HHS Public Access

Author manuscript

Autism. Author manuscript; available in PMC 2022 October 01.

Published in final edited form as:

Autism. 2021 October ; 25(7): 2101–2111. doi:10.1177/13623613211015003.

Measuring Parent Strategy Use in Early Interventions: Reliability and Validity of the NDBI-Fi Across Strategy Types

Bailey J. Sone[1], **Aaron J. Kaat**[2], **Megan Y. Roberts**[1]

[1]Department of Communication Sciences and Disorders, Northwestern University

[2]Department of Medical Social Sciences, Northwestern University

Abstract

Children with autism spectrum disorder benefit from early, intensive interventions to improve social communication, and parent-implemented interventions are a feasible, family-centered way to increase treatment dosage. The success of such interventions is dependent on a parent's ability to implement the strategies with fidelity. However, measurement of parent strategy use varies across studies. Most studies use one of two types of observational coding measures (macro and micro-coding). Macro-codes are known for being efficient while micro-codes are known for being precise. The present study evaluates the reliability and validity of the *NDBI-Fi*, a macro-code, compared to a micro-code. Parent-child interaction videos for 177 participants were used to compare these measures. Results demonstrated that the *NDBI-Fi* had strong interrater reliability. It also had strong convergent validity with the micro-code after intervention. In addition, the *NDBI-Fi* was sensitive to change, and it demonstrated precision comparable to the micro-code. Furthermore, a novel scoring procedure detected differences in parents who learned different intervention strategy types. However, the *NDBI-Fi* did not demonstrate strong validity before intervention, particularly when measuring responsive intervention strategies. Taken together, findings support the use of the *NDBI-Fi* as an outcome measure, and future work should focus on continued development of valid pre-intervention macro-codes.

Introduction

It is widely recognized that early, intensive interventions have the potential to increase developmental outcomes for children with autism spectrum disorder (ASD; McManus et al., 2019; Zwaigenbaum et al., 2015). For such interventions to be effective, they must target the core deficits exhibited by children with ASD, such as social communication (Zwaigenbaum et al., 2015). Naturalistic Developmental Behavioral Interventions (NDBIs) have been suggested as particularly promising for improving such outcomes (Sandbank et al. 2019). NDBIs are frequently implemented by parents, thus they have an added benefit of including a family-centered component that is a cornerstone of early intervention. Critical to successful parent-implemented NDBIs is a way to measure changes in parent behavior that are likely to improve child social communication.

Correspondence regarding this article should be addressed to Bailey Sone, 2240 Campus Drive, Evanston, IL 60208. bailey.sone@u.northwestern.edu.

We have no known conflicts of interest to disclose.

Fidelity in Parent-Implemented Intervention

Parent-implemented interventions emphasize the active role of the parent as the primary teacher and communication partner for their child (Roberts et al., 2019). Systematically teaching parents to implement intervention strategies increases treatment dosage allowing children with ASD to receive the recommended 10–25 hours of weekly services (McManus et al., 2019; Virués-Ortega, 2010). However, parents must implement the intervention often and accurately in order for child communication to improve (Barton & Fettig, 2013; Haring Biel et al., 2019). In other words, parent fidelity as measured by quantity and quality of intervention delivery is a crucial component that contributes to child intervention outcomes.

Meta-analyses have demonstrated positive communication outcomes for children with ASD following parent-implemented intervention (Hampton & Kaiser, 2016; Roberts et al., 2019). Given the integral role of parent fidelity in these interventions, it is surprising that reporting of parent outcomes is inconsistent across individual studies (Hampton & Kaiser, 2016; Roberts et al., 2019). In fact, parent outcomes are reported in less than half of group design studies on parent-implemented interventions (Roberts et al., 2019). This is particularly problematic because study-level differences in fidelity could account for different study outcomes, leading to a lack of clarity about which intervention strategies and features are most beneficial for children with ASD. For example, one meta-analysis showed that children with ASD make the greatest gains in spoken language following a combination of parent and clinician-implemented intervention, positing that the presence of the clinician contributes to higher fidelity (Hampton & Kaiser, 2016). Furthermore, when parent fidelity is measured, the methods for its measurement are variable or not adequately described (Roberts et al., 2019).

Measuring Parent Outcomes

Observational measures are the gold standard for evaluating parent-child interactions (Gridley et al., 2019). Such measures can be broadly described as using one of two types of coding methods. Micro-coding allows for the analysis of fine-grained, specific details that may not otherwise be apparent, but it is time-consuming and requires extensive training (Dishion et al., 2017). This level of detail is achieved through coding discrete events for the constructs of interest (Dishion et al., 2017). In measuring parent-child interactions, parent fidelity can be measured with micro-level precision using count coding (Yoder et al. 2018), in which the coder decides each time a strategy is used. In contrast, macro-coding involves making broad, global judgements. It requires minimal time and less training, and thus it can be more cost effective (Dishion et al., 2017; Rosenberg et al., 1986). In measuring parent-child interactions, parent fidelity can be measured at a macro-level using rating scales (Yoder et al. 2018). These two methods represent an inverse relationship between time and precision, such that one system cannot be ideal (Rosenberg et al., 1986). This trade-off calls into question the extent to which studies that use these different measurement methods are truly comparable and the extent to which they capture the same constructs.

Such measurement concerns, paired with an increased interest in feasibility and efficiency in both research and practice, have prompted comparisons of micro and macro-codes with varying results. For example, Adamson and colleagues (2012) demonstrated a strong

relationship between micro-coding and macro-coding when evaluating parent-child joint engagement. In contrast, Dishion and colleagues (2017) found that micro-coding and macro-coding did not measure the same constructs when evaluating parenting skills. Most notably, Suhrheinrich and colleagues (2019) showed that some macro-codes (5-point Likert scales) demonstrated strong reliability with micro-codes but other macro-codes (3-point Likert scales) were less reliable when evaluating provider fidelity in a communication intervention for children with ASD. However, it remains unclear the extent to which these findings extend to observational measures of parent fidelity in parent-implemented communication interventions.

