Research in Autism Spectrum Disorders 5 (2011) 1430–1440

Contents lists available at ScienceDirect

Research in Autism Spectrum Disorders

Journal homepage: http://ees.elsevier.com/RASD/default.asp

Assessing implementation of comprehensive treatment models for young children with ASD: Reliability and validity of two measures

Kara Hume [a,*], Brian Boyd [b], Matt McBee [a], Drew Coman [c], Anibal Gutierrez [c], Evelyn Shaw [a], Laurie Sperry [d], Michael Alessandri [c], Samuel Odom [a]

[a] Frank Porter Graham Child Development Institute, University of North Carolina, Chapel Hill, 517 South Greensboro Street, Carrboro, NC 27510, USA
[b] Department of Allied Health Sciences, University of North Carolina, Chapel Hill, UNC-CH Bondurant Hall, Chapel Hill, NC 27599-7120, USA
[c] Department of Psychology, University of Miami, P.O. Box 248185, Coral Gables, FL 33124-0751, USA
[d] School of Education and Human Development, University of Colorado, Denver, P.O. Box 173364, Campus Box 106, Denver, CO 80217-3364, USA

ARTICLE INFO

Article history:
Received 5 October 2010
Accepted 3 February 2011
Available online 31 March 2011

Keywords:
Autism spectrum disorder
Comprehensive treatment models
Treatment implementation
Treatment integrity

ABSTRACT

Treatment implementation is an under-studied and under-reported aspect of intervention studies involving individuals with autism spectrum disorder (ASD). One primary area of concern is the lack of reliable and valid implementation measures, which allows a conclusive association to be drawn between the intervention and participant outcomes. This study examined the psychometric properties of two implementation measures developed for comprehensive treatment models serving preschoolers with ASD (i.e., LEAP and TEACCH). Both of the measures were completed in classrooms using LEAP or TEACCH instructional approaches as well as in classrooms in which a business-as-usual or non-model specific treatment approach was used. Across four months of one school year, a maximum of 4 observations were conducted in each of the 34 classrooms involved in the study. Results indicated that both implementation tools are reliable and valid, and that particular subscales of these measures allowed for discrimination of the three types of classrooms from each other. This step of psychometrically validating implementation measures as part of conducting efficacy studies may yield more robust associations between implementation and intervention effects.

1. Introduction

To promote positive developmental outcomes for children with autism spectrum disorders (ASD), over the past decades researchers have developed comprehensive treatment models of service (National Research Council, 2001). A prominent question in the field has been about the relative efficacy of these models (Odom, Boyd, Hall, & Hume, 2010). Any examination of efficacy, however, is built on the presumption that models are well implemented (Durlak, 2010), which, in turn, requires that implementation, be systematically assessed. Systematic assessments of implementation in scientific studies of efficacy must be reliable and valid, and to date, there are been few reports of the psychometric characteristics of implementation measures designed for programs for students with ASD. The purpose of this study was to analyze the reliability and validity of implementation measures for two prominent CTMs for children with ASD. Implementation was defined in the current study as "*the extent to which the critical components of an intended program are present when that program is enacted*" (Century, Rudnick, & Freeman, 2010).

* Corresponding author. Tel.: +1 919 843 2291; fax: +1 919 966 1786.
E-mail address: kara.hume@unc.edu (K. Hume).

1750-9467/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.rasd.2011.02.002

In their examination of educational practices for children with autism, a committee convened by the National Academy of Sciences identified a set of comprehensive models of intervention, which they defined as a set of practices designed to achieve a broad learning or developmental impact on the core deficits of ASD (National Research Council, 2001). Further, they occurred over an extended period of time (e.g., one year or multiple years), are intense in their application (e.g., 25 h or more per week), usually have multiple components targeting skills across multiple developmental domains, and many have strong parent involvement or training components. CTMs have been in existence for over 30 years and new models continue to be created. In a recent review, Odom et al. (2010) identified 30 CTMs that have been developed over the last three decades and are still in operation. Examples of historic CTMs are the UCLA Young Autism Project (which we will call the Lovaas model) (Lovaas, 1987), Treatment and Education of Autistic and Communication Handicapped Children (TEACCH) (Mesibov, Shea, & Schopler, 2005), the LEAP model (Hoyson, Jamieson, & Strain, 1984), and the Denver Model (Rogers et al., 2006).

A key movement in the field of educational and psychological interventions for students with ASD and other disabilities has been to establish the evidence-base for practices (National Autism Center's National Standards Project, 2009; National Professional Development Center on ASD, 2007), with a primary source of evidence for CTMs being supplied by efficacy studies. Yet, there remains a scarcity of high-quality, model-specific or comparative evaluation information about CTMs (Hume & Odom, in press). More than half of the 30 models reviewed by Odom et al. (2010) had no evidence of efficacy published in a peer-reviewed journal. In a recent critical review of CTMs for young children with ASD and their families, Rogers and Vismara (2008) evaluated the current research on comprehensive treatments for young children with ASD, finding limited evidence of efficacy for all but the Lovaas model, with some limited support for Pivotal Response Treatment (PRT) (Koegel, Koegel, Harrower, & Carter, 1999). Even the Lovaas model has been questioned. In recent evaluation report by the What Works Clearinghouse (2010), evaluators rated the Lovaas model as having small effects. Importantly, this evaluation was based on only two articles (from the more than 50 publications reviewed) that met the methodological inclusion criteria.

To respond to the need for high-quality treatment efficacy research, NIH convened a panel of investigators convened by NIH to determine the key needs around designing efficacy research, and they concluded that conducting comprehensive treatment research in the community settings and analyzing the relative effects of different CTMs were important (Lord et al., 2005). In a follow-up to that panel discussion, Smith, Scahill, et al. (2007) proposed a process for developing a program of research that began with manualization of treatment procedures and establishing fidelity/implementation protocols. In their review of comprehensive treatment programs, Odom et al. (2010) found as a group, CTMs were strongest in the operationalization (i.e., providing manualized descriptions of the content and procedures involved in model implementation) of their models; however, the actual measurement of implementation was relatively weak in comparison. Only one CTM had a fidelity measure with preliminary psychometric data, while 25% had informal or no methods of measuring fidelity.

