Autism at a glance: A pilot study optimizing thin-slice observations

Lauren H. Hampton

Philip R. Curtis

Megan Y. Roberts

Northwestern University

Corresponding Author:
Lauren H. Hampton
2240 Campus Drive
Frances Searle, Room 3-365
Evanston, IL
60208
lauren.hampton@northwestern.edu

Abstract

Borrowing from a clinical psychology observational methodology, thin slice observations were used to assess autism characteristics in toddlers. Thin slices are short observations taken from a longer behavior stream which are assigned ratings by multiple raters using a 5-point scale. The raters' observations are averaged together to assign a 'thin-slice' value for each observation. In the current study, a total of 60 toddlers were selected from a video archive: 20 children with typical development, 20 children with developmental language disorder, and 20 children with autism. In this first part of this study, 20 raters observed small play segments between toddlers and an assessor. Raters assigned scores to each of the 60 toddlers on items related to autism symptomatology. Item analysis and Generalizability and Decision studies were conducted to determine the factor structure and optimal number of raters to achieve a stable estimate of autism characteristics. In the second part of the study, generalizability and decision studies were conducted to determine the most efficient and optimal combination of raters and naturalistic contexts. This pilot study provides recommendations for optimizing the utility of thin-slice observations for measuring autism symptomatology in young children.

Keywords: Autism spectrum disorders, screening, thin-slice

Autism at a glance: A pilot study optimizing thin-slice observations

Despite the US Preventative Services Task Force conclusion that there is no clear benefit from a universal autism screening, multiple agencies, parent groups, and the American Academy of Pediatrics (AAP) all concluded that identifying autism and providing early intervention services is a public health priority and identifying efficient and effective methods for optimizing screenings is necessary (Pierce, Courchesne, & Bacon, 2016). Nevertheless, pediatricians do not screen for autism red flags on a routine basis (Roux et al., 2012; Arunyanart et al., 2012) or readily follow the AAP implementation recommendations (Self et al., 2015), despite the fact that children with autism can be reliably identified before the age of 2 years (Cox et al., 1999; Kleinman et al., 2008; Lord et al., 2006). Families frequently wait between 1 to 3 years to receive an autism diagnosis (ADEP, 2017; Crane, Chester, Goddard, Henry, & Hill, 2016), which results in most children not receiving a diagnosis until after age 4 (Brett, Warnell, McConachie, & Parr, 2016; Soke, Maenner, Christensen, Kurzius-Spencer, & Schieve, 2017). Families often identify concerns about their child's communication development much younger than they receive a diagnosis (Bolton, Golding, Emond, & Steer, 2012; Kozlowski, Matson, Horovitz, Worley, & Neal, 2011). This is particularly problematic given that that earlier intervention is effective for children with autism (Zwaigenbaum et al., 2015) and, receiving a diagnosis is critical to gaining access to these services. Children who receive an autism diagnosis are more likely to access and benefit from early interventions and experience more positive outcomes over time (Suma, Adamson, Bakeman, Robins, & Abrams, 2016).

Access to a skilled diagnostician is one of the greatest barriers to receiving a diagnosis (Bisgaier, Levinson, Cutts, & Rhodes, 2011). Current screening practices rely primarily on parent report or short assessments with a novel person (Matson & Tureck, 2012; Robins, Fein, Barton, & Green, 2001; Wendy L. Stone, Coonrod, & Ousley, 2000). The Modified Checklist for Autism in Toddlers- Revised, a widely used screening tool adapted for many different cultures, relies on parent report (MCHAT-R; Robins et al., 2001). Although this measure may be a reasonable Level 1 screener, with sensitivity and specificity up to .90, it fails to be useful as a Level 2 screener, discriminating between children with autism and children with other disabilities, such as language delays. Similarly, the Baby and Infant Screen for Children with aUtIsm Traits (BISCUIT; Matson & Tureck, 2012) and Screening Tool for Autism in Toddlers (STAT; Stone, Coonrod, Turner, & Pozdol, 2004), have high sensitivity and moderate specificity but again may not differentiate between autism and other disabilities well when used as the only measure. Since access to a skilled diagnostician for autism is limited (Bisgaier et al., 2011), it is critical during the time between initial screening and diagnosis (level 2) to differentiate between children who show signs of autism and need an autism-specific diagnostic testing battery, from those children who show developmental delays but no signs of autism and require a comprehensive language and developmental assessment battery. Adding a simple naturalistic direct observation to the level two screener stage may improve positive predictive value (PPV) and increase specificity for referrals.

Thin-slice behavior observations are short samples from a behavior stream, such as a child-assessor interaction, that contain dynamic information (Ambady & Rosenthal, 1992). Ratings are made on a series of questions by multiple raters using a 5-point scale. The ratings are averaged across raters into one thin-slice score. Within this methodology, there are three parameters that can vary: (1) the *length of segments* that raters code; (2) the *number of raters* that are averaged together to create a thin-slice score; and (3) the *number of segments watched by each rater for each child*. Although segments that are less than 5 minutes are recommended, the optimal length and number of observations for ratings is unknown (Carney, Colvin, & Hall, 2007). It has been suggested that as few as 5 raters or as many as 150 raters are necessary to get a useful thin-slice rating. Additionally, raters from a more similar social group tend to rate more similarly than those from different social groups (Wang, Toosi, & Ambady, 2009; Wright & Drinkwater, 1997; Young & Hugenberg, 2010). The simplicity of thin-slice observations may make direct observation as a screening measure more feasible if the length and number of raters is optimized.