Observational measures are the most commonly used metric following parent-implemented communication intervention in studies that report parent outcomes (Frost et al., 2020). Problematically, many studies use fidelity measures or coding methods that are lab or intervention-specific without reporting psychometric properties (Frost et al., 2020). As such, there are two substantial measurement barriers to effective intervention for children with ASD: (a) potential differences in measured constructs across studies (e.g., lab-specific measures) and (b) aforementioned differences in observational coding methods (e.g., micro-coding and macro-coding). The resultant variety in measurement created by both of these barriers highlights the need for published tools that quantify the nature of parent-child interactions, particularly those that have applications in measuring parent fidelity.

To meet this need, Frost and colleagues (2020) developed the *Naturalistic Developmental Behavioral Intervention Fidelity Rating Scale* (*NDBI-Fi*). This macro-code was created to rate the quantity and quality of NDBI strategies. NDBIs are a group of evidence-based treatments for children with ASD that are broadly established from the same theoretical framework and share common intervention strategies (Schreibman et al., 2015). Often described as having hybrid characteristics, NDBIs merge the developmental sciences with features of traditional applied behavior analysis. The combination of these characteristics results in interventions that are not only child-led and naturalistic, but also provide structure that facilitates learning for children with ASD (Schreibman et al., 2015). The *NDBI-Fi* is the first published measure that isolates common features of NDBIs and as such is not specific to any single intervention program (Frost et al., 2020). Having a singular, feasible measure could aid in both meta-analysis and research on active intervention ingredients. Promising initial reliability and validity for the *NDBI-Fi* was established using multiple empirically based NDBIs (Frost et al., 2020). While the goal of using a singular measure that only includes common features is promising, it is also possible that such a measure may fail to detect intervention-specific nuances. In fact, some items that individual researchers consider essential intervention elements were not included in the *NDBI-Fi* (Frost et al., 2020). For this reason, it is crucial that psychometric analysis of the *NDBI-Fi* is replicated across different NDBIs and applications.

Applications of Macro-Coding

In addition to offering a common measure for all NDBIs, the constructs measured in the *NDBI-Fi* may apply to many parent-implemented interventions due to their hybrid characteristics. A recent meta-analysis found that the majority of parent-implemented

communication interventions for children with and at-risk for developmental delays included a naturalistic framework (Roberts et al., 2019). Because the *NDBI-Fi* includes parent strategies informed by both theories of natural learning (i.e., responsiveness) and behavioral theory (i.e., antecedent-behavior-consequence), its items could have broader applications to interventions that address these constructs independently. However, to determine the broader applications of the *NDBI-Fi*, it is necessary to understand its reliability and validity for interventions that include some, but not all, components of NDBIs. Likewise, development of a differential scoring procedure for separate theoretical constructs may improve the utility of the *NDBI-Fi*.

The use of a macro-code may also have applications in clinical settings due to its feasibility. Addressing the research-to-practice gap through implementation science is a necessary next step in improving outcomes for children with ASD (Barton & Fettig, 2013; Haring Biel et al., 2019; Vivanti et al., 2018). Parent-implemented interventions are used at varying degrees in clinical practice, and few clinicians who provide these communication interventions report using parent observation (Douglas et al., 2019). This is not surprising because to our knowledge, there is no widely disseminated tool for clinicians to use to measure parents' use of intervention strategies. A tool such as the *NDBI-Fi* could promote structured parent-child observations that aid in the provision of treatment planning and progress monitoring. Progress monitoring happens on an ongoing basis, and clinicians require a tool that is not only useful in measuring strategy use when parents have learned an NDBI, but also over the course of an intervention program when parents may use strategies at varying degrees.

Such practice-based considerations are vital to intervention research for children with ASD. Designing interventions that are feasible and usable from the outset improves the translation from research to practice (Vivanti et al., 2018). As such, the use of a more clinically feasible *NDBI-Fi* as opposed to time-consuming micro-coding as a common outcome measure across intervention studies is a critical next step to reduce this research-to-practice gap. However, it is first necessary to evaluate the validity of the *NDBI-Fi* before recommending its use in either research or clinical settings.

Purpose

The purpose of this study is to extend the initial reliability and validity of the *NDBI-Fi*. Specifically, the present study adds to prior work by evaluating the validity of the *NDBI-Fi* in comparison to a different coding method than used in the original study: a precise micro-code. As such, the research questions and methods were informed by those used in the *NDBI-Fi*'s development and validation (Frost et al., 2020). The present study also examines the *NDBI-Fi's* utility in measuring different intervention strategy types in order for it to be applied in a broader context. The following research questions guided this study:

1. What is the inter-rater reliability of the *NDBI-Fi*?
2. How do the *NDBI-Fi* and a micro-coded measure compare in measuring parent strategy use (a) before intervention and (b) after intervention?
3. Is the *NDBI-FI* sensitive to change during intervention?

4. Can the *NDBI-Fi* detect differences between parents who learn different types of intervention strategies?

Methods

Study Design

This cross-sectional, longitudinal study used extant video data from two samples. Both groups were recruited by the Early Intervention Research Group at Northwestern University. The first group (n = 60) included a baseline only observation. The second group participated in a randomized clinical trial of two parent-implemented interventions. This sample included 117 participants with baseline data and 95 (out of 117) with baseline and post-intervention data. Data for both groups were combined, such that the full sample included 177 videos at baseline and 95 videos post-intervention. These videos were scored using the *NDBI-Fi* and a micro-code.