The limited evidence of implementation measurement is also evidence in the ASD literature on focused intervention practices (i.e., single intervention designed to address outcome more limited in scope than would be addressed in CTMs). Wolery and Garfinkle (2002), in their review of the intervention literature from 1970 to the early 1990s, found that only 13% of the studies including students with ASD reported procedural fidelity information. When reviewing only single case design studies involving young children with ASD published from 1990 to 2003, Odom et al. (2003) found that 32% of the studies included implementation measures. A more recent finding indicated that only 18% of intervention studies for students with ASD published between the years of 1993–2004 assessed and reported treatment fidelity data (Wheeler, Baggett, Fox, & Blevins, 2006).

Both the limited use of implementation measures and the even more limited documentation of the reliability and validity of these measures is particularly problematic for several reasons. First, failure to ensure the integrity or fidelity with which an intervention is delivered poses a number of threats to drawing valid inferences about treatment effects (Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000). If the components of the treatment are not well measured, no definitive conclusions can be drawn regarding the effects of the independent variables on the outcome measures. Therefore, the perceived presence of a functional relationship between the dependent (e.g., child outcomes) and independent variables (e.g., specific CTM) may be faulty (Wheeler et al., 2006). Further, in efficacy research, without a measure of fidelity, researchers cannot (a) fully account for differences between experimental and control groups (e.g., the degree to which control classrooms are using intervention components); (b) determine whether unsuccessful outcomes are due to an ineffective intervention; or (c) determine if poor outcomes are a result of a failure to implement the intervention as intended (O'Donnell, 2008). Perhaps most importantly, is the relationship between implementation and intervention outcomes. In a review of meta-analyses of community based health programs, Durlak and DuPre (2008) summarized that programs with stronger adherence resulted in mean effect sizes that were 2–3 times higher than programs with poorer implementation (up to 12 times higher in several studies). In a review of a subset of the studies (59 studies related to program implementation), Durlak and DuPre found that in 76% of the studies there was a significant positive association between the level of implementation and at least half of the program outcomes (2008). Researchers can postulate that this direct relationship between degree of treatment integrity and degree of treatment outcome is applicable to the ASD population and related intervention research.

The purpose of this study was to analyze the reliability and validity of implementation measures for two established CTMs for students with ASD, the TEACCH model and the LEAP model. The measures are currently being used in an examination of relative efficacy of the two models, and the implementation measure development is consistent with the Smith, Scahill, et al. (2007) recommendation concerning the need for establishing measurement of treatment fidelity in efficacy research. The specific research questions are: (1) Are the two implementation measures reliable as judged by the assessment of inter-rater agreement, test–retest reliability, and internal consistency? and (2) Are the implementation

measures valid as judged by the utility in discriminating between the classrooms employing the two CTMs as well as classrooms following "business as usual" (i.e., not formally employing either model)?

2. Methods

Researchers have described three major steps in developing and evaluating fidelity measures (Mowbray, Holter, Teague, & Bybee, 2003). The first step is to identify critical components, both structural and instructional, of a given treatment model based on an expert consensus process and explicit description. This step includes describing data sources for behavioral indicators, establishing operational definitions, and developing indicators that assist in anchoring points on rating scales so they are objective and measureable (Mowbray et al., 2003). This step is described below (Section 2.1). The second step is to collect data to measure the indicators through a multi method, multi-informant approach. This step, which includes both classroom observation and interviews conducted by multiple research staff, is described in Section 2.4. The third step is to examine the indicators in terms of their reliability as well as validity (Moncher & Prinz, 1991). This final step is addressed in Sections 2.5 and 3.

2.1. Measures and measure development

LEAP program and fidelity measure. The LEAP model was established in 1981 by Dr. Phil Strain. The prevailing theoretical/conceptual foundation for the model is applied behavior analysis. The model, itself, has five programmatic features that make it a natural comparison to the TEACCH model: (a) typically developing children are full-time class members, (b) naturalistic teaching strategies are used (e.g., embedded learning opportunities), (c) the classroom arrangement mirrors that of a typical early childhood setting, (d) a co-teaching model of instruction is used with classrooms having a general and special educator, and (e) the use of a parent training component.

The original LEAP fidelity measure, titled *LEAP-USA Fidelity Form*, and previously titled *Quality Program Guidelines*, was developed by Dr. Strain and the LEAP outreach staff in 2002. The measure has 8 subscales and 38 total items. Subscales are identified in Table 3. Each item has 3–7 behavioral indicators to provide a more in-depth description of the item. Each item is scored on a 5-point scale, with 5 indicating full implementation of the item, 3 indicating partial implementation, and 1 indicating the item "Needs Work". Strain reported test–retest reliability of 0.88 on the original measure and predictive validity of the measure that explained 60% of the variance in child outcomes in a previously unpublished study (personal communication, 2007).

Several modifications were made to the measure for use in this study. An interview protocol was developed to ensure that all items could be accurately and consistently scored across project staff, as observers may not have an opportunity to observe all indicators/items while in the classroom (e.g., family training sessions; family involvement in IEP meetings). For scoring purposes, direct observation of any item did supersede a staff interview. The interview protocol had 13 questions, along with additional probes that could be used to ensure that each item/indicator was thoroughly assessed. Further, a column was added to the measure for observers to indicate whether the information obtained was via observation (as indicated by the "O" on the measure) and/or report from the teacher or classroom staff ("R"). One scoring benchmark also was modified—a score of 1 now indicates "Minimal/No implementation." In addition, based on pilot testing the fidelity measure, a number of specific guidelines were added to assist observers in better scoring specific items that had lower inter-rater reliability (e.g., "To score this section, observers should ask to see the IEPs of two students with ASD").