Ratings from thin-slice observations in psychology have been used to reliably rate a variety of behaviors and are valid for predicting meaningful outcomes. These ratings have been found to be reliable on a variety of behavioral ratings of depression, Parkinson’s disease, marital status or life satisfaction (Ambady & Gray, 2002; Bonanno & Keltner, 1997; Hatfield, Cacioppo, & Rapson, 1994; Pentland, Pitcairn, Gray, & Riddle Jr, 1987). Thin-slice ratings have also been used to predict important outcomes such as divorce, suicide attempts, and disease progression (Archinard, Haynal-Reymond, & Heller, 2000; Gottman & Levenson, 1992; Heller & Haynal, 1997; Hertenstein, Hansel, Butts, & Hile, 2009; Kleiman & Rule, 2013; Mason, Sbarra, & Mehl, 2010). More recently, studies have used thin-slice observations of typically developing children to measure personality traits (Tackett et al., 2017) and to measure outcomes for children with autism (Walton & Ingersoll, 2016).

There is a clinical bottleneck in autism diagnostics, and although there are useful models across the country, the diagnostic process is still taking longer than it should and families are waiting too long for services that depend on a diagnosis (Gordon-Lipkin, Foster, & Peacock, 2016). By optimizing the efficiency and effectiveness of the screening process, we can reduce clinical wait-time and better prioritize referrals for full diagnostic evaluations. Some recent advancements in technology have led to multiple recommendations to use technology to address these needs and maximize autism discoveries. For example, crowdsourcing is one method that may accelerate our discoveries in autism (David, Babineau, & Wall, 2016).” Additionally, researchers have utilized content analysis of public videos to reliability detect autism (Fusaro et al., 2014), and easy-to-use mobile apps have been developed to collect videos from families for diagnostic purposes (Illingworth, Thomas, Rozga, & Smith, 2017; J. Smith et al., 2016; Knutsen et al., 2016; Nazneen et al., 2015). Despite the potential of these innovative video-based behavioral observation methods to improve the autism diagnostic process, there is not clear evidence indicating how to best gain a stable estimate of autism symptoms from these videos. We propose that by applying the thin-slice observation methodology to these existing observational methodologies, the autism diagnostic process may be improved.

Current study

The current study aim of the study is to provide a preliminary assessment of the stability of multiple thin-slice ratings, provide an initial optimization of this tool for developmental screenings, and provide a set of recommendations to the field when considering thin-slice video observations as a diagnostic tool. To accomplish this aim, we conducted a two-part study to answer the following research questions: Study 1: (1) What is the dimensionality, validity, and reliability of the proposed thin-slice rating measure? Study 2: (2) What is the most efficient combination of raters and contexts to achieve a stable thin-slice value for autism symptomatology in toddlers? (3) Using an optimized thin-slice score, how does the thin-slice observation relate to similar and dissimilar measures? And finally, an exploratory question; (4) Does the optimized thin-slice value result in a meaningful positive predictive value for identifying children with autism?

Methods

Participants

A subset of participants from a larger randomized controlled trial, were used in this retrospective study (NCT02632773). Toddlers in this larger trial were included in part by not exceeding the clinical cut off score on the Screening Tool for Autism in Toddlers at intake into the study (STAT; Stone & Ousley, 1997), yet throughout the 15-month period of the RCT study, 20 participants were identified as having ASD through clinical judgement, parent report, and administration of the Autism Diagnostic Observation Scale (ADOS; Lord, Rutter, Dilavore, & Risi, 2008). As such, these participants' autism symptoms were more subtle (mean ADOS severity score of 6.45, standard deviation of 1.88); this means that these children, who did not flag on a level-2 screening measure but *did* go on to receive a diagnosis of ASD, are likely the most difficult children to screen. All 20 participants diagnosed with ASD during the course of the study were included and were matched with 20 toddlers with developmental language disorder (DLD) and 20 typically developing toddlers on age, gender, and receptive language ability from the same trial. Children were matched from different single time points based on age in months, receptive/expressive language skills (for children with autism or DLD alone) and on gender. Participants are described in detail in Table 1. All assessments came from one time point within each participant, but different time points were selected across participants to improve age-matching.

Raters

Raters were undergraduate or graduate speech-language pathology students or master's level research staff. Raters reported how much experience they had caring for, observing, or working with toddler-aged children, and how much experience they had rating behavioral observations. Raters had moderate experience coding videos for language and behavior, but all were new to this method of rating and naïve to the study's purpose and to each child's diagnosis. In Study 1, 20 raters independently rated each video using the thin-slice rating scale described in Table 3. In Study 2, a subset of 5 raters independently rated each video using the thin-slice rating scale across 2 novel observational contexts. All raters are described in Table 2. Raters received no formal training in the measure other than definitions and examples provided in order to benefit from the thin-slice logic relying on initial impressions.

Measures

Assessments. Participants were assessed by a master's level clinician who was unfamiliar with the study purpose and hypothesis. The ADOS (Lord, Rutter, Dilavore, & Risi, 2008) and Preschool Language Scale-4th edition (PLS-4; Zimmerman, Steiner, & Pond, 2002) were administered concurrently with thin-slice observations. The Autism Severity score, on a scale from 1-10, was calculated from the ADOS so that scores could be compared across modules. Scores from the ADOS are scored in a reverse manner, such that a lower score indicates a more typical performance. Additionally, scores from the STAT (Stone & Ousley, 1997) were collected at entry into the study when participants were 2-years-old.