Participants

Participants were 177 parent-child dyads. The mean age of the children in the full sample was 33.08 months (*SD* = 6.14). To be eligible for either study, children were required to have a diagnosis of ASD. Diagnoses were confirmed by a research reliable clinician using the Autism Diagnostic Observation Schedule – Second Edition (Lord et al., 2012). Consistent with the prevalence of ASD, children were primarily male (76%). Dyads lived in the Chicago area, and participants were excluded if English was not the primary language spoken in the home. On average, the parents reported their race as Caucasian (53%) and education level as a college degree or higher (56%). As such, about half the sample was diverse with respect to race, ethnicity, or socioeconomic status. Demographic data are shown in Table 1.

Intervention

Of the 177 total dyads, 95 dyads completed an 8-week intervention as part of a clinical trial in which mothers were randomized to a parent-implemented intervention strategy type common in NDBIs (responsive or directive; 1R01DC014709). Both conditions used the same instructional procedure, and parents have demonstrated the ability to learn both types of intervention strategies (Roberts et al., 2014).

Responsive Strategy Condition.—Responsive strategies were defined as strategies that were based on developmental, naturalistic frameworks. Parents in the responsive condition (n = 46) were taught to respond to child communication, to engage with their child, and to follow their child's lead. Parents were also taught to notice and respond to non-verbal and verbal communication and to interact by taking turns with their child.

Directive Strategy Condition.—Directive strategies were defined as strategies that were based on behavioral theory. Parents in the directive condition (n = 49) were taught to elicit child communication through the use of communication temptations and prompts. Parents were taught to arrange the environment to encourage their child to communicate and to scaffold prompts to teach and reinforce language.

Measures

Sampling Context.—Dyads were filmed during a naturalistic Parent-Child Interaction (PCX) using a standard set of toys. Before filming, parents were instructed to play with their child as they normally would. Ten-minute PCXs were recorded either in a research space at Northwestern University (n = 222; 152 at baseline, 70 post-intervention) or in the home (n = 50; 24 at baseline, 26 post-intervention), depending on the needs of the family.

NDBI-Fi.—All PCXs were macro-coded using a modified version of the *NDBI-Fi*. The original *NDBI-Fi* is an eight-item rating scale. Each item is scored on a five-point Likert scale, with the average of all items representing overall fidelity. Scores are assigned based on a global assessment quality and/or quantity, such that the overall fidelity score is representative of both features. The *NDBI-Fi* has an intraclass correlation between raters of 0.80, demonstrating good reliability (Frost et al., 2020). In addition, the *NDBI-Fi* is positively correlated ($r = 0.60$) with the global fidelity scales or interval macro-codes collected for three NDBIs: Project ImPACT, Pivotal Response Training, and Social ABCs (Frost et al., 2020), demonstrating strong construct validity.

The *NDBI-Fi* was constructed using an iterative process with experts on NDBIs (Frost et al., 2020). Since this process was used to ensure that the essential components of NDBIs were measured, minor modifications were made to ensure that the *NDBI-Fi* accurately reflected the intervention participants received in the present study. As such, one additional item was added (Pace Verbal Models) and minor changes were made in the scoring guidelines for three items (Responding to Communication, Communication Temptations, and Frequency of Direct Teaching). The full scale and modification description are available in Supplement A.

Along with the overall fidelity score, two additional fidelity scores were derived to evaluate parent learning of different strategy types. *NDBI-Fi* items that were theoretically based in responsiveness and taught in the responsive condition were averaged to create a responsive composite score, and *NDBI-Fi* items that were theoretically based in direct teaching and taught in the directive condition were averaged to create a directive composite score. Some *NDBI-Fi* items were not explicitly taught in either condition but still may be reflective of overall progress. These items were included in the overall fidelity score, hereafter referred to as the overall composite score. Rating items and composite scores are shown in Table 2.

NDBI-Fi Rating Procedure.—Raters were two speech-language pathologists (SLPs) with over three years of experience working with young children and their families as well as two speech-language pathology graduate students. One rater, a doctoral student and SLP, was trained in the interventions used in the larger trial. The second SLP rater was not trained in the specific intervention but had prior training in another NDBI. The clinical graduate students did not have any prior experience delivering NDBIs. The purpose of the differing experiences was to ensure clinical usability and to assess reliability in the context of raters with different intervention backgrounds. Raters were kept naïve to intervention condition to the greatest extent possible. However, this was not always possible for one of the four raters due to involvement in other elements of the study.

All raters were trained to reliability with a standard set of consensus-rated videos using the recommendations from the original article (Frost et al., 2020). As such, raters were reliable once ratings on three consecutive videos met the following criteria compared to the training samples: (a) seven items were within one point, (b) no items were greater than two points apart, and (c) the overall composite score was within 0.5 points.

Due to the continuous nature of the larger clinical trial, it was not possible for raters to be naïve to all timepoints. However, raters were naïve to timepoint for 20% of videos from the intervention sample (n = 40). Overall composite scores on these naïve ratings did not differ significantly from overall composite scores on non-naïve ratings at baseline ($t = 0.30$, $p = 0.77$) or post-intervention ($t = 0.22$, $p = 0.83$), suggesting that knowledge of timepoint did not compromise the integrity of the ratings. There was also no significant effect of knowledge of timepoint for responsive or directive composites scores.

Micro-Code.—All PCXs were micro-coded using a method that was developed as the primary outcome measure in the larger clinical trial. It was designed by clinicians and researchers with training and expertise in NDBIs to capture parent use of target strategies. Further, the micro-code was created in accordance with widely accepted recommendations for observational measures of behavior (Yoder et al., 2018). Similar micro-codes have been used to measure parent outcomes in previous trials on parent-implemented NDBIs with demonstrated interrater reliability (Roberts, 2019; Roberts & Kaiser, 2015). Taken together, the micro-code included in the present study is an ideal example of those commonly used in NDBI studies.