TEACCH program and fidelity measure. The TEACCH program was formally established in 1972 by Dr. Eric Schopler. The primary theoretical model for TEACCH is cognitive social learning theory. The programmatic features that distinguish TEACCH from LEAP as well as other CTMs include (a) self-contained classrooms for preschool-aged children with ASD often are used, (b) adults in the environment structure learning opportunities for the child, (c) the classroom environment is arranged to address the characteristics of autism children display, (d) a special education teacher is the primary instructor, and (e) the model, as with LEAP, includes a strong parent involvement component.

There was no formal measure of treatment fidelity for the TEACCH model prior to this study. Several checklists had been developed by TEACCH staff and practitioners but no measure included all of the main components of the model (as identified by Dr. Gary Mesibov, model director, personal communication, 2007). Using the available versions (primarily a checklist developed in 2005), project staff with expertise in the model and in collaboration with the directors of the 9 regional TEACCH centers in North Carolina, developed a draft version of the current measure. This version was then reviewed by experts in the field at Division TEACCH and staff members with expertise in autism and CTMs at the Frank Porter Graham Child Development Institute, University of Miami, and University of Colorado-Denver. The instrument was revised based on expert recommendations and circulated again to the TEACCH directors and project staff. After additional revisions, the instrument included 31 items across nine subscales. Subscales are identified in Table 4. As with the LEAP fidelity measure, behavioral indicators were included to provide additional descriptors for each item; additional guidelines were added to items to assist in accurate scoring; and the measure was scored using the same 5-point scale (5 = Full Implementation, 1 = Minimal/No Implementation) and interview protocol. An "observe" or "report" column also was provided.

BAU programs. In this study, typical classes from the community were included to serve as a counterfactual in our examination of discriminant validity, described subsequently. Identified as "business as usual" (BAU), these classes followed an eclectic conceptual orientation and staff did not subscribe to one primary theoretical foundation or program philosophy. Often teachers in these classrooms used a variety of strategies to educate children with ASD. As with TEACCH and LEAP

classrooms, BAU classrooms for this study were selected based on school district administrator's nomination of programs as well as input from local project staff.

2.2. Procedures

After measure development, research staff began a systematic training process to achieve reliability in using and scoring the measures.

Phase 1-video review. First, videos of children in LEAP and TEACCH classrooms were obtained from project staff who were previously affiliated with those models, or from individuals who are currently employed by the model programs.

These individuals were asked to provide videos that represented a range of implementation, and the videos were predetermined by them to be examples of high or low adherence. The qualifications of project staff that made these judgments on classroom adherence included: a former trainer/staff member of the LEAP training team, a certified trainer in the TEACCH model, and a former teacher in the TEACCH demonstration classroom at the University of North Carolina. Four videos (high and low fidelity LEAP and high and low fidelity TEACCH classrooms) were then scored using the fidelity measures. Prior to scoring these videos, all project staff had either directly observed a TEACCH or LEAP classroom or had seen video of a high adherence TEACCH or LEAP classroom. The staff scored both fidelity measures (i.e., TEACCH, LEAP) for each video.

Phase 2- live scoring and reliability checks. Four research staff from three states/sites (NC, CO, and FL) met at one of the study sites to review their scores from the videos and observe and score the measures live in four classrooms across two school districts (1 TEACCH, 1 LEAP, 2 BAU). Prior to conducting the live observations, the staff reviewed their scores from the videos to resolve disagreements and reach consensus on scoring difficult items. Again, staff scored both fidelity measures in each classroom and one of the staff conducted teacher interviews for the four classrooms. Staff scored the fidelity measures independently then met to calculate interobserver agreement (IOA) on each measure for each classroom. Percentage of interobserver agreement was calculated by dividing the number of agreements (scores of the two raters had to be within +/− 1 point) by the number of agreements plus disagreements and multiplying by 100. Mean percentage agreement for each measure across the four classrooms in the training phase were: 85% for the TEACCH measure (range across domains 75–100%) and 85% for the LEAP measure (range across domains 69–95%). After the reliability checks, research staff revised 13 items across the three measures to provide additional clarity for observers.

Phase 3-video review and reliability checks. After the live scoring and measure revisions, research staff re-scored the videos from Phase 1. Again, both measures were scored for each of the four videos. Percentage of interobserver agreement was calculated using the same formula from Phase 2. Mean percentage agreement for each measure across the four videos and staff members increased to: 94% for the TEACCH measure (range across domains 84–100%) and 86% for the LEAP measure (range across domains 72–100). *Phase 4-training of additional staff*. Research staff then developed consensus coded fidelity measures based on the work completed in Phase 3. These consensus forms and videos were used to train additional research staff. Each staff member was required to reach 80% agreement across all videos and measures.

2.3. Participants

Research staff conducted observations in 34 classrooms (11 TEACCH, 10 LEAP, and 13 BAU classrooms) in the same three states. The following steps were followed in classroom selection: (1) LEAP model developer, Phil Strain, and TEACCH model director, Gary Mesibov, assisted research staff in locating school districts with classrooms using the respective models, (2) school district staff, with knowledge and understanding of the comprehensive models needed for this study identified appropriate classrooms that matched the model descriptions provided by project staff, (3) BAU classrooms were identified by school district staff who had familiarity with non-model specific preschool classrooms, and (4) potential classrooms/teachers met specific inclusion criteria. Teachers/classrooms met these inclusion criteria: (a) classrooms were in public schools, (b) teachers were certified to teach preschool in the state, (c) TEACCH/LEAP teachers had attended formal training in the model, (d) TEACCH/LEAP teachers had taught using the model for at least two years, and (e) BAU teachers had taught students with ASD for at least two years.

LEAP and TEACCH classrooms across the continuum of adherence were included in this study, and teachers across the three classroom types had varied levels of experience and training in the models and/or working with students with disabilities. See Table 1 for demographic information on teachers and classrooms.

