and 850 ms (50% of a time window lasting 1,700 ms; Venker et al., 2013). Similar issues arise regarding the number of trials a participant must contribute to be included in the analyses. In general, at least two or three trials is considered a minimum (Grieco-Calub, Saffran, & Litovsky, 2009; Venker, 2013), but some studies have required that children contribute data on more than half of test trials, with missing trials replaced by mean values for that age and condition (Piotroski & Naigles, 2012).

Data cleaning is an especially important issue in studies of NDD because participants may accrue higher levels of missing data than typically developing children, leading to exclusion of trials, conditions, and participants (Venker, 2013). This is particularly true when comparisons are made between conditions, and participants must contribute adequate data for each condition. How valid trials are defined and which participants contribute data to reported analyses are closely linked, leading to differences in the conclusions that might be drawn from a study and the interpretations that readers may make. In addition, correlational analyses require accuracy at the level of individual children to detect accurate patterns of behavior.

Although data cleaning is rarely discussed at length in published manuscripts, we recommend that researchers report cleaning criteria at the level of trials, conditions, and participants, as well as any changes to a priori cleaning

criteria. In addition, reporting patterns of data loss will aid interpretation, inform the design of future studies, and suggest new avenues for research. For example, in one study, children with ASD who were excluded from the analyses due to data loss had significantly lower receptive vocabulary levels than children who were retained (Venker, 2013). This is unfortunate because exclusion of children with severe language deficits precludes examination of the factors that led to such delays. If future studies continue to demonstrate that children with lower skills are most likely to be excluded, researchers will be faced with the challenge of designing tasks better suited for these participants.

Analyzing Eye-Gaze Data

Group Matching

In addition to the cascading effects of study design, data processing, and data cleaning, researchers must make decisions about data analysis that directly affect the results of a study. One issue pertains to whether the analyses utilize group-matching design. Such studies compare performance among groups of children that differ on one variable (e.g., diagnostic status) but are matched on another variable (e.g., nonverbal cognitive ability). Researchers should carefully consider the dimension of ability that is most appropriate as a matching variable when the dependent measures are drawn from eye-gaze tasks because the lower level behaviors underlying higher level skills are not fully understood. If the final matching variable is different from the matching variable selected a priori, this should be reported. The matching process should be clearly described with details regarding the selection criteria for participants and the extent of equivalence between groups, including effect sizes and variance ratios for the group difference on the matching variable (Kover & Atwood, 2013).

Establishing group equivalence can be a methodological hurdle in any study (e.g., Burack, Iarocci, Bowler, & Mottron, 2002; Mervis & Klein-Tasman, 2004), but concerns about group matching are exacerbated when data loss is substantial and differs across groups. Excluding participants on the basis of chronological age or nonverbal cognitive ability to match groups may lead to small sample sizes, variable numbers of trials, or poor group matches when also accounting for valid eye-gaze data (see Gillespie-Lynch et al., 2013). It is unclear how to interpret instances in which outcomes are correlated with number of valid trials contributed, particularly if this is the case in one participant group but not another. Is the number of valid trials an artifact of the task or a meaningful indicator of engagement that mirrors the ways in which a child interacts with visual and auditory stimuli in the natural environment? In addition, when groups differ considerably in the number of trials contributed, the accuracy of estimates of their eye-gaze performance may also differ.

Group comparisons can aid interpretation by demonstrating when performance differs from what would be expected for a particular age or ability level. However, researchers should also be aware of potential alternatives to traditional group comparison procedures. One alternative is a type of cross-sectional trajectories analysis in which the trajectory of the comparison group is applied to the data for the group(s) of individuals with NDD; performance is standardized by calculating the differences between the observed and modeled values (i.e., residuals; Jarrold & Brock, 2004). Another alternative is using comparison group performance as a benchmark against which to compare performance in the groups of participants with NDD by transforming raw data for individuals with NDD into referenced *z*-scores (Yoder, Stone, Walden, & Malesa, 2009). Development of norm-referenced scores for experimental eye-gaze tasks would also help circumvent some of the issues that complicate group comparisons. Depending on the research question, within-group designs and correlational analyses may lend themselves more naturally than group-matching designs to uncovering underlying mechanisms or qualitatively different patterns of behavior in individuals with NDD.