Each PCX was simultaneously transcribed and micro-coded for parent strategy use such that individual micro-code items were assigned at the utterance level. This micro-code allows for parent strategy use to be quantified by both frequency and percentage (i.e., frequency/opportunities). This type of count coding inherently accounts for the quantity of strategy use, and quality is also considered as codes are only assigned when the strategy meets a predetermined quality criterion. Thus, the items on both the micro-code and the *NDBI-Fi* account for features of quantity and quality. The present study included eight items from this micro-code quantified by percentage (score range = 0.00–1.00) that measured strategies specific to the intervention conditions. These eight items have strong reliability demonstrated by intraclass correlations ranging from 0.897 to 0.997. Similar to the *NDBI-Fi*, items were averaged to determine a responsive composite score, a directive composite score, and an overall composite score. These composite scores were critical in comparing the two measures because at the composite level, the micro-code and the *NDBI-Fi* capture the same theoretical constructs. While they share similarities at the item level, a single micro-code item may be represented in different ways on multiple NDBI-Fi items. Likewise, a single *NDBI-Fi* item may be represented in different ways on multiple micro-code items. Micro-code items, composite scores, and item correspondence with the NDBI-Fi are available in Supplement B (Supplement B Table 1).

Coding Procedure.—Coders were full-time research assistants trained to 80% reliability across each micro-code item. To ensure ongoing reliability, 20% of all PCXs were double-coded by a master coder, and discrepancies were discussed during weekly coding meetings.

All coders were naïve to intervention condition but were not naïve to timepoint, as baseline to post-intervention comparisons were not an aim of the larger study.

Micro-coding usually occurred prior to *NDBI-Fi* ratings. For each video, micro-coders and *NDBI-Fi* raters were not aware of the scores given on the other measure. In addition, no videos were micro-coded and rated on the *NDBI-Fi* by the same person. These steps ensured that scoring on one measure did not influence scoring on the other measure.

Analysis

Reliability.—To evaluate the interrater reliability of the *NDBI-Fi*, 25% of videos were randomly selected for double rating (n = 72). Reliability videos were equally distributed between baseline videos (n = 46) and post-intervention videos (n = 26) with respect to the total number of videos at each timepoint. Raters were unaware of which videos were selected for reliability calculations.

Interrater reliability was calculated using Krippendorff's alpha (Hayes & Krippendorff, 2007). Although intraclass correlations were calculated in the development of the *NDBI-Fi* (Frost et al., 2020), the present study seeks to extend this work. Krippendorff's alpha is determined by the data from each rater, and data is not added or omitted to calculate reliability (Hayes & Krippendorff, 2007). As such, Krippendorff's alpha poses a distinct advantage in fitting the level of measurement of the data, a consideration that is important for ordinal scales such as the *NDBI-Fi*. Krippendorff's alpha calculates the percent of disagreements and is interpreted on a 0.00–1.00 scale, such that 1.00 represents perfect agreement (Hayes & Krippendorff, 2007). Strong agreement is shown by alpha values exceeding 0.80, and alpha values should not be lower than 0.667 for a measure to demonstrate reliability (Krippendorff, 2018). For these analyses, composite scores were kept as sums instead of averages to maintain the true ordinal structure of the data.

Convergent Validity.—Validity was assessed by comparing the *NDBI-Fi* to the micro-code. Comparisons were made by calculating Pearson correlations at baseline and post-intervention. These analyses were conducted separately to determine the extent to which the association between the two measures varied by timepoint. Separate analyses at each timepoint also ensured that correlations were not related to repeated measures within the same participants.

Sensitivity to Change.—The *NDBI-Fi* was evaluated for sensitivity to change in three ways. First, baseline responsive composite scores and post-intervention responsive composite scores were compared for participants who learned responsive strategies. Second, baseline directive composite scores and post-intervention directive composite scores were compared for participants who learned directive strategies. The first two analyses tested the sensitivity of the new responsive and directive composite scores. The responsive and directive groups were analyzed separately for these first two analyses. This method limited the analyses to participants who were predicted to change on each composite score based on the intervention they received, thus accurately capturing sensitivity based on the study hypotheses. This method aligns with the methods used to analyze sensitivity in the development of the *NDBI-Fi*, in which sensitivity was analyzed for only those participants

expected to change on the measure (i.e., participants in the treatment condition, but not participants in the control condition). Third, baseline overall composite scores and post-intervention overall composite scores were compared for the entire intervention sample. This analysis tested the sensitivity of the entire measure when participants varied in the strategy type they learned. Paired t-tests were used for all three analyses. For contrast, micro-code composite scores were compared using the same process. Standardized mean difference (Cohen's *d*) between baseline scores and post-intervention scores was calculated for each measure to determine if the *NDBI-Fi* was comparable to the micro-code in the magnitude of change it detected.

Known Group Validity.—If parents who learn responsive strategies and parents who learn directive strategies systematically differ on responsive and directive composite scores, this may indirectly demonstrate that the *NDBI-Fi* items measure the intended constructs (Virues-Ortega et al., 2011). Participants who learned the responsive intervention were expected to have significantly greater responsive composite scores compared to parents who learned the directive intervention. Similarly, participants who learned the directive intervention were expected to have significantly greater directive composite scores compared to participants who learned the responsive intervention. Unpaired t-tests were used for these analyses. As in the previous analyses, micro-code composite scores were compared using the same method. Additionally, standardized mean difference (Cohen's *d*) between the responsive and directive groups was calculated for the responsive and directive composite scores for both the *NDBI-Fi* and the micro-code.

Patient and Public Involvement

No community members for whom this measure was developed to evaluate (e.g., parents of children with ASD, individuals with ASD) were involved in the production of this study. However, the first and last authors are certified speech-language pathologists and as such contributed a stakeholder perspective related to the efficiency and clinical usability of the *NDBI-Fi*.