2.4. Data collection

The current study was then conducted to obtain psychometric data on the measures (i.e., reliability and validity). Across the four-month period, 128 observations were conducted by a primary observer across sites (2–4 observations per classroom, conducted every 4–8 weeks). In addition, 66 observations were conducted with a reliability observer across sites (1–2 per classroom). At each observation both fidelity measures were completed, along with teacher interviews. To ensure consistency across sites, all observations occurred during the first 3 h of the classroom day, as some classrooms had a shorter instructional day (half day sessions versus full day sessions). Teacher interviews took 20–30 min to complete and were conducted on the day of the observation. See Table 2 for additional information about number of classrooms and observations conducted across models.

Table 1
Teacher and classroom demographic data.

Variable	Level	Number ($n = 34$)		
		BAU	LEAP	TEACCH
Education	AA	0	1	0
	BS/BA	4	4	5
	MEd/MS/MA	8	5	5
	Above MEd/MS/MA	1	0	1
Ethnicity	Non-Hispanic	8	7	8
	Hispanic	5	3	3
Race	White	12	10	11
	(Missing)	1	0	0
Gender	Female	13	9	11
	Male	0	1	0

Variable	Mean (SD)		
	BAU	LEAP	TEACCH
Years teaching	13.7 (11.1)	11.6 (6.2)	7.7 (5.46)
Years teaching children with ASD	5.6 (4.3)	7.5 (4.5)	6.4 (3.5)
Children with ASD per class	2.9 (2.9)	3.3 (0.95)	5.8 (2.9)
Typically developing children per class	4.7 (4.2)	7.5 (1.5)	1.1 (2.4)

2.5. Data analysis

For each of the fidelity measures, four psychometric properties were examined. First, interrater agreement analysis examined the stability of scores across raters. Interrater agreement was determined through the computation of near-agreement ratios (de Vet, Terwee, Knol, & Bouter, 2006). Second, the test–retest reliability was computed in order to determine if the scores were stable over a 2–4 month period. Test–retest reliability was assessed via the computation of intra-class correlation coefficients (ICCs) assessing the proportion of common variance in scores across four rating occasions (Shrout & Fleiss, 1979). Third, the internal consistency reliability of each measure and its subscales were examined via the calculation of Cronbach's alpha coefficients. Finally, the discriminant validity of each measure was examined. This analysis tested the degree to which each fidelity form could discriminate between the two treatment models and BAU comparison group. Because the sample in this study was too small to use logistic regression, we used descriptive discriminant analysis (Huberty & Olenjik, 2006) to perform this analysis. Though we were interested in the construct validity of the measure, our sample was insufficiently sized for factor analysis; yet, discriminant analysis provided an empirical approach to examine validity of the three measures.

3. Results

3.1. Interrater agreement

Interrater agreement was determined by the following procedure. Pairs of raters completed the fidelity forms for each classroom on the same day during two out of four observation sessions with those sessions separated by two to three months. The pairs of raters did independently score the measures. Due to logistical reasons, two of the 34 classrooms were visited by a pair of raters in only one session. Each fidelity measure was completed for each classroom. Subscale and total scores for each pair of raters were examined. A proportion of agreement index was calculated for each pair. If the paired raters' scores differed by one point or less, the raters were coded as having agreed; otherwise a disagreement was recorded. Sixty four pairs of ratings were recorded for each instrument. Interrater agreement was high for the overall score on the LEAP (96.9%), and TEACCH (95.3%) measures.

Table 2
Number of classroom types and observations per model.

	# of classrooms across models	# of observations across models
NC	TEACCH: 6	TEACCH: 21 P (11 R)
	LEAP: 0	LEAP: 0
	BAU: 4	BAU: 15 P (8 R)
CO	TEACCH: 1	TEACCH: 1 P (1 R)
	LEAP: 5	LEAP: 20 P (10 R)
	BAU: 3	BAU: 11 P (6 R)
FL	TEACCH: 4	TEACCH: 16 P (8 R)
	LEAP: 5	LEAP: 20 P (10 R)
	BAU: 6	BAU: 24 P (12 R)
	Total: TEACCH: 11	Total: TEACCH: 38
	LEAP: 10	LEAP: 40
	BAU: 13	BAU: 50
	Total: 34	Total: 128

Note: P = primary observation; R = reliability observation.

3.2. Test–retest reliability

A trained rater visited each classroom a maximum of four times over the course of the school year and completed each of the three fidelity measures for each classroom. Test–retest reliability was measured by fitting an unconditional multilevel model (Raudenbush & Bryk, 2002) to scores for each fidelity measure. Fitting the multilevel model led to estimates of two variance components. A within-classroom variance component represented the variability of each classroom's score about its grand mean, while a between-classrooms variance component represented the variability between classrooms that was stable over time. The test–retest reliability coefficients were computed by dividing the between-classrooms variance by the total variance, which is the sum of the within- and between-classroom variance components. The resulting ratio is called an intra-class correlation coefficient (ICC), and represents the proportion of variance in the outcome that varies between, rather than within, classrooms. Therefore, the ICC represents the proportion of variance in the scores that is stable over time. Again, the ICC for the total score was comparable across measures with test–retest scores for the LEAP measure = 0.748 (range: 0.457–0.860), and for the TEACCH measure = 0.805 (range: 0.437–0.861).

3.3. Internal consistency reliability

The internal consistency of each fidelity measure, including subscale scores and total scores, was examined. Subscale scores and total scores were calculated by summing item responses and dividing by the number of items on the scale to yield scores that could range from one to five. Classrooms were rated multiple times by multiple raters in the study. The ratings selected for the internal consistency analysis were from the final rating session and from the rater designated as the primary rater. Because only one rating session could be used, the final rating session was selected arbitrarily; however, sensitivity analysis revealed that the reliability estimates did not vary substantially across rating sessions. Cronbach's alpha was then computed for each subscale score and total score for the LEAP, TEACCH, and BAU fidelity measures. As indicated in Table 3, the reliability for the total item correlation was comparable across measures. The overall internal consistency for LEAP = 0.934 (range for subscales: 0.560–0.904), and for TEACCH = 0.932 (range: 0.428–0.964).