Outcome Variables

Another decision at the point of data analysis pertains to outcome variables. Depending on the experimental design, eye-gaze methods are capable of producing many different outcome variables, including the proportion or duration of gaze to various AOIs, the length and location of first look, the latency to look to a certain location, and the number of shifts between various AOIs. Even data loss could serve as a meaningful outcome variable. In some cases, the fact that eye-gaze methods yield multiple, distinct variables is a scientific advantage. For example, a participant group could show intact processing accuracy, as reflected by relative gaze to a named image, but impaired efficiency, as reflected by the speed of gaze shift to the named image. However, it can be challenging to interpret a study's findings when similar outcome variables produce inconsistent results (Kedar, Casasola, & Lust, 2006). Researchers should carefully select the outcome variables that best represent the construct(s) of interest and remember that different outcome variables will not always produce the same results.

Because eye-gaze methods can address a broad range of questions, there may be differences in terminology used to describe the reported variables. On the basis of the content area, gaze behaviors could be described as looking, gazing, attending, recognizing, fixating, shifting, or orienting; in addition, the same terms may mean different things to different people (Falck-Ytter et al., 2013). We encourage researchers to precisely define in published manuscripts the terminology they use to describe their variables, to prevent ambiguity or miscommunication.

Although outcome variables should be selected a priori, researchers may examine alternative outcome variables for a number of reasons. For example, it may not be possible to analyze the proposed outcome variable due to data loss (see "Limited Numbers of Trials"). In such cases, researchers may report results for an alternative outcome variable in addition to reporting the inability to analyze the

intended outcome. In addition, researchers may conduct exploratory analyses on different outcome variables to determine whether the results align. Assuming honesty and transparency, reporting different sets of results is valuable because it sheds light on which variables are most useful for measuring specific constructs in individuals with NDD.

Outcome variables from eye-gaze methods need not serve solely as dependent variables. These variables can also serve as predictors of later outcomes or as mediators or moderators of causal relationships among different developmental factors. For example, Weisleder and Fernald (2013) found that children's lexical processing efficiency mediated the relationship between parental language input and children's later expressive vocabulary. Auditory and/or visual processing measured by eye-gaze methods may also represent potential moderators of treatment effectiveness—that is, children with a particular characteristic at the start of intervention may benefit more from one type of intervention than another, as has been found for other child characteristics (Yoder & Compton, 2004; Yoder, Molfese, & Gardner, 2011).

Test Windows

In many cases, researchers are interested in analyzing gaze not during an entire trial, but during a window of time that begins at some point after the presentation of meaningful information (e.g., an informative word) and ends sometime before the next trial. This period of time—often called the test window—is important to consider because making small changes in the length and/or placement of this window can produce different results. Children with NDD may need more processing time than children in a comparison group, meaning that their peak performance may only be captured in later windows. It is unfortunate that there is limited empirical evidence researchers can use to guide their selection of a test window. One option is to examine and report exploratory test windows, which may reveal group differences not apparent when averaging eye-gaze behavior over a long portion of the trial or before most individuals in a particular group have shown the anticipated effect.

Limited Numbers of Trials

Another issue is that some outcome variables can only be derived from a subset of trials; these variables often incorporate more unsystematic variability (i.e., noise) than variables averaged across more trials. In lexical processing studies, for example, researchers are often interested in how long it takes children to look toward a named image. At best, researchers are working with approximately half of the trials available because latency data for gaze shifts toward the named image can be obtained only from trials in which the child happened to be looking at the unnamed image at noun onset (Fernald et al., 2008). At worst, there are only a handful of valid trials that contain high levels of noise and may not provide a meaningful signal across participants (Venker et al., 2013). Furthermore, some researchers omit trials in which participants are looking away or shifting at a certain point during the trial (Fernald et al., 2008); this practice may be justified, or it may result in unnecessary data loss.