A 20-minute semi-structured and standardized naturalistic language sample was used to observe spontaneous communication with a novel assessor in a clinic setting. Assessors, unfamiliar with the child, played with the child using a standard set of toys (a farm set, car ramp and cars, baby doll set, play dough set, play kitchen, a book, and bubbles) and using no specific vocabulary (Miller, 1981). The adult imitated any language the child used, responded using utterances without specific language (uh-huh, wow, uhoh), and engaged in play with the child. Alternatively, parent-child interactions included a different standard set of toys (doll house, ball, school bus set, book, puzzle, and blocks) and parents were instructed to play and communicate with their child as they usually do (Roberts & Kaiser, 2012). Language samples and parent-child interactions were video recorded and transcribed and coded for number of different words (NDW) spontaneously produced by the child. Fidelity of administration and verification of standardized scoring was assessed on 20% of administrations and exceeded 90% across all measures. Twenty percent of the language samples were coded by a second coder, and point-by-point agreement exceeded 90% across language samples.

Thin slices. Thin-slice observations were created by selecting a 2-minute segment from each of three contexts: the play portion of the ADOS module 1 and 2, a naturalistic language sample conducted by an unfamiliar assessor that mirrors the play portion of the ADOS, and a parent-child, play-based interaction. Three different contexts were selected to retain independence of the samples; the first interaction was selected to validate the rating scale, and the second two interactions were selected to determine stability and validity of thin-slice scores across contexts and for greatest applicability as a screening measure. Two-minute segments were selected starting at 3 minutes into the original video recording, or the next 2-minutes of uninterrupted free-play, whichever came first. Interruptions included 5 seconds or longer of the child or adult being off camera or segments where the adult was speaking to someone other than the child.

Raters were randomly assigned a unique assignment order such that no two raters observed the videos in the same sequence. The raters viewed each 2-minute segment without pausing or rewinding the video. Raters assigned codes for each video on 11 unique items, rated on a 5-point Likert scale (see Table 3). These ratings from all 11 items were averaged together so that each rater had one thin-slice score for each video. In Study 1, raters observed the thin-slice from the ADOS; in Study 2, the raters observed the language sample and parent interaction videos intermixed together. Each item in the 11-item thin-slice rating scale was defined and raters were instructed to rate the quality of the skill or behavior they observed in the child on the anchored scale from 1-5 (Table 3). The raters were also instructed to use clinical judgement based on experiences with typically developing 2 to 3-year-olds.

Thin slice rating items. The 11 unique items were derived from multiple sources for this preliminary study. We considered 10. First, the 5-item scale used by Walton and Ingersoll (2016) to rate communicative behaviors in children with ASD were replicated (items 1-5). Second, additional behaviors that may be relevant to discriminate children with DLD from children with ASD were added (items 6-8). Third, three items were created using unique variables from the ADOS and reworded to fit with the other items in the scale (items 9-11). Raters were instructed to rate each item based on how much they agreed that the statement was either characteristic of the child in the short clip (1=strongly disagree/never observed, 2=disagree/observed <5% of observation, 3=neither agree or disagree/observed <50% of observation, 4=agree/observed >50% of observation, 5=strongly agree and >80% of observation).

Analysis

Item reliability and validity. To determine the structure of the 11 items used to code the thin slices, average ratings from 20 raters on 60 slices from the ADOS were used in a factor analysis. All analyses were run in R using the psych (Revelle, 2017) and paran packages (Dinno, 2009). In order to determine the number of factors within the scale, Horn's Parallel Analysis was used and results supported the retention of only one factor (Horn, 1965). Likewise, a Very Simple Structure analysis also suggested that only one factor was supported from the current data (Revelle & Rocklin, 1979). A factor analysis was run to extract one factor, estimated using the minimum residual method. The solution was subjected to an oblimin rotation. All items loaded highly on this one extracted factor (factor loadings are given in Table 3). These results support the unidimensionality of the items used in coding the thin slices. Reliability of the scale was assessed using both Chronbach's α, as well as the more conservative omega-hierarchical coefficient (ω_h; Zinbarg, Revelle, Yovel, & Li, 2005). Omega-hierarchical uses a hierarchical factor model, in which both a general factor and lower-order factors can account for variance in observed scores; the ω_h coefficient represents the general factor saturation.

G-study. Generalizability theory (G-theory) allows us to determine how much variability across facets (e.g. observations, contexts, observers) of a study is due to true score relative to error (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). A g-coefficient uses the partitioned variance attributed to each measurement facet, similar to an interclass correlation, to give an estimate of measurement stability. Specifically, the g-coefficient is calculated using components form a traditional analysis of variance (ANOVA; Equation 1). This ratio separates variance due to the individual person from the total combined variance components of the person, person by rater, person by session and total combined error. Increased stability in measurement increases confidence that the measure is representative and therefore more likely generalizable, thus increasing external validity. A G-coefficient between 0.6 and 0.8 is considered an acceptable level of stability (Bakeman, McArthur, Quera, & Robinson, 1997).