3.4. Discriminant analysis

A descriptive discriminant analysis (DDA) was performed for each fidelity measure to determine the degree to which the subscales of each could distinguish between the three classroom types. The scores were collected from primary raters only and were averaged across the four observation sessions in order to minimize the impact of measurement error on the analysis. The DDA results indicated that each instrument significantly separated the three classroom types. Because there were three groups, two canonical variates (one less than the total number of groups), which represent dimensions of separation between classroom types, were extracted.

The analysis began by extracting two canonical variates for each measure. These canonical variates represent the underlying dimensions that most effectively separated the three groups. Next, the locations of the centroids for the LEAP, TEACCH, and BAU groups on the two canonical variates were computed. The distance between centroids could be measured by calculating the squared Mahanobis distance between each pair of classroom types. These distances are assumed to follow an F distribution, so statistical tests of the separation of the group centroids could be performed. Finally, inspection of the pooled within-classroom canonical coefficients, which are somewhat similar to factor loadings from factor analysis, were examined to determine which subscales contributed most strongly to each canonical variate.

LEAP measure. Examination of the canonical coefficients revealed that the first canonical variate was defined primarily by the promoting social interactions subscale, while the second canonical variate was defined primarily by the teaching strategies, promoting social interactions, and teaching communication skills subscales The LEAP fidelity measure significantly discriminated between the three models, Wilks' $\lambda = .094$, $F(16, 48) = 6.77$. Both canonical roots were statistically significant. In

Table 3
Reliability for LEAP fidelity measure.

Subscale (*n* items)	Internal consistency (Cronbach's alpha)	Test–retest (ICC)	Interrater agreement
Total score (38)	0.93	0.75	96.9%
Interactions with families (5)	0.72	0.62	96.9%
Interactions with children (4)	0.76	0.65	95.3%
Measuring progress (4)	0.81	0.46	93.7%
Providing positive behavioral guidance (5)	0.87	0.67	93.8%
Promoting social interactions (6)	0.90	0.86	82.3%
Teaching communication skills (4)	0.58	0.50	98.4%
Teaching strategies (5)	0.83	0.62	90.6%
Organization and planning (5)	0.56	0.58	95.3%
Mean	0.77	0.63	93.7%
SD	0.13	0.12	4.8%

Note: Internal consistency assessed at final time point. Test–retest reliability assessed over four time points. ICC represents proportion of variance that did not vary across time.

Table 4
Reliability for TEACCH fidelity measure.

Subscale (*n* items)	Internal consistency (Cronbach's alpha)	Test–retest (ICC)	Interrater agreement
Total (31)	0.93	0.81	95.3%
Family involvement (2)	0.88	0.68	96.9%
Behavior management (3)	0.43	0.44	89.1%
Social leisure (3)	0.82	0.60	76.6%
Communication (4)	0.69	0.59	93.8%
Assessment and teaching (4)	0.86	0.44	85.9%
Visual structure (2)	0.93	0.72	89.1%
Work systems (4)	0.96	0.86	90.6%
Visual schedules (4)	0.94	0.85	87.5%
Physical structure (5)	0.72	0.55	93.8%
Mean	0.81	0.65	89.9%
SD	0.17	0.16	5.9%

Note: Internal consistency assessed at final time point. Test–retest reliability assessed over four time points. ICC represents proportion of variance that did not vary across time.

discriminant analysis, tests of significance for canonical variates are performed on a cumulative basis. Thus, the first test is a joint significance test for both variates, and its test statistic is identical to the overall multivariate test reported above. The second test is a test of the second canonical variate after removing the variance explained by the first canonical variate. The test of the second variate was significant $F(7,25) = 3.38$, $p = .011$. The coordinates for the centroids of each group on the two canonical variates were as follows: LEAP (2.592, 0.802), TEACCH (−2.533, 0.664), and BAU (0.149, −1.178). The LEAP fidelity measure significantly separated all three classroom types from one another. BAU classrooms were separated from LEAP classrooms, $F(8, 24) = 5.407$, $p < .001$; BAU classrooms were separated from TEACCH classrooms, $F(8, 24) = 6.104$, $p < .001$; and LEAP classrooms were separated from TEACCH classrooms, $F(8, 24) = 13.321$, $p < .001$. The first variate separated all three groups (all $p < .001$), but maximized the distance between LEAP and TEACCH classrooms. The second variate separated both TEACCH and LEAP from BAU (both $p < .001$), but did not separate TEACCH from LEAP ($p = .754$) (Fig. 1).

TEACCH measure. Examination of the canonical coefficients revealed that the first canonical variate was defined primarily by the visual schedules, physical structure, assessment and teaching, and visual structure subscales, while the second canonical variate was defined primarily by the work systems subscale. The TEACCH fidelity measure significantly discriminated treatment models, Wilks' $\lambda = .051$, $F(18, 48) = 8.75$, $p < .001$; as described previously, this represents a joint test of the two canonical variates. The second canonical variate significantly discriminated groups as well, $F(8, 24) = 8.68$, $p < .001$. The coordinates for the centroids of each group on the two canonical variates were as follows: LEAP (1.530, −2.156), TEACCH (1.488, 1.981), and BAU (−2.435, −0.018). The TEACCH fidelity measure significantly separated all three classroom types from one another. BAU classrooms were separated from LEAP classrooms, $F(9, 23) = 9.455$, $p < .001$; BAU classrooms were separated from TEACCH classrooms, $F(9, 23) = 9.522$, $p < .001$; and LEAP classrooms were separated from TEACCH classrooms, $F(9, 23) = 7.392$, $p < .001$. The first variate separated LEAP and TEACCH classrooms from BAU classrooms (both $p < .001$) but did not separate TEACCH from LEAP ($p = .924$). The second variate separated all three classrooms (all $p < .001$), but maximized the distance from TEACCH to LEAP classrooms (Fig. 2).