Time Course Analysis

One analytical approach that may be particularly sensitive to subtle differences across participants and groups is a type of hierarchical linear modeling or linear mixed effects modeling called growth curve analysis (GCA; Barr, 2008; Mirman, 2014). GCA analyzes nested data in which each participant contributes multiple responses over time (e.g., each time frame during a 5-s trial, each trial in an experiment). In contrast to *t* tests or analyses of variance, which typically compare participants' performance during a test window against their baseline performance or against chance, GCA allows researchers to examine change in gaze patterns over time by incorporating linear, quadratic, and other higher order polynomials. GCA can quantify both fixed effects (e.g., diagnostic group) and random effects (e.g., slope values for individual participants) and can test associations between these effects and child-level factors (e.g., language ability). Researchers who use GCA are still faced with decisions about issues such as data cleaning, bin size, and test window; for example, the shape of a modeled curve changes considerably depending on the time window that is selected. Nonetheless, studies using GCA to analyze looking behaviors may shed light on underlying mechanisms that inform our understanding of behavioral phenotypes (McMurray, Munson, & Tomblin, 2014).

Interpretation

The most important part of any study is interpreting its findings and determining the corresponding clinical and theoretical implications. However, interpretation is challenging because it requires readers to consider how methodological and analytical decisions affect the results. Even seemingly trivial decisions—excluding invalid trials, requiring two versus three valid trials per condition, selecting a shifted test window, changing the duration of the test window, or adopting an alternate analysis—can change the statistical significance and/or effect size of the findings and, thus, the ultimate conclusions drawn.

One issue is how participant exclusion affects the generalizability of results. As in any empirical study, results can only be generalized to the population of children from which those in the participant sample with analyzed data were drawn, so readers should note which children were eliminated from the analyses because they did not contribute adequate data. Readers should remember that exploratory analyses increase the potential for Type I error and thus may produce sample-dependent results that are less likely to be independently replicated. Although we believe that exploratory analyses are critical to advance the use of eye-gaze methods with individuals with NDD, researchers must transparently report their procedures and rationale so readers can make an informed interpretation of their findings.

Another complication in interpreting eye-gaze data is that the skills underlying task performance are not necessarily well understood. Researchers should exercise caution when basing conclusions about high-level cognitive abilities on lower level oculomotor behavior that may develop atypically in individuals with NDD, conflating differences in the ability of interest (Karatekin, 2007). For example, it is unclear how differences across groups in volitional eye-movement control affect eye-gaze responses to stimuli designed to assess higher level cognitive skills (Kelly, Walker, & Norbury, 2013). Developmental trajectories of some aspects of eye-gaze measures for typical development are still under debate and so are unlikely to be resolved for children with NDD for some time (Karatekin, 2007). Researchers might acknowledge, and when possible, directly assess, lower level abilities potentially related to visual perception and eye movement that could explain or share variance with the eye-gaze task or variables of interest.

Although eye gaze provides a window into underlying cognitive processes, in many cases it does not represent the cognitive processes themselves. Different cognitive processes can produce similar gaze patterns, and, likewise, similar cognitive processes can produce different gaze patterns (Aslin, 2007). In addition, differences in gaze patterns may in some cases be caused by an unintended third factor, such as data quality (Wass et al., 2014). In other words, there may not be a one-to-one mapping between observed gaze behaviors and the construct of interest indicated by the research question. Acknowledging these limitations is crucial for appropriately interpreting findings, developing new research questions, and developing more advanced techniques that are needed to better understand NDD.