D-study. While a G-study results in an estimate of achieved stability, a determination study (D-study) is used to estimate the expected achieved g-coefficient with increased samples (Shavelson, Webb, & Rowley, 1989; Equation 2). A D-study uses the fixed variance from the G-study and applies it to additional samples, and therefore assumes the variability attributed to the original samples remains stable over time. The relative error variance (σ^2_{rel}) is a division of the individual variances by different combinations hypothesized sample sizes for number of sessions and raters (see formula 2 and the calculator available from Webb & Shavelson, 2005). A D-study allows the researcher to optimize a measure by estimating the ideal measurement samples necessary to achieve an optimal G-coefficient. For example, when considering how many different raters are needed to achieve a stable G-coefficient, a D-study may estimate 2 raters at g=.59, 3 raters at g=.72 and 4 raters at g=.80. Since 3 raters and 4 raters both fall in the recommended range, the research can select to average 3 rater observations together to optimize resources and outcomes. All G and D-studies were completed using the EduG software (*EduG*, 2012).

Equations.

Equation 1. $g = \frac{\sigma_c^2}{\sigma_c^2+\sigma_{cr}^2+\sigma_{cs}^2+\sigma_{crs.,e}^2}$;

Equation 2. D study: $\sigma_{rel}^2 = \left(\frac{\sigma_{cr}^2}{n_r'}\right) + \left(\frac{\sigma_{cs}^2}{n_s'}\right) + \left(\frac{\sigma_{crs,e}^2}{n_r' n_s'}\right)$

$\sigma^2 =$ variance components; where c= child level variance components, r= raters, and s= sessions, e= error variance, n= sample size (Yoder &. Symons, 2010, p.29).

Measurement validity. Once an optimal G-coefficient was identified, the optimal number of raters and contexts of the thin-slice were averaged together to result in an optimized thin-slice score. Correlations between thin-slice scores and measures of similar constructs (language and ASD measures) were estimated to demonstrate validity. Additionally, thin-slice scores were correlated with divergent variables (the child’s age, and socioeconomic status) expected to not be highly correlated with the outcome to further increase validity of the measure.

Screening utility. To estimate the utility of the optimized thin-slice score as a screening tool, the positive predictive and negative predicative values were calculated to estimate the utility of the thin-slice measure in correctly categorizing toddlers with and without autism as compared to children with DLD. Sensitivity, specificity, negative predictive value, and positive predictive value were also calculated across all three populations to compare the utility to other screeners. These likelihood ratios consider the proportion of accurately classified children with autism to those accurately classified as not having autism. A ROC curve analysis, with bootstrapped confidence intervals, was used to estimate the best cut-score for optimizing sensitivity, specificity, NPV and PPV using pROC package in R (Robin et al., 2011).

Results

Study 1.

To determine the structure of the 11 items used to code the thin slices, average ratings from 20 raters on 60 slices from the ADOS were used in a factor analysis. All analyses were run in R using the psych (Revelle, 2017) and paran packages (Dinno, 2009). In order to determine the number of factors within the scale, Horn's Parallel Analysis was used and results supported the retention of only one factor (Horn, 1965). Likewise, a Very Simple Structure analysis also suggested that only one factor was supported from the current data (Revelle & Rocklin, 1979). When a factor analysis was run with one factor, all items loaded highly on this one factor (all factor loadings > .62). These results support the unidimensionality of the items used in coding the thin slices.

Reliability of the scale was assessed using both Chronbach's α, as well as the more conservative omega-hierarchical coefficient (ω_h;Zinbarg, Revelle, Yovel, & Li, 2005). Omega-hierarchical models a hierarchical factor model, in which both a general and item-specific factors can account for variance in observed scores; the ω_h coefficient represents the degree to which items load onto the *general factor*. Both α and ω_h estimates suggested excellent reliability ($\alpha = .95$, $\omega_h = .82$). In order to test the validity of the new scale in assessing autism symptomatology, average sum scores across raters from the 60 thin slices were correlated with the STAT at baseline (r=0.57) and time-point specific ADOS severity scores (r=0.35). Correlations were moderate and significant and support the validity of the new scale.

Using the initial 60 ratings from the ADOS, a single facet G-study was conducted to optimize the number of raters required to establish a stable estimate of the thin-slice average score for each child. The single facet G-study resulted in an absolute G-coefficient of 0.95 when using 20 raters. Interestingly, the D-study, which uses a bootstrap-like resampling method to generalize to larger samples, found that an optimal rating could be achieved with as few as 2 (*g*=0.69) and 5 (*g*=0.82) raters of a semi-structured observation. Due to the subjectivity of the rating and the potential for increased variability when more observations are introduced, 5 raters were selected for study 2 to ensure maximal optimization.

Study 2.

Optimized across contexts. The two facet G-study, used to determine the optimal number of contexts and raters, resulted in an optimal stable estimate when 5 raters rated 2 novel naturalistic contexts (g=0.73). This was a large improvement over the G-coefficient obtained when 5 raters rated only one naturalistic context (g=0.60). During the second study, two new contexts were selected to reduce bias, yet greater instability was observed when 5 raters were used as compared to Study 1. Therefore, we selected the most optimized measure that promised the greatest generalizability across naturalistic contexts and averaged the 5 raters' scores across both observations (language sample and parent-child interaction), using the optimized thin-slice score for the subsequent analyses. Additionally, the language sample and parent-child interaction have the most utility as a screening measure as opposed to the ADOS thin-slice.