Results

Reliability

The *NDBI-Fi* demonstrated good interrater reliability on all three composite scores. Each fell above the minimum acceptable standard, with the responsive composite score (α = 0.774), directive composite score (α = 0.704), and overall composite score (α = 0.752) all showing moderate to strong agreement.

At the item level, interrater reliability was more variable. Individual items ranged from having weak interrater reliability (e.g., Responding to Attempts to Communicate, α = 0.389) to having strong interrater reliability (e.g., Pace Verbal Models, α = 0.806). Five of the nine total items fell below the minimum acceptable standard for interrater reliability. However, four of these five items were close to that standard (α = 0.603 – 0.653), with only one item falling much below it (α = 0.389). Item-level and composite reliability are shown in Table 3.

Convergent Validity

Baseline.—The *NDBI-Fi* had variable convergent validity with the micro-code at baseline. The measures strongly correlated on directive composite measures ($r = 0.54$, $p < 0.001$) and the overall composite measures ($r = 0.58$, $p < 0.001$). However, responsive composite measures demonstrated a weaker correlation ($r = 0.30$, $p < 0.001$).

Post-Intervention.—The *NDBI-Fi* demonstrated convergent validity with the micro-code post-intervention. The measures strongly correlated on responsive composite measures ($r = 0.58$, $p < 0.001$), directive composite measures ($r = 0.63$, $p < 0.001$), and overall composite measures ($r = 0.57$, $p < 0.001$). All correlations are presented in Table 4 and scatterplots are available in the Supplement B (Supplement B Figure 1, Supplement B Figure 2).

Sensitivity to Change.—The *NDBI-Fi* detected significant differences from baseline to post-intervention across all composite scores. For participants who learned responsive strategies, there was a significant difference in responsive composite scores between baseline and post-intervention with a large effect size ($p < 0.001$, $d = 1.83$, 95% CI [1.13, 2.51]). For participants who learned directive strategies, there was a significant difference in directive composite scores between baseline and post-intervention with a large effect size ($p < 0.001$, $d = 0.79$, 95% CI [0.20, 1.37]). Additionally, there was a significant difference in overall composite scores between baseline and post-intervention with a large effect size ($p < 0.001$, $d = 0.81$, 95% CI [0.39, 1.23]) for the intervention sample. Micro-code analyses revealed similarly large effect sizes. Baseline and post-intervention data are shown in Table 5.

Known Group Validity.—The *NDBI-Fi* responsive composite score detected a significant difference between parents who learned the responsive intervention and parents who learned the directive intervention ($p < 0.001$, $d = 1.13$, 95% CI [0.66, 1.59]). Similarly, the directive composite score detected a significant difference between parents who learned the directive intervention and parents who learned the responsive intervention ($p < 0.001$, $d = 1.12$, 95% CI [0.65, 1.58]). The magnitude of this difference was strong on both the *NDBI-Fi* and the micro-code. Group data are shown in Table 6.

Discussion

The results of this study support the use of the *NDBI-Fi* to measure parent outcomes in parent-implemented interventions for children with ASD. Further, these results suggest that efficient macro-codes can serve as reliable, valid, and precise measures. First, the *NDBI-Fi* demonstrated reliability in measuring overall parent strategy use. This finding replicates the strong reliability of the overall composite score from the original study using a reliability coefficient well-suited for ordinal data. The present study also validates two newly derived composite scores such that it captures parent responsive strategy use and parent directive strategy use. Reliability was demonstrated using four coders with varying levels of experience with NDBIs, including clinical graduate students with no prior training or experience in parent observation, suggesting that reliability may be attainable in clinical practice settings.

In addition to being reliable, the *NDBI-Fi* demonstrated convergent validity compared to a micro-code following intervention. At the post-intervention timepoint, results indicated there was a strong, positive association between the two measures on parent responsive composite scores, directive composite scores, and overall composite scores. Because the micro-code is considered the gold-standard for precise, accurate measurement, this strong convergence poses a distinct advantage for the already efficient macro-code. Further, the convergent validity of the responsive and directive composite scores suggest that the *NDBI-Fi* may not only be applicable to NDBIs but may also be more broadly applicable across many parent-implemented interventions.

Results also indicate that the *NDBI-Fi* is sensitive to changes made during a brief intervention. This finding was consistent in responsive composite scores of parents who learned responsive intervention strategies, directive composite scores of parents who learned directive strategies, and even in the overall composite scores for the full group in which parents learned some, but not all, of the strategies measured on the scale. Effect sizes from baseline to post-intervention were comparable to the micro-code, suggesting that there is not a substantial methodological disadvantage to using the *NDBI-Fi*. Finally, the responsive and directive composite scores appropriately differentiated between these groups, adding confidence that these constructs are appropriately defined.

However, the results of this study also reveal several disadvantages of the *NDBI-Fi*. First, it may not be precise in measuring responsive strategies at baseline. Based on our data, we posit that this finding may be due to the fact that parents often use responsive strategies to some degree even without instruction. In contrast, directive strategies are rare prior to instruction (i.e., scores of or near zero occurred on the directive composite but not on the responsive composite). This observation may be due to the fact that responsive strategies are child-led, such that playing with the child would necessitate the use of responsiveness to some degree, while directive strategies are adult-led, and therefore depend on a parent's use of that specific skill. A micro-code may be better at detecting subtle differences between parents' use of responsive strategies when they occur at lower rates or are of lower quality. However, baseline levels of directive strategies are likely measured with similar precision by micro and macro measures due to their rarity.

A second disadvantage is that the *NDBI-Fi* had inconsistent interrater reliability at the item-level, with one item demonstrating poor reliability (Responding to Attempts to Communicate, $\alpha = 0.389$). Notably, the original article also found inconsistent item-level reliability, as one item demonstrated poor reliability (Quality of Direct Teaching, ICC = 0.33; Frost et al. 2020). An implication for this finding is that, at present, the *NDBI-Fi* may not be suitable for research on active ingredients of interventions because individual strategy use cannot be reliability measured.