Final Comments

In this tutorial, we have presented a discussion of considerations for designing and conducting eye-gaze studies of individuals with NDD as well as a number of conventions for reporting the findings of such studies. It is our hope that the transparency brought about by this conversation will save time and effort in creating new studies; increase the likelihood of replicability; facilitate the sharing of data processing and analysis procedures; foster new collaborations, research ideas, and methodological studies; and lead to more informed decisions about issues such as missing data, data cleaning, and analytical approach. Improving the quality and pace of research on NDD ultimately will positively affect the lives of the children and families affected by them.

Acknowledgments

This work was supported by National Institutes of Health Grant F31DC012451 (awarded to Courtney E. Venker), Grant F31DC010959 (awarded to Sara T. Kover), Grant R01DC012513 (awarded to Susan Ellis Weismer), and Grant P30HD003352 (awarded to Marsha Mailick) as well as a Student and Early Career Council Dissertation Funding Award from the Society for Research in Child Development and a Dissertation Research Award from the American Psychological Association (both awarded to Sara T. Kover). Thanks to Paul Yoder and Rob Olson for providing comments on an earlier draft of this manuscript. We thank our mentors for providing us with the opportunity to access and learn about eye-gaze methodologies and our lab-mates for helping us brainstorm and problem-solve. Most of all, we gratefully acknowledge the children and families who participated in our eye-gaze studies and from whom we learned so much.

References

Abbeduto, L., Kover, S. T., & McDuffie, A. S. (2012). Studying the language development of children with intellectual disabilities. In E. Hoff (Ed.), *Research methods in child language* (pp. 330–346). Oxford, United Kingdom: Wiley-Blackwell. doi:10.1002/9781444344035

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10,* 48–53. doi:10.1016/j.biotechadv.2011.08.021.Secreted

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59,* 457–474. doi:10.1016/j.jml.2007.09.002

Benjamin, D. P., Mastergeorge, A. M., McDuffie, A. S., Kover, S. T., Hagerman, R. J., & Abbeduto, L. (2014). Effects of labeling and pointing on object gaze in boys with fragile X syndrome: An eye-tracking study. *Research in Developmental Disabilities, 35,* 2658–2672. doi:10.1016/j.ridd.2014.06.021

Brady, N. C., Anderson, C. J., Hahn, L. J., Obermeier, S. M., & Kapa, L. L. (2014). Eye tracking as a measure of receptive vocabulary in children with autism spectrum disorders. *Augmentative and Alternative Communication, 30,* 147–159. doi:10.3109/07434618.2014.904923

Brock, J., Norbury, C., Einav, S., & Nation, K. (2008). Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition, 108,* 896–904. doi:10.1016/j.cognition.2008.06.007

Burack, J. A., Iarocci, G., Bowler, D., & Mottron, L. (2002). Benefits and pitfalls in the merging of disciplines: The example of developmental psychopathology and the study of persons with autism. *Development and Psychopathology, 14,* 225–237.

Charman, T., Drew, A., Baird, C., & Baird, G. (2003). Measuring early language development in preschool children with autism spectrum disorder using the MacArthur Communicative Development Inventory (Infant Form). *Journal of Child Language, 30,* 213–236. doi:10.1017/S0305000902005482

Eigsti, I.-M., de Marchena, A. B., Schuh, J. M., & Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders, 5,* 681–691. doi:10.1016/j.rasd.2010.09.001

Elsabbagh, M., Fernandes, J., Webb, S. J., Dawson, G., Charman, T., & Johnson, M. H. (2013). Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood. *Biological Psychiatry, 74,* 189–194. doi:10.1016/j.biopsych.2012.11.030

Falck-Ytter, T., Bölte, S., & Gredebäck, G. (2013). Eye tracking in early autism research. *Journal of Neurodevelopmental Disorders, 5,* 28. doi:10.1186/1866-1955-5-28

Farris-Trimble, A., & McMurray, B. (2013). Test–retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research, 56,* 1328–1346. doi:10.1044/1092-4388(2012/12-0145)group-level

Farzin, F., Scaggs, F., Hervey, C., Berry-Kravis, E., & Hessl, D. (2011). Reliability of eye tracking and pupillometry measures in individuals with fragile X syndrome. *Journal of Autism and*

Developmental Disorders, 41, 1515–1522. doi:10.1007/s10803-011-1176-2

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development, 83,* 203–222.