Correlations. The correlations between the optimized thin-slice score and related measures of language and autism were moderate (r=-0.58- -0.47, p<0.05; See Table 4). Even though NDW also significantly related to the optimized thin-slice score (r=-0.59, p<0.05), NDW and autism severity were not significantly correlated with one another (r=-0.23, p>0.1). Divergent scores (SES, age) were not significantly or meaningfully correlated with the optimized thin-slice score (r=-0.11, -0.02, p>0.1).

Positive Predictive Value. A ROC curve analysis was used to estimate an optimized cut score of 34.3 (AUC=0.82, Robin et al., 2011). This cut score resulted in high sensitivity and specificity for distinguishing autism from DLD and typical development (Table 6). Additionally, moderate positive predictive value and negative predictive value resulted from identifying children with autism from DLD alone (see Table 6). Thus, the proposed thin-slice rating tool may be a useful tool towards estimating autism in young children, however this may be further optimized when combined with additional level-2 screening measures.

Discussion

The thin-slice measure resulted in stable estimates across a relatively few number of raters and contexts. Using short observations of naturalistic interactions between young children and their parents and a short interaction with an assessor may provide a useful perspective on the child's development and risk for autism. Optimized thin-slice ratings were significantly correlated with the severity score from the ADOS, the gold-standard (but time-intensive) assessment tool in autism diagnostics, and the STAT measure for screening. Unrelated variables were not correlated with the thin-slice score (age, family income). However, the relationship between the thin-slice scores and language measures may indicate that the thin-slice scores may overemphasize language and communication skills. Finally, the optimized thin-slice score was relatively effective in identifying children with autism as compared to children with DLD (PPV = .70; NPV=0.83) in this preliminary study. This PPV is interesting given that the children in this study had lower autism severity scores and were compared to children with DLD, a comparison that is often difficult for differential diagnosis (Bishop, Dorothy V.M. & Norbury, Courtenay Frazier, 2002; Loucas et al., 2008). Although this PPV is not high enough to recommend using the thin-slice score as a screener alone, it suggests that the thin-slice score could add meaningful value to a battery of screening assessments for children suspected of having autism.

Implications

This study is an important first step towards optimizing the thin-slice measure as a useful tool for characterizing autism symptomatology in young children. One of the important results of this study is the practicality of utilizing the thin-slice methodology in clinical practice for toddlers with autism. Only five raters' scores on two 2-minute contexts (4 minutes per coder, 20 minutes of total coding time across all raters) were needed to arrive at a stable estimate of children's social communication scores. This means that utilizing the thin-slice scoring methodology may be an efficient and cost-effective measurement tool. Although the thin-slice is an interesting new measure, it is important to understand how to best optimize and utilize a measure that relies on short observations to maximize generalizability.

Limitations

The results of this study must be considered in light of three primary limitations. First, the ratings were conducted by well-trained research staff who are familiar with rating and coding child communicative behavior. The results of the G-study and D-studies may only generalize to other raters with experience working with young children. Second, the ratings during Study 2 came from different contexts than Study 1. Although all contexts were play-based, the variability in G-coefficients across contexts suggests further replication is necessary to provide a truly optimized measure. Third, this study used a video archive, and was limited by the participants available. A prospective screening study is necessary to replicate this work.

Next Steps

Thin-slice rating requires the use of a multi-item, Likert-type scale. The current study used a novel, 11-item scale for this purpose. The scale showed strong reliability and acceptable validity within the current sample. The 11 items in this study, though not exhaustive, were selected to replicate previous work using thin-slice measures, and to provide a context for optimizing the thin-slice approach. Further expansion and refinement of the 11-item scale may result in improved predictive validity. Additionally, to further validate the thin-slice observation scale, a predictive validity study should be conducted. Additionally, the G- and D-studies should be replicated to increase the optimization and generalizability across populations and contexts. The accuracy of ratings for less experienced raters should also be examined in order to determine the needed qualifications and experience for reliable thin-slice coding in clinical practice. Future studies should consider the thin-slice measure along with existing screening tools and the additive value of this observational measure. Finally, the thin-slice measure should be explored for other important uses, such as an outcome measure for intervention studies aiming to reduce autism symptomatology and increase social communication.

Conclusions

This study provides preliminary support for the use of thin-slice coding procedures using short, 2-minute video-taped interactions in order to assess children’s social communication and autism symptomatology. Short segments from a behavioral stream, or thin-slices, can be optimized and reliability rated when averaged across 5 raters and 2 contexts. The thin-slice observation is an exciting and useful tool for making snap judgement that may be effective for rating autism characteristics. Although additional validation is needed, this short observational measure may be an effective tool to increase efficiency in autism screenings.

References

ADEP. (2017). Ohio Center for Autism and Low Incidence | Autism Diagnosis Education Project. Retrieved November 14, 2017, from http://www.ocali.org/project/adep

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology*, *83*(4), 947.

Ambady, N., & Rosenthal, R. (1992). *Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.* American Psychological Association.

Archinard, M., Haynal-Reymond, V., & Heller, M. (2000). Doctor's and patients' facial expressions and suicide reattempt risk assessment. *Journal of Psychiatric Research*, *34*(3), 261–262.

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*(4), 357.