The results of this study should be interpreted in light of its limitations. First, it was not possible to keep raters and coders naïve to timepoint on either of the measures. Although no bias due to timepoint was detected on the *NDBI-Fi*, it is possible that sensitivity analyses may have been impacted by knowledge of timepoint. Second, we did not implement a video viewing protocol in the present study. This may have led to differences in viewing and

scoring practices between our raters, impacting reliability, and could lead to replicability concerns in future studies.

Future work might first seek to improve both the disadvantages of the *NDBI-Fi* and the limitations of the present study. For example, item-level reliability may be improved by developing a structured viewing and scoring system. It may be that watching each video multiple times improves item-level reliability, or it could be that dividing videos into smaller segments and then averaging scores across items improves item-level reliability. In fact, similar procedures were used in a recent study on another macro-code, the *Measure of NDBI Strategy Implementation-Caregiver Change (MONSI-CC)* and yielded good item-level interrater reliability across all items in its initial development (Vibert et al., 2020). It is also possible that reliability is influenced by the diverse participant sample in our present study. Previous work has shown that macro-codes are more likely to be subject to cultural and racial bias than micro-codes (Yasui & Dishion, 2008). Follow-up work may explore the extent to which such bias is present when scoring the *NDBI-Fi,* and if such bias exists, future work should develop rater training to reduce it.

Our finding that the *NDBI-Fi* did not precisely measure parent strategy use at baseline is an important one, given the goal of implementation in clinical practice. There remains a need for common, efficient, and feasible measures that can support both treatment planning and progress monitoring. Future work should expand the *NDBI-Fi* to include items that refine the responsiveness composite such that it can better capture both learned strategies and naturally occurring responsiveness in parents.

Finally, these promising initial results may prompt future work on both broader applications of macro-coding to parent-implemented interventions that share intervention features with NDBIs across other populations of toddlers with developmental delays as well as implementation in clinical practice. A next step towards this goal is to determine the reliability of the *NDBI-Fi* when used by practicing clinicians. Although the present study used coders of varying experience levels, suggesting the clinical utility of the *NDBI-Fi*, these coders were trained to use the measure in a research setting, and this training process may not be feasible or accessible in practice settings. Taken together, results from this study suggest that both the continued development of macro-codes and their current and future applications have the potential to significantly advance early intervention research and practice for children with ASD and beyond.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Adamson LB, Bakeman R, Deckner DF, & Nelson PB (2012). Rating parent–child interactions: Joint engagement, communication dynamics, and shared topics in autism, Down syndrome, and typical development. Journal of Autism and Developmental Disorders, 42(12), 2622–2635. 10.1007/s10803-012-1520-1 [PubMed: 22466689]

Barton EE, & Fettig A (2013). Parent-implemented interventions for young children with disabilities: A review of fidelity features. Journal of Early Intervention, 35(2), 194–219. 10.1177/1053815113504625

Dishion TJ, Mun CJ, Tein J-Y, Kim H, Shaw DS, Gardner F, Wilson MW, & Peterson J (2017). The validation of macro and micro observations of parent–child dynamics using the relationship affect coding system in early childhood. Prevention Science, 18(3), 268–280. 10.1007/s11121-016-0697-5 [PubMed: 27620623]

Douglas SN, Meadan H, & Kammes R (2019). Early interventionists' caregiver coaching: A mixed methods approach exploring experiences and practices. Topics in Early Childhood Special Education, 0271121419829899. 10.1177/0271121419829899

Frost KM, Brian J, Gengoux GW, Hardan A, Rieth SR, Stahmer A, & Ingersoll B (2020). Identifying and measuring the common elements of naturalistic developmental behavioral interventions for autism spectrum disorder: Development of the NDBI-Fi. Autism, 1362361320944011. 10.1177/1362361320944011

Gridley N, Blower S, Dunn A, Bywater T, Whittaker K, & Bryant M (2019). Psychometric properties of parent–child (0–5 years) interaction outcome measures as used in randomized controlled trials of parent programs: A systematic review. Clinical Child and Family Psychology Review, 22(2), 253–271. 10.1007/s10567-019-00275-3 [PubMed: 30734193]

Hampton LH, & Kaiser AP (2016). Intervention effects on spoken-language outcomes for children with autism: A systematic review and meta-analysis. Journal of Intellectual Disability Research, 60(5), 444–463. 10.1111/jir.12283 [PubMed: 27120988]

Haring Biel C, Buzhardt J, Brown JA, Romano MK, Lorio CM, Windsor KS, Kaczmarek LA, Gwin R, Sandall SS, & Goldstein H (2019). Language interventions taught to caregivers in homes and classrooms: A review of intervention and implementation fidelity. Early Childhood Research Quarterly, 50, 140–156. 10.1016/j.ecresq.2018.12.002

Hayes AF, & Krippendorff K (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1(1), 77–89. 10.1080/19312450709336664

Kaiser AP, Hancock TB, & Nietfeld JP (2000). The effects of parent-implemented enhanced milieu teaching on the social communication of children who have autism. Early Education and Development, 11(4), 423–446. 10.1207/s15566935eed1104_4

Krippendorff K (2018). Content analysis: An introduction to its methodology. Sage publications.

Lord C, Luyster R, Gotham K, & Guthrie W (2012). Autism diagnostic observation schedule, 2nd edition (ADOS-2) Manual (Part II): Toddler module. Torrence, CA: Western Psychological Services.