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

Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the development of social attention using free-viewing. *Infancy, 17,* 355–375. doi:10.1111/j.1532-7078.2011.00086.x

Gillespie-Lynch, K., Elias, R., Escudero, P., Hutman, T., & Johnson, S. P. (2013). Atypical gaze following in autism: A comparison of three potential mechanisms. *Journal of Autism and Developmental Disorders, 43,* 2779–2792.

Gredebäck, G., Johnson, S., & von Hofsten, C. (2010). Eye tracking in infancy research. *Developmental Neuropsychology, 35,* 1–19. doi:10.1080/87565640903325758

Grieco-Calub, T. M., Saffran, J. R., & Litovsky, R. Y. (2009). Spoken word recognition in toddlers who use cochlear implants. *Journal of Speech, Language, and Hearing Research, 52,* 1390–1400. doi:10.1044/1092-4388(2009/08-0154).Spoken

Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews, 42,* 279–297. doi:10.1016/j.neubiorev.2014.03.013

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, United Kingdom: Oxford University Press.

Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders, 34,* 81–86. doi:10.1023/B:JADD.0000018078.82542.ab

Karatekin, C. (2007). Eye tracking studies of normative and atypical development. *Developmental Review, 27,* 283–348. doi:10.1016/j.dr.2007.06.006

Kedar, Y., Casasola, M., & Lust, B. (2006). Getting there faster: 18- and 24-month-old infants' use of function words to determine reference. *Child Development, 77,* 325–338.

Kelly, D. J., Walker, R., & Norbury, C. F. (2013). Deficits in volitional oculomotor control align with language status in autism spectrum disorders. *Developmental Science, 16,* 56–66. doi:10.1111/j.1467-7687.2012.01188.x

Kim, S. H., & Lord, C. (2010). Restricted and repetitive behaviors in toddlers and preschoolers with autism spectrum disorders based on the Autism Diagnostic Observation Schedule (ADOS). *Autism Research: Official Journal of the International Society for Autism Research, 3,* 162–173. doi:10.1002/aur.142

Kover, S. T. (2012). *Language and learning in boys with fragile X syndrome: Syntactic processing and the role of phonological memory* (Unpublished doctoral dissertation). University of Wisconsin–Madison.

Kover, S. T., & Atwood, A. K. (2013). Establishing equivalence: Methodological progress in group-matching design and analysis. *American Journal on Intellectual and Developmental Disabilities, 118,* 3–15. doi:10.1352/1944-7558-118.1.3

Kylliäinen, A., Jones, E. J. H., Gomot, M., Warreyn, P., & Falck-Ytter, T. (2014). Practical guidelines for studying young children with autism spectrum disorder in psychophysiological experiments. *Review Journal of Autism and Developmental Disorders, 1,* 373–386. doi:10.1007/s40489-014-0034-5

Martinez-Conde, S., Otero-Millan, J., & Macknik, S. L. (2013). The impact of microsaccades on vision: Towards a unified theory of saccadic function. *Nature Reviews Neuroscience, 14,* 83–96. doi:10.1038/nrn3405

McMurray, B., Munson, C., & Tomblin, J. B. (2014). Individual differences in language ability are related to variation in word recognition, not speech perception: Evidence from eye movements. *Journal of Speech, Language, and Hearing Research, 57,* 1344–1362. doi:10.1044/2014

Mervis, C. B., & Klein-Tasman, B. P. (2004). Methodological issues in group-matching designs: Alpha levels for control variable comparisons and measurement characteristics of control and target variables. *Journal of Autism and Developmental Disorders, 34,* 7–17.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.