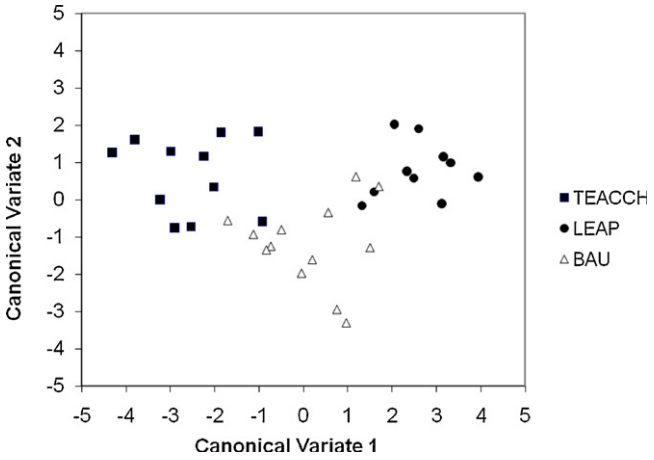

Fig. 1. LEAP descriptive discriminant analysis results ($n = 34$).

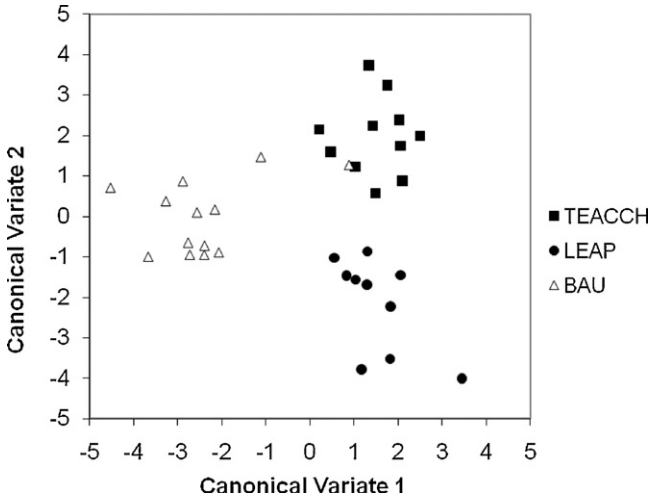

Fig. 2. TEACCH descriptive discriminant analysis results ($n = 34$).

4. Discussion and conclusions

The purpose of this study was to analyze the psychometric properties of assessments of implementation for two established CTMs for children with ASD. The results indicated that the TEACCH and LEAP measures were reliable based on several reliability analyses. The total intra-class correlation (ICC) coefficients indicated excellent test–retest reliability across measures and time points (per Cicchetti, 1994, excellent ICC is .75–1.00). Inter-rater reliability across observers was high, in general, (range for subscales: 76.6–98.4%) as was the overall internal consistency of the measures (.93 for both measures).

The results also indicated that specific canonical variables that assessed critical components of the CTM effectively discriminated the models from one another and from the BAU classrooms. On the TEACCH measure, canonical variate 2, which was primarily defined by the work system subscale, significantly separated TEACCH classrooms from both LEAP and BAU classrooms. It appears that work systems, which are visual instructions designed to help the learner perform skills independently, are unique structural features of this CTM (Hume & Odom, 2007; Mesibov et al., 2005), and are unlikely to be found in non-TEACCH classrooms serving young children with ASD. Canonical variate 1, which included subscales related to physical structure, assessment and teaching, visual schedules, and visual structure, significantly separated TEACCH from BAU and LEAP from BAU classrooms, but not TEACCH from LEAP. This is not surprising, as critical components of both TEACCH and LEAP (the two autism-specific models), include physical organization of the classroom environment, and include direct and explicit instruction using a variety of methods (e.g., picture schedules) throughout the school day (Mesibov et al., 2005; Strain & Hoyson, 2000). Further, both models prioritize instruction based on the core deficit areas of ASD (Mesibov et al., 2005; Strain & Hoyson, 2000); however, the teaching strategies and instructional emphases are quite different (a more structured approach in TEACCH classrooms and more naturalistic approach in LEAP classrooms).

On the LEAP measure, canonical variate 1, which was primarily defined by the promoting social interaction subscales, significantly separated LEAP from both TEACCH and BAU classrooms, and maximized the separation of LEAP and TEACCH classrooms. The results likely reflect the differences between the models in how social skill instruction is approached. LEAP classroom teachers are trained in peer-mediated instruction and intervention, a social skills training approach developed to support interactions between children with ASD and their typical classroom peers (Goldstein, Kaczmarek, Pennington, & Shafer, 1992) and this is a critical component of the LEAP model; whereas, the TEACCH approach emphasizes the use of structured, cooperative play tasks to promote social interaction (Reynolds & Hume, 2009). Further, the method of social skill instruction likely differs between the two models because the make-up of the children in the classrooms vastly differs, with LEAP classrooms containing both children with ASD and their typical peers while TEACCH classrooms in our study primarily contained only children with ASD. Canonical variate 2, subscales related to teaching strategies, promoting social interaction, and teaching communication skills, significantly separated both models from BAU classrooms. This separation may have resulted from teachers in BAU classrooms using fewer explicit behavioral, or perhaps autism-specific, teaching techniques (e.g., task analysis, total communication, hierarchy of behavioral prompts) and instead using more general intervention methods that meet the needs of students across disabilities (i.e., embedded instruction in ongoing activities, verbal interaction during routines and activities, acknowledging efforts and positive behaviors).

As discussed previously, the use of implementation measures in intervention studies has been quite limited, both in the field of ASD and across the broader field of education. However, recent requirements by funding agencies and a sense of urgency in the field of ASD to conduct more rigorous efficacy studies, require that implementation measures become integral components of research proposals and plans (Albro & O'Donnell, 2010; Smith, Scahill, et al., 2007). For example, the Institute of Education Sciences (IES), an agency of the Department of Education, is currently funding eight large scale intervention studies focused on individuals with ASD. The agency now requires researchers to conceptualize fidelity during intervention

development, assess fidelity during efficacy and scale-up studies, and examine the impact of fidelity in data analysis (Albro & O'Donnell, 2010). This requires that researchers begin by carefully integrating implementation measurement during the planning stages of efficacy research (Smith, Daunic, & Taylor, 2007), as modeled in this study and the larger treatment comparison project. In addition, several groups of researchers in the field of ASD have developed recommendations related to study design and methodology when conducting efficacy studies (Charman & Howlin, 2003; Smith, Scahill, et al., 2007). In their summary of recommendations from18 autism researchers in the UK to improve the overall quality of autism intervention research, Charman and Howlin (2003) state "measures of treatment fidelity. . .should be standardized and included for control and intervention groups" (p. 222). An additional working group of autism researchers published similar recommendations in 2007 highlighting the need for "systematic monitoring of intervention fidelity" throughout efficacy studies (Smith, Scahill, et al., 2007, p. 361). The recommendations related to fidelity measurement in both active treatment and control groups were adhered to in the current study and the larger ongoing efficacy research study.