McManus BM, Richardson Z, Schenkman M, Murphy N, & Morrato EH (2019). Timing and intensity of early intervention service use and outcomes among a safety-net population of children. JAMA Network Open, 2(1), e187529–e187529. 10.1001/jamanetworkopen.2018.7529 [PubMed: 30681716]

Roberts MY (2019). Parent-implemented communication treatment for infants and toddlers with hearing loss: A randomized pilot trial. Journal of Speech, Language, and Hearing Research, 62(1), 143–152. 10.1044/2018_JSLHR-L-18-0079

Roberts MY, Curtis PR, Sone BJ, & Hampton LH (2019). Association of parent training with child language development: A systematic review and meta-analysis. JAMA pediatrics, 173(7), 671–680. 10.1001/jamapediatrics.2019.1197 [PubMed: 31107508]

Roberts MY, & Kaiser AP (2015). Early intervention for toddlers with language delays: a randomized controlled trial. Pediatrics, 135(4), 686–693. 10.1542/peds.2014-2134 [PubMed: 25733749]

Roberts MY, Kaiser AP, Wolfe CE, Bryant JD, & Spidalieri AM (2014). Effects of the teach-model-coach-review instructional approach on caregiver use of language support strategies and children's expressive language skills. Journal of Speech, Language, and Hearing Research, 57(5), 1851–1869. 10.1044/2014_JSLHR-L-13-0113

Rosenberg SA, Robinson CC, & Beckman PJ (1986). Measures of parent-infant interaction: An overview. Topics in Early Childhood Special Education, 6(2), 32–43. 10.1177/027112148600600204

Sandbank M, Bottema-Beutel K, Crowley S, Cassidy M, Dunham K, Feldman JI, Crank J, Albarran SA, Raj S, Mahbub P, & Woynaroski TG (2020). Project AIM: Autism intervention meta-analysis for studies of young children. Psychological Bulletin, 146(1), 1–29. 10.1037/bul0000215 [PubMed: 31763860]

Schreibman L, Dawson G, Stahmer AC, Landa R, Rogers SJ, McGee GG, Kasari C, Ingersoll B, Kaiser AP, Bruinsma Y, McNerney E, Wetherby A, & Halladay A (2015). Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder. Journal of Autism and Developmental Disorders, 45(8), 2411–2428. 10.1007/s10803-015-2407-8 [PubMed: 25737021]

Suhrheinrich J, Dickson KS, Chan N, Chan JC, Wang T, & Stahmer AC (2020). Fidelity assessment in community programs: An approach to validating simplified methodology. Behavior Analysis in Practice, 13(1), 29–39. 10.1007/s40617-019-00337-6 [PubMed: 32231965]

Vibert BA, Dufek S, Klein CB, Choi YB, Winter J, Lord C, & Kim SH (2020). Quantifying caregiver change across early autism interventions using the measure of NDBI strategy implementation: Caregiver change (MONSI-CC). Journal of Autism and Developmental Disorders, 1–16. 10.1007/s10803-019-04342-0 [PubMed: 31729599]

Virués-Ortega J (2010). Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. Clinical Psychology Review, 30(4), 387–399. 10.1016/j.cpr.2010.01.008 [PubMed: 20223569]

Virues-Ortega J, Montaño-Fidalgo M, Froján-Parga MX, & Calero-Elvira A (2011). Descriptive analysis of the verbal behavior of a therapist: A known-group validity analysis of the putative behavioral functions involved in clinical interaction. Behavior Therapy, 42(4), 547–559. 10.1016/j.beth.2010.12.004 [PubMed: 22035985]

Vivanti G, Kasari C, Green J, Mandell D, Maye M, & Hudry K (2018). Implementing and evaluating early intervention for children with autism: Where are the gaps and what should we do? Autism Research, 11(1), 16–23. 10.1002/aur.1900 [PubMed: 29206358]

Yasui M, & Dishion TJ (2008). Direct observation of family management: Validity and reliability as a function of coder ethnicity and training. Behavior Therapy, 39(4), 336–347. 10.1016/j.beth.2007.10.001 [PubMed: 19027430]

Yoder PJ, Lloyd BP, & Symons FJ (2018). Observational measurement of behavior: Second edition. Paul H. Brooks Publishing Co.

Zwaigenbaum L, Bauman ML, Choueiri R, Kasari C, Carter A, Granpeesheh D, Mailloux Z, Roley SS, Wagner S, Fein D, Pierce K, Buie T, Davis PA, Newschaffer C, Robins D, Wetherby A, Stone WL, Yirmiya M, Estes N, Hansen R, McPartland JC, & Natowicz MR (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. 10.1542/peds.2014-3667E [PubMed: 26430170]

Table 1.

Baseline Participant Characteristics

		Full Sample[a]	Intervention sample[b]	
Characteristic	**Definition**	***n* = 177**	**Responsive *n* = 46**	**Directive *n* = 49**
Child				
Age, *M (SD)*	Months	33.08 (6.14)	32.41 (5.99)	33.60 (6.21)
Gender, *n* (%)	Male	135 (76)	31 (67)	41 (84)
	Female	42 (24)	15 (33)	8 (16)
Race, *n* (%)	African American	19 (11)	5 (11)	6 (12)
	American Indian/Alaskan	3 (2)	0 (0)	0 (0)
	Asian	16 (9)	2 (4)	1 (2)
	Caucasian	84 (47)	23 (50)	30 (61)
	Multiple[c]	34 (19)	10 (22)	11 (22)
	Native Hawaiian/PI[d]	0 (0)	0 (0)	0 (0)
	No Response	21 (12)	6 (13)	1 (2)
Ethnicity, *n* (%)	Hispanic or Latinx	62 (35)	15 (33)	16 (33)
	Not Hispanic or Latinx	105 (59)	28 (61)	32 (65)
	No Response	10 (6)	3 (7)	1 (2)
Parent				
Gender, *n* (%)	Male	13 (7)	0 (0)	0 (0)
	Female	164 (93)	46 (100)	49 (100)
Race, *n* (%)	African American	21 (12)	5 (11)	7 (14)
	American Indian/Alaskan	4 (2)	0 (0)	1 (2)
	Asian	22 (12)	2 (4)	3 (6)
	Caucasian	93 (53)	29 (63)	31 (63)
	Multiple	9 (5)	4 (9)	2 (4)
	Native Hawaiian/PI[c]	1 (0.06)	1 (2)	0 (0)
	No Response	27 (15)	5 (11)	5 (10)
Ethnicity, *n* (%)	Hispanic or Latinx	51 (29)	14 (30)	11 (22)
	Not Hispanic or Latinx	116 (66)	30 (65)	36 (73)
	No Response	10 (6)	2 (4)	2 (4)
Education, *n* (%)	>High School	3 (2)	1 (2)	1 (2)
	High School	14 (8)	2 (4)	4 (8)
	Some College	44 (25)	10 (22)	13 (27)
	Special Training	12 (7)	5 (11)	2 (4)
	College Degree	46 (26)	14 (30)	17 (35)
	Graduate Degree +	53 (30)	14 (30)	12 (24)
	No Response	5 (3)	0 (0)	0 (0)