Murphy, M. M., Abbeduto, L., Schroeder, S., & Serlin, R. (2007). Contribution of social and information-processing factors to eye-gaze avoidance in fragile X syndrome. *American Journal on Mental Retardation, 112,* 349–360. doi:0895-8017-112-5-349 [pii] 10.1352/0895-8017(2007)112[0349:COSAIF]2.0.CO;2

Naigles, L., Kelty, E., Jaffery, R., & Fein, D. (2011). Abstractness and continuity in the syntactic development of young children with autism. *Autism Research, 4,* 422–437. doi:10.1002/aur.223

Naigles, L., & Tovar, A. T. (2012). Portable intermodal preferential looking (IPL): Investigating language comprehension in typically developing toddlers and young children with autism. *Journal of Visualized Experiments: JoVE, 70,* e4331

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. doi:10.3758/s13428-012-0247-4

Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy, 15,* 1–5. doi:10.1111/j.1532-7078.2010.00030.x

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy, 17,* 1–8. doi:10.1111/j.1532-7078.2011.00101.x

Papagiannopoulou, E. A., Chitty, K. M., Hermens, D. F., Hickie, I. B., & Lagopoulos, J. (2014). A systematic review and meta-analysis of eye-tracking studies in children with autism spectrum disorders. *Social Neuroscience, 9,* 610–632. doi:10.1080/17470919.2014.934966

Piotroski, J., & Naigles, L. (2012). Intermodal preferential looking. In E. Hoff (Ed.), *Research methods in child language* (pp. 17–28). Oxford, United Kingdom: Wiley-Blackwell.

Sasson, N. J., & Elison, J. T. (2012). Eye tracking young children with autism. *Journal of Visualized Experiments: JoVE, 61,* 1–5. doi:10.3791/3675

Shic, F., Chawarska, K., & Scassellati, B. (2008). The amorphous fixation measure revisited: With applications to autism. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1–6). Austin, TX: Cognitive Science Society. doi:10.1.1.145.8123

Swensen, L. D., Kelley, E., Fein, D., & Naigles, L. R. (2007). Processes of language acquisition in children with autism: Evidence from preferential looking. *Child Development, 78,* 542–557.

Swingley, D. (2012). The looking-while-listening procedure. In E. Hoff (Ed.), *Research methods in child language* (pp. 29–41). Oxford, United Kingdom: Wiley-Blackwell.

Tenenbaum, E. J., Amso, D., Abar, B., & Sheinkopf, S. J. (2014). Attention and word learning in autistic, language delayed and typically developing children. *Frontiers in Psychology, 5,* 490. doi:10.3389/fpsyg.2014.00490

Venker, C. (2013). *Statistical word learning and non-social visual attention in children with autism* (Unpublished doctoral dissertation). University of Wisconsin–Madison.

Venker, C. E., Eernisse, E. R., Saffran, J. R., & Ellis Weismer, S. (2013). Individual differences in the real-time comprehension of children with ASD. *Autism Research, 6,* 417–432.

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., & 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

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science, 24,* 2143–2152. doi:10.1177/0956797613488145

Yoder, P., & Compton, D. (2004). Identifying predictors of treatment response. *Mental Retardation and Developmental Disabilities Research Reviews, 10,* 162–168.

Yoder, P. J., Molfese, D., & Gardner, E. (2011). Initial mean length of utterance predicts the relative efficacy of two grammatical treatments in preschoolers with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54,* 1170–1181.

Yoder, P., Stone, W. L., Walden, T., & Malesa, E. (2009). Predicting social impairment and ASD diagnosis in younger siblings of children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 39,* 1381–1391.

Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science, 14,* 165–180. doi:10.1111/j.1467-7687.2010.00958.x