Bayley, N. (2006). Bayley scales of infant and toddler development. Pearson.

Bisgaier, J., Levinson, D., Cutts, D. B., & Rhodes, K. V. (2011). Access to Autism Evaluation Appointments With Developmental-Behavioral and Neurodevelopmental Subspecialists. *Archives of Pediatrics & Adolescent Medicine*, *165*(7), 673–674. https://doi.org/10.1001/archpediatrics.2011.90

Bishop, Dorothy V.M., & Norbury, Courtenay Frazier. (2002). Exploring the boarderlands of autisitic disorder and specific language impairment: A study using standardised diagnostic instruments. *Journal of Child Psychology and Psychiatry*, *43*(7), 917–929.

Bolton, P. F., Golding, J., Emond, A., & Steer, C. D. (2012). Autism spectrum disorder and autistic traits in the Avon Longitudinal Study of Parents and Children: precursors and early signs. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(3), 249–

260.e25. https://doi.org/10.1016/j.jaac.2011.12.009

Bonanno, G. A., & Keltner, D. (1997). Facial expressions of emotion and the course of conjugal bereavement. *Journal of Abnormal Psychology*, *106*(1), 126.

Brett, D., Warnell, F., McConachie, H., & Parr, J. R. (2016). Factors Affecting Age at ASD Diagnosis in UK: No Evidence that Diagnosis Age has Decreased Between 2004 and 2014. *Journal of Autism and Developmental Disorders*, *46*(6), 1974–1984. https://doi.org/10.1007/s10803-016-2716-6

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*(5), 1054–1072.

Cox, A., Klein, K., Charman, T., Baird, G., Baron-Cohen, S., Swettenham, J., … Wheelwright, S. (1999). Autism spectrum disorders at 20 and 42 months of age: Stability of clinical and ADI-R diagnosis. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(5), 719–732.

Crane, L., Chester, J. W., Goddard, L., Henry, L. A., & Hill, E. (2016). Experiences of autism diagnosis: A survey of over 1000 parents in the United Kingdom. *Autism*, *20*(2), 153–162. https://doi.org/10.1177/1362361315573636

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of measurements*. New York: John Wiley & Sons.

David, M. M., Babineau, B. A., & Wall, D. P. (2016). Can we accelerate autism discoveries through crowdsourcing? *Research in Autism Spectrum Disorders*, *32*(Supplement C), 80–83. https://doi.org/10.1016/j.rasd.2016.09.001

EduG. (2012). (Version EduG version 6.1-e, generalizability study) [Société Suisse pour la Recherche en Éducation, Groupe de travail Edumétrie - Qualité de l’évaluation en éducation]. Longueuil, Qc: software prepared by Maurice Dalois and Léo Laroche.

Fusaro, M., Vallotton, C. D., & Harris, P. L. (2014). Beside the point: Mothers' head nodding and shaking gestures during parent–child play. *Infant Behavior & Development*, *37*(2), 235–247. https://doi.org/10.1016/j.infbeh.2014.01.006

Gordon-Lipkin, E., Foster, J., & Peacock, G. (2016). Whittling Down the Wait Time. *Pediatric Clinics of North America*, *63*(5), 851–859. https://doi.org/10.1016/j.pcl.2016.06.007

Gottman, J. M., & Levenson, R. W. (1992). Marital processes predictive of later dissolution: behavior, physiology, and health. *Journal of Personality and Social Psychology*, *63*(2), 221.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). Emotional contagion: Cambridge studies in emotion and social interaction. *Cambridge, UK: Cambridge University Press. Errors-in-Variables Regression Model When the Variances of the Measurement Errors Vary between the Observations. Statistics in Medicine*, *21*, 1089–1101.

Heller, M., & Haynal, V. (1997). Depression and suicide faces. *What the Face reveals'(Oxford, 1997)*, 398–407.

Hertenstein, M. J., Hansel, C. A., Butts, A. M., & Hile, S. N. (2009). Smile intensity in photographs predicts divorce later in life. *Motivation and Emotion*, *33*(2), 99–105.

Illingworth, D. A., Thomas, R. P., Rozga, A., & Smith, C. J. (2017). Cue Use in Distal Autism Spectrum Assessment: A Lens Model Analysis of the Efficacy of Telehealth Technologies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 170–170. https://doi.org/10.1177/1541931213601526

J. Smith, C., Rozga, A., Matthews, N., Oberleitner, R., Nazneen, N., & Abowd, G. (2016). *Investigating the Accuracy of a Novel Telehealth Diagnostic Approach for Autism Spectrum Disorder* (Vol. 29). https://doi.org/10.1037/pas0000317

Kleiman, S., & Rule, N. O. (2013). Detecting suicidality from facial appearance. *Social Psychological and Personality Science*, *4*(4), 453–460.

Kleinman, J. M., Ventola, P. E., Pandey, J., Verbalis, A. D., Barton, M., Hodgson, S., … Fein, D. (2008). Diagnostic Stability in Very Young Children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *38*(4), 606–615. https://doi.org/10.1007/s10803-007-0427-8

Knutsen, J., Wolfe, A., Burke, B. L., Hepburn, S., Lindgren, S., & Coury, D. (2016). A Systematic Review of Telemedicine in Autism Spectrum Disorders. *Review Journal of Autism and Developmental Disorders*, *3*(4), 330–344. https://doi.org/10.1007/s40489-016-0086-9

Kozlowski, A. M., Matson, J. L., Horovitz, M., Worley, J. A., & Neal, D. (2011). Parents' first concerns of their child's development in toddlers with autism spectrum disorders. *Developmental Neurorehabilitation*, *14*(2), 72–78.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of General Psychiatry*, *63*(6), 694–701.