There are three additional reasons that the process outlined in this study is beneficial as the initial phase in an efficacy study, and that the use of validated implementation measures should be an integral component throughout the research process. First, treatment implementation should be explored as a potential mediating variable to account for individual differences in participant's outcomes. Research indicates that implementation is a significant influence on outcomes (Durlak & DuPre, 2008), however little is known about the impact of varied levels of implementation (i.e. high, intermediate, low levels), how implementation data relate to gains achieved by different subgroups of participants (e.g., functioning level, gender), how far one can deviate from implementation protocol and still achieve optimal outcomes, and which aspects and levels of implementation are necessary to achieve the best results (Durlak & DuPre, 2008; Gresham et al., 2000). It also may be important to measure the relationship between changes in implementation quality over the course of a planned treatment period, for example a school year, and participant outcomes (Wolery & Garfinkle, 2002).

Second, replication studies are important to confirm the reliability and validity of participant outcomes across investigators and settings. Replication requires comprehensive and clear specifications of intervention procedures and an evaluation of whether the procedures were implemented as anticipated—best captured by valid implementation measures. In addition, CTMs are often replicated across sites, independent of the original model developer. In Odom et al.'s review (2010) of CTMs, 14 of the 30 models reported that two or more independent sites had replicated their model. What is concerning, however, is that a number of CTMs reported replication of their models in community settings while also reporting that only brief or no instrumentation for assessing implementation were developed or used (Odom et al., 2010). Unless a researcher knows precisely what was done, how it was done, and how long it was done, replication would be difficult to achieve (Gresham et al., 2000).

Third, implementation data also provides researchers valuable information about intervention feasibility and acceptability- key issues when "scaling up" or conducting community effectiveness studies. Studying practitioners' fidelity to an intervention in practice exposes important information about the likelihood that an intervention can and will be implemented with fidelity in the broader community (O'Donnell, 2008). Research indicates that acceptable treatments are more likely to be implemented with greater integrity, providing researchers with additional information about the likelihood of continued implementation without their ongoing involvement (Gresham et al., 2000).

This study is one of the few that has attempted to establish the psychometric properties of implementation measures, and while relatively unique in this regard there are limitations associated with the study. A primary limitation of the study is that the implementation raters were not blind to the type of classroom in which they observed. We purposefully decided not to use blind raters for two reasons (1) to accurately score the measure the rater must have some knowledge of TEACCH and LEAP classroom features and practices, and (2) it would have been difficult to maintain blindness given that the actual physical layout of the classrooms look different (e.g., work systems in TEACCH classrooms) and the types of children in those classroom are different (e.g., presence vs. no presence of typical peers). Other limitations exist when conducting in vivo observations, including behavioral reactivity and difficulty ensuring independent ratings. The frequency of observations (monthly visits) and the regularity of reliability observations (conducted for at least 50% of observations) may have assisted in reducing these threats and ensuring accuracy in scoring the measures. Another limitation is the small sample size, which limited our ability to use more advanced data analysis strategies such as logistic regression. Achieving a larger sample was made more complicated by the fact that the analyses had to be performed at the classroom level versus the child level, that is, even though there may have been three children with ASD in a particular class, the fidelity score was based on classroom wide implementation of the CTM. Finally, we have only validated implementation measures for two CTMS when as many as 30 have been identified (Odom et al., 2010); however, this paper provides a reasonable approach for validation of other implementation tools. The next step is to examine the association between these psychometrically validated implementation measures and child and family outcomes.

Author note

Full measures are available from the first author and excerpts can be found in Appendix A.

Acknowledgments

Development of this paper was partially supported with funding from the Institute of Education Sciences, U.S. Department of Education (R324B070219). The opinions expressed by the authors are not necessarily reflective of the position, or endorsed by, the U.S. Department of Education.

Appendix A

Excerpt from TEACCH Fidelity Measure

Social and Leisure	Full Implementation		Partial Implementation		Minimal /No Implementation	O	R
24. Activities addressing leisure and social skills are designed to match student's developmental level, strengths, and needs • Informal assessment is used to determine student's social and leisure goals, as well as in the design of social/leisure activities • Social skills activities are appropriate to student's developmental level (i.e. proximity, parallel, turn taking, rules) • Leisure and social activities are planned around individual student interests • Leisure and social activities incorporate appropriate elements of visual structure	5	4	3	2	1		
25. Leisure activities are taught to facilitate student's independent use of free time • Students are actively engaged in leisure activities during free time	5	4	3	2	1		
26. Social skills training focuses on facilitating positive experiences with others • Social skills training facilitates interaction with others • Social skills training occurs in a variety of contexts across the school day (i.e. group activities, paired activities) • Social skills training involves typically developing peers if appropriate peer models are available	5	4	3	2	1		

Comments:

Excerpt from LEAP Fidelity Measure

Promoting Social Interactions

	Full Implementation		Partial Implementation		Minimal /No Implementation	O	R
1. Capitalizes on the presence of typically developing peers	5	4	3	2	1		
• utilizes peers as models of desirable social behavior							
• encourages peer partners/buddies (i.e., hold hands during transitions, play partner, clean up buddy, etc.)							
• demonstrates sensitivity to peer preferences and personalities							
• shows an understanding of developmental levels of interactions and play skills							
2. Utilizes effective environmental arrangements to encourage social interactions	5	4	3	2	1		
• considers peer placement during classroom activities							
• effectively selects and arranges materials that promote interactions							
• effectively selects and arranges activities that promote interactions							
• plans for consistent social opportunities within classroom routines (i.e., table captain, clean-up partner, snack set-up, etc.)							
3. Uses prompting and reinforcement of interactions effectively	5	4	3	2	1		
• provides sincere, enthusiastic feedback to promote and maintain social interactions							
• waits until interactions are finished before reinforcing; does not interrupt interactions							
• models phrases children can use to initiate and continue interactions							
• gives general reminders to "play with your friends"							
• facilitates interactions by supporting and suggesting play ideas							
• ensures interactions are mostly child-directed not teacher-directed during free play							
						O	R

References

Albro, E., & O'Donnell, C. (2010, March). Progressing towards a shared set of methods and standards for developing and using measures of implementation fidelity. *Paper presented at the annual meeting of the Society for Research on Educational Effectiveness*.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*, 199–218.

Charman, T., & Howlin, P. (2003). Research into early intervention for children with autism and related disorders: Methodological and design issues. *Autism, 7*, 217–225.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*, 1033–1039.

Durlak, J. (2010). The importance of doing well in whatever you do: A commentary on the special section: "Implementation research in early childhood education". *Early Childhood Research Quarterly, 25*, 348–357.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.

Goldstein, H., Kaczmarek, L., Pennington, R., & Shafer, K. (1992). Peer-mediated intervention: Attending to, commenting on, and acknowledging the behavior of preschoolers with Autism. *Journal of Applied Behavior Analysis, 25*, 289–305.

Gresham, F, MacMillan, D., Beebe-Frankenberger, M., & Bocian, K. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice, 15*, 198–205.

Hoyson, M., Jamieson, B., & Strain, P. S. (1984). Individualized group instruction of normally developing and autistic-like children: The LEAP Curriculum Model. *Journal of the Division for Early Childhood, 8*, 157–172.

Huberty, C. J., & Olenjik, S. (2006). *Applied MANOVA and Discriminant Analysis* (2nd ed.). NJ: Wiley.

Hume, K., & Odom, S. (2007). Effects of an individual work system on the independent functioning of students with autism. *Journal of Autism and Developmental Disorders, 37*, 1166–1180.

Hume, K., & Odom, S. Best practice, policy, and future directions: Behavioral and psychosocial interventions. In D. Amaral, G. Dawson, & D. Geschwind (Eds.), *Autism Spectrum Disorders*. Oxford University Press, in press.

Koegel, L., Koegel, R. L., Harrower, J., & Carter, C. M. (1999). Pivotal response intervention I: An overview of the approach. *The Journal of the Association for Persons with Severe Handicaps, 24*(3), 174–185.

LEAP-USA Fidelity Form. (2008). *Originally developed by Phil Strain; modified by Autism Spectrum Disorder Treatment Comparison Study*.

Lord, C., Wagner, A., Rogers, S., Szatmari, P., Aman, M., Charman, T., et al. (2005). Challenges in evaluating psychosocial interventions for autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 35*, 695–708.

Lovaas, I. O. (1987). Behavioral intervention and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55*, 3–9.

Mesibov, G., Shea, V., & Schopler, E. (2005). *The TEACCH approach to autism spectrum disorders*. New York: Plenum Press.

Moncher, F., & Prinz, R. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11*, 247–266.

Mowbray, C., Holter, M., Teague, G., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315–340.

National Autism Center's National Standards Project (NSP). (2009). *The National Standards Report*. Retrieved from http://www.nationalautismcenter.org/affiliates/model.php

National Professional Development Center on Autism Spectrum Disorders. (2007). *Evidence Based Practice Briefs*. Retrieved from http://autismpdc.fpg.unc.edu/content/briefs

National Research Council. (2001). Educating Children with Autism. Committee on Educational Interventions for Children with Autism. In C. Lord & J. P. McGee (Eds.), *Division of Behavioral and Social Sciences and Education*. Washington, DC: National Academy Press.

Odom, S. L., Boyd, B., Hall, L. J., & Hume, K. (2010). Evaluation of comprehensive treatment models for individuals with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 40*, 425–436.

Odom, S. L., Brown, W. H., Frey, T., Karasu, N., Smith-Carter, L., & Strain, P. S. (2003). Evidence-based practices for young children with autism: Evidence from single subject design research. *Focus on Autism, 18*, 176–181.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*, 33–84.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reynolds, B., & Hume, K. (2009, May). An emerging technology: Using structured teaching to enhance joint attention in young children with ASD. *Lecture presentation at Association for Behavior Analysis*.

Rogers, S., & Vismara, L. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child and Adolescent Psychology, 37*, 8–38.

Rogers, S. J., Hayden, D., Hepburn, S., Charlifue-Smith, R., Hall, T., & Hayes, A. (2006). Teaching young nonverbal children with autism useful speech: A pilot study of the Denver model and PROMPT interventions. *Journal of Autism and Developmental Disorders, 36*, 1007–1024.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Smith, S., Daunic, A., & Taylor, G. (2007). Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children, 30*, 121–134.

Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., Rogers, S., & Wagner, A. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders, 37*, 354–366.

Strain, P., & Hoyson, M. (2000). The need for longitudinal, intensive social skill intervention: LEAP follow-up outcomes for children with autism. *Topics in Early Childhood Special Education, 20*, 11–122.

What Works Clearinghouse. (2010). *WWC intervention report: Lovaas model of Applied Behavior Analysis*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_lovaas_082410.pdf

Wheeler, J., Baggett, B., Fox, J., & Blevins, L. (2006). Treatment integrity: A review of intervention studies conducted with children with autism. *Focus on Autism and Other Developmental Disabilities, 21*, 45–54.

Wolery, M., & Garfinkle, A. N. (2002). Measures in intervention research with young children who have autism. *Journal of Autism and Developmental Disorders, 32*, 463–478.