Note.

[a]Participants from the full sample are from two larger clinical trials: (1) 60 participants, (2) 117 participants

[b] Participants from the intervention sample are from trial (2); data reflects participants with post-intervention data only; participants without post-intervention data are included in the full sample

[c] Multiple = parent indicated that they belonged to more than one of the categories presented

[d] PI = Other Pacific Islander

Table 2.

NDBI-Fi Rating Items

			Composite[b]		
Item Number	**Rating Item**	**Score Range[a]**	**Responsive**	**Directive**	**Overall**
1	Face to Face	1–5			X
2	Follow the Child's Lead	1–5	X		X
3	Positive Affect	1–5			X
4	Modeling Language	1–5	X		X
5	Responding to Communication	1–5	X		X
6	Pace Verbal Models	1–5	X		X
7	Communication Temptations	0–5		X	X
8	Frequency of Direct Teaching	0–5		X	X
9	Quality of Direct Teaching	0–5		X	X

Note.

[a]all items were scored on a 1–5 Likert scale, but score ranges containing a 0 indicate that NA was selected when the strategy never occurred;

[b]composite scores are calculated by averaging all applicable rating items

Table 3.

Item and Composite Reliability

Item	Krippendorff's alpha
Face to Face	0.742
Follow the Child's Lead	0.620
Positive Affect	0.650
Modeling Language	0.603
Responding to Communication	0.389
Pace Verbal Models	0.806
Communication Temptations	0.668
Frequency of Direct Teaching	0.781
Quality of Direct Teaching	0.653
Composite	
Responsive	0.774
Directive	0.704
Overall	0.752

Note. Values above 0.80 were considered to have strong reliability; values above 0.667 were considered adequate

Table 4.

Composite Score Correlations

	Micro Responsive	Micro Directive	Micro Overall
Baseline Correlations			
NDBI-Fi Responsive	0.30***	0.15*	0.23**
NDBI-Fi Directive	0.22**	0.54***	0.52***
NDBI-Fi Overall	0.36***	0.54***	0.58***
Post-Intervention Correlations			
NDBI-Fi Responsive	0.58***	−0.11	0.20*
NDBI-Fi Directive	−0.18	0.63***	0.55***
NDBI-Fi Overall	0.22*	0.44***	0.57***

Note.

* $p < 0.05$;

** $p < 0.01$;

*** $p < 0.001$

Table 5.

NDBI-Fi and Micro-Code Sensitivity

	Baseline		Post-Intervention				
	Mean (SD)	**Range**	**Mean (SD)**	**Range**	***t***	***d***	***p***
Full Intervention							
Sample n = 95							
Overall Micro	0.28 (0.08)	0.10 – 0.51	0.36 (0.11)	0.15 – 0.65	7.26	0.88	< 0.0001
Overall *NDBI-Fi*	2.24 (0.44)	1.33 – 3.33	2.65 (0.56)	1.56 – 4.00	6.56	0.81	< 0.0001
Responsive Group n = 46							
Responsive Micro	0.30 (0.07)	0.18 – 0.54	0.48 (0.12)	0.20 – 0.80	12.19	1.86	< 0.0001
Responsive *NDBI-Fi*	2.84 (0.50)	1.75 – 3.75	3.78 (0.52)	2.50 – 4.75	10.08	1.83	< 0.0001
Directive Group n = 49							
Directive Micro	0.28 (0.12)	0.00 – 0.52	0.42 (0.19)	0.00 – 0.78	5.00	0.87	< 0.0001
Directive *NDBI-Fi*	0.82 (0.93)	0.00 – 2.67	1.76 (1.43)	0.00 – 4.33	4.59	0.79	< 0.0001

Table 6.

Known Group Validity

	Responsive Group n = 46		Directive Group n = 49				
***NDBI-Fi* Score**	**Mean (SD)**	**Range**	**Mean (SD)**	**Range**	***t***	***d***	***p***
Responsive Micro	0.48 (0.12)	0.20 – 0.80	0.34 (0.06)	0.23 – 0.47	6.95	1.45	< 0.0001
Responsive *NDBI-Fi*	3.78 (0.52)	2.50 – 4.75	3.19 (0.51)	2.25 – 4.50	5.50	1.13	< 0.0001
Directive Micro	0.20 (0.20)	0.00 – 0.75	0.42 (0.19)	0.00 – 0.78	5.36	1.10	< 0.0001
Directive *NDBI-Fi*	0.48 (0.73)	0.00 – 2.67	1.76 (1.43)	0.00 – 4.33	5.55	1.12	< 0.0001

Note. Analyses at post-intervention