Lord, C., Rutter, M., Dilavore, P. C., & Risi, S. (2008). *ADOS: Autism diagnostic observation schedule*. Hogrefe Boston.

Loucas, T., Charman, T., Pickles, A., Simonoff, E., Chandler, S., Meldrum, D., & Baird, G. (2008). Autistic symptomatology and language ability in autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, *49*(11), 1184–1192.

Mason, A. E., Sbarra, D. A., & Mehl, M. R. (2010). Thin-slicing divorce: Thirty seconds of information predict changes in psychological adjustment over 90 days. *Psychological*

Science, *21*(10), 1420–1422.

Matson, J. L., & Tureck, K. (2012). Early diagnosis of autism: Current status of the Baby and Infant Screen for Children with aUtIsm Traits (BISCUIT-Parts 1, 2, and 3). *Research in Autism Spectrum Disorders*, *6*(3), 1135–1141. https://doi.org/10.1016/j.rasd.2012.02.009

Miller, J. F. (1981). *Assessing language production in children: Experimental procedures* (Vol. 1). Univ Park Press.

Nazneen, N., Rozga, A., Smith, C. J., Oberleitner, R., Abowd, G. D., & Arriaga, R. I. (2015). A novel system for supporting autism diagnosis using home videos: Iterative development and evaluation of system design. *JMIR mHealth and uHealth*, *3*(2).

Pentland, B., Pitcairn, T. K., Gray, J. M., & Riddle Jr, W. (1987). The effects of reduced expression in Parkinson's disease on impression formation by health professionals. *Clinical Rehabilitation*, *1*(4), 307–312.

Pierce, K., Courchesne, E., & Bacon, E. (2016). To Screen or Not to Screen Universally for Autism is not the Question: Why the Task Force Got It Wrong. *The Journal of Pediatrics*, *176*, 182–194. https://doi.org/10.1016/j.jpeds.2016.06.004

Revelle W, & Rocklin, T (1979) Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behavioral Research 14:403-414

Roberts, M. Y., & Kaiser, A. P. (2012). Assessing the effects of a parent-implemented language intervention for children with language impairments using empirical benchmarks: A pilot study. *Journal of Speech, Language, and Hearing Research*, *55*(6), 1655-1670.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 77.

Roux, A. M., Herrera, P., Wold, C. M., Dunkle, M. C., Glascoe, F. P., & Shattuck, P. T. (2012). Developmental and Autism Screening Through 2-1-1. *American Journal of Preventive Medicine*, *43*(6 0 5), S457–S463. https://doi.org/10.1016/j.amepre.2012.08.011

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922.

Soke, G. N., Maenner, M. J., Christensen, D., Kurzius-Spencer, M., & Schieve, L. A. (2017) Brief Report: Estimated Prevalence of a Community Diagnosis of Autism Spectrum Disorder by Age 4 Years in Children from Selected Areas in the United States in 2010: Evaluation of Birth Cohort Effects. *Journal of Autism and Developmental Disorders*, *47*(6), 1917–1922. https://doi.org/10.1007/s10803-017-3094-4

Stone, W. L., & Ousley, O. Y. (1997). STAT Manual: Screening tool for autism in two-year-olds. *Unpublished Manuscript, Vanderbilt University*.

Suma, K., Adamson, L. B., Bakeman, R., Robins, D. L., & Abrams, D. N. (2016). After early autism diagnosis: Changes in intervention and parent–child interaction. *Journal of Autism and Developmental Disorders*, *46*(8), 2720–2733.

Tackett, J. L., Smack, A. J., Herzhoff, K., Reardon, K. W., Daoud, S., & Granic, I. (2017). Measuring child personality when child personality was not measured: Application of a thin-slice approach. *Personality and Mental Health*, *11*(1), 4–13.

Walton, K. M., & Ingersoll, B. R. (2016). The utility of Thin Slice ratings for predicting language growth in children with autism spectrum disorder. *Autism*, *20*(3), 374–380.

Wang, E. J., Toosi, N. R., & Ambady, N. (2009). Nonverbal dialects: Culture and person perception. *Understanding Culture: Theory, Research, and Application*, 289–98.

Webb NM, Shavelson RJ. Generalizability theory: Overview. In: Everitt BS, Howell DC, editors. Encyclopedia of statistics in behavioral science. Vol. 2. 2005. pp. 717–719.

Wright, J. C., & Drinkwater, M. (1997). Rationality vs. accuracy of social judgment. *Social Cognition*, *15*(4), 245–273.

Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. Springer Publishing Company.

Young, S. G., & Hugenberg, K. (2010). Mere social categorization modulates identification of facial expressions of emotion. *Journal of Personality and Social Psychology, 99*(6), 964.

Zinbarg RE, Revelle, W, Yovel, I, & Li, W (2005) Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika 70:123-133

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *PLS-4: Preschool Language Scale* (Fourth Edition). San Antonio, TX: The psychological corporation.

Zwaigenbaum, L., Bauman, M. L., Choueiri, R., Kasari, C., Carter, A., Granpeesheh, D., … Fein, D. (2015). Early intervention for children with autism spectrum disorder under 3 years of age: recommendations for practice and research. *Pediatrics*, *136*(Supplement 1), S60–S81.

Table 1

Participant description

Mean (SD)	Typical	DLD	ASD
Age in months	30.0 (4.6)	31.6	32.9
% male	90%	90%	90%
% minority	35%	20%	25%
Autism Severity score	1.15 (0.49)	1.25 (0.44)	6.45 (1.88)
PLS, Total Language	115.95 (17.0)	74.8 (15.1)	72.6 (16.0)
Family income	$89,236 (64,306)	$76,336 (68,957)	$63,501 (23,344)
NDW	86.5 (36.9)	17.0 (21.6)	37.9 (40.4)
IQ	101.25 (9.1)	93.25 (9.9)	86.5 (8.6)

Note. DLD: Developmental Language Disorder; ASD: Autism Spectrum Disorder. Minority status indicates families who reported race as Black/African American, Asian, or Other and families who indicated Hispanic ethnicity. Autism Severity score, higher score indicating more symptoms of autism, calculated from the Autism Diagnostic Observation Scale, Lord et al., 2008. PLS: Preschool Language Scale 4th edition, Zimmerman, & Pond, 2002. NDW: Number of Different Words from a language sample. IQ: Intelligence Quotient from the Bayley Scales of Infant and Toddler Development, Bayley, 2006.

Table 2

Rater description

	Study 1: 20-raters	Study 2: 5-raters
Mean age (SD)	24.4 (4.1)	26.4 (4.3)
Male	15%	0%
Minority	10%	0%
Some graduate school or higher	45%	60%
3 years or more experience with toddlers	40%	80%
Experience rating videos	75%	80%

Table 3
Thin slice average ratings

Item	Factor loading	Typical	DLD	ASD
1. Shows interest in the adult	0.93	2.2 (0.8)	2.5 (0.5)	3.4 (0.8)
2. Plays appropriately with toys	0.80	1.9 (0.7)	2.1 (0.4)	2.5 (0.5)
3. Uses language appropriately	0.80	2.2 (0.9)	3.2 (0.7)	3.1 (0.8)
4. Uses appropriate gestures in play/communication	0.89	2.4 (0.8)	2.7 (0.7)	3.2 (0.6)
5. Imitates actions modeled by the adult	0.63	2.7 (0.8)	2.8 (0.4)	3.2 (0.3)
6. Behaves appropriately (does not engage in problem/disruptive behaviors)	0.66	1.5 (0.2)	1.8 (0.5)	1.8 (0.5)
7. Uses many different speech sounds	0.70	2.5 (1.2)	3.9 (0.9)	3.4 (1.1)
8. Understands what the adult says	0.91	2.0 (0.7)	2.5 (0.5)	2.8 (0.6)
9. Coordinates 2 forms of communication (eye contact, gestures, vocalizations or words)	0.90	2.2 (0.9)	2.6 (0.8)	3.2 (0.9)
10. Uses appropriate eye contact	0.82	2.3 (0.8)	2.5 (0.7)	3.4 (0.7)
11. Engages in unusual behaviors	0.90	2.1 (0.5)	2.5 (0.6)	3.0 (0.4)
Total thin-slice score		**24.2 (7.7)**	**28.6 (5.5)**	**33.0 (6.6)**

Note. All items are scored such that lower scores are closer to typical development and higher scores indicate greater impairment.
DLD: Developmental Language Delay, ASD: Autism Spectrum Disorder

Table 4. Correlations between measures and thin slice score

	Age, months	Family income	ADOS: Severity[1]	PLS: Total standard	NDW	Total Slice score
Age, months	1					
Family income	-0.03	1				
ADOS: Severity[1]	0.03	-0.09	1			
PLS: Total standard	-0.10	0.21	-0.45*	1		
NDW	0.35*	0.16	-0.23	0.77*	1	
Total Slice score	-0.07	-0.09	0.47*	-0.59*	-0.53*	1

Note. [1]*Higher scores indicates greater autism severity; *<.05; ADOS: Autism Diagnostic Observation Scale, Lord et al., 2008; PLS: Preschool Language Scale 4th edition, Zimmerman, Steiner, & Pond, 2002; NDW: Number of Different Words;*

Table 5.

Number of participants flagged as ASD-risk using 34.3 cut-score

		Slice rating	
		ASD-risk	No ASD-risk
Clinical diagnosis	ASD	16	4
	DLD	7	13
	Typical	0	20

Note: ASD: Autism Spectrum Disorder, DLD: Developmental Language Delay

Table 6.

Diagnostic accuracy (95% bootstrapped confidence interval)

	Sensitivity	Specificity	PPV	NPV
ASD vs DLD	0.80 [0.60-0.95]	0.65 [0.45-0.85]	0.70 [0.57-0.84]	0.76 [0.60-0.94]
ASD vs DLD + Typical	0.80 [0.60-0.95]	0.83 [0.70-0.93]	0.70 [0.56-0.86]	0.89 [0.81-0.97]

Note. 95% Confidence intervals bootstrapped with 10,000 stratified replicates. ASD: Autism Spectrum Disorder, DLD: Developmental Language Delay, PPV: Positive Predictive Value, NPV: Negative Predictive Value