Journal on Developmental Disabilities, Volume 13, Number 3, 2007

The Reliability and Validity of a Pilot Version of the York Measure of Quality of Intensive Behavioural Intervention

Helen E. Penn, E. Alice Prichard, and Adrienne Perry

Abstract

Professional consensus panels have strongly recommended Intensive Behavioural Intervention (IBI) as the treatment of choice for young children with autism. Researchers have linked treatment quantity to better results; however, few attempts have been made to link other treatment factors (i.e., quality) to outcome. This study presents information about the reliability and validity of a pilot version of a measure to evaluate the quality of IBI (the York Measure of Quality of Intensive Behavioural Intervention or YMQI). Inter-rater reliability and internal consistency were generally adequate, although some items and categories were weak. Similarly, criterion related validity and construct validity were adequate for total scores and most categories when a subjective approach to coding was used. Although results were promising, future research is needed before the YMQI can be used as a reliable and valid measure or treatment quality. Research implications and directions for future research are discussed.

Autism is a complex developmental disability that involves impairments in communication, difficulties in social interaction, and restrictive or repetitive behaviours (American Psychiatric Association, 1994). Although a variety of treatments have been proposed to treat individuals with autism (see Perry & Condillac, 2003), intensive behavioural intervention (IBI) has received by far the most empirical support (Green, 1996; New York State Department of Health, 1999; Schreibman, 2000). IBI is an intensive treatment (typically one-on-one for 20 to 40 hours per week) for young children with autism based on the principles of applied behaviour analysis (Schreibman, 2000). In a landmark study, Lovaas (1987) reported that 9 of 19 preschoolers

with autism who received at least two years of IBI for 40 hours per week became indistinguishable from their peers by age six. Although this study is not without controversy, critiques have been competently addressed (e.g., Eikeseth, 2001), and a number of more recent studies have supported the effectiveness of IBI in comparison to other interventions (e.g., Smith, Groen, & Wynn, 2000; Eikeseth, Smith, Jahr, & Eldevik, 2002; Howard, Sparkman, Cohen, Green, & Stanislaw, 2005). However, outcomes are variable both within and across studies, and a minority of children have not improved significantly despite treatment (e.g., Lovaas, 1987; Bibby, Eikeseth, Martin, Mudford, & Reeves, 2001).

A number of factors may contribute to differential outcomes for children receiving IBI. To date, research suggests that child factors such as younger age and higher IQ may predict better outcome (e.g., Harris & Handleman, 2000). Furthermore, a higher intensity of teaching per week has also been linked to better results, with previous research emphasizing the importance of children receiving a high quantity of IBI (20-40 hours per week) (New York State Department of Health, 1999). However, it is likely that the quality of treatment received during these hours is also important to consider (Green, 1996; Schreibman, 2000). Initial support for the importance of high quality programs arises from research indicating that programs which may be of lower quality (owing to poor supervision and training, etc.) produce poorer than expected outcomes (e.g., Bibby et al., 2001). Unfortunately, little research has been carried out to determine the key characteristics of quality IBI and to suggest how these characteristics should be measured, limiting the ability of researchers to take the construct of quality into consideration when evaluating treatment effectiveness (Kasari, 2002). Serious concerns have been expressed about the future of the IBI field if quality is not defined and measured (e.g., Jacobson, 2000) and several reviews and critiques of the IBI literature emphasize the importance of linking treatment quality to outcome (Kasari, 2002; National Research Council, 2001; Schreibman, 2000; Wolery & Garfinkle, 2002).

In recognition of this need, our research team has developed an observational tool to assess the quality of IBI (the York Measure of Quality of IBI, or YMQI). A reliable and valid measure of treatment quality would be valuable in a range of clinical and research contexts. The YMQI focuses on the technical quality of one on one teaching and does not address broader systems factors affecting quality (e.g., program design, staff training). In the present study a pilot version of the YMQI is described and its reliability and validity are assessed.

The Development of the YMQI

The YMQI was created based on accepted practices for the development of observational measures (see Hartmann & Wood, 1982). These involve outlining behavioral categories, carrying out pilot observations, creating operational definitions, determining response dimensions, and outlining the measurement context. We consulted a number of sources to determine behavioural categories relevant to quality teaching. These sources included: (a) training manuals published by expert clinicians (e.g., Lovaas, 2003); (b) experimental research suggesting that specific teaching procedures facilitate learning (e.g., Stokes & Baer, 1977); (c) rating scales used by treatment providers to monitor staff performance (Hundert, Walton-Allen, Earle-Williams, Sim, & Cope-Scott, 2000; Leaf, & McEachin, 1999; Provincial Regional Trainers Network, 2004); (d) two empirically supported measures designed to evaluate staff competence within prescribed teaching situations using a specific type of IBI (Davis, Smith, & Donahoe, 2002; Koegel, Russo, & Rincover, 1977). There were no existing empirically-supported measures used to evaluate a wide range of characteristics of IBI in a range of contexts. In order to incorporate clinical judgment into the design of our measure, we consulted with psychologists experienced with IBI during measure design and piloting. Furthermore, we carried out a survey asking parents and professionals about important characteristics of high quality IBI and how they should be measured (see Perry, Prichard, & Penn, in press).

Description of the YMQI

The pilot version of the YMQI evaluated in the present study is applied to videotaped footage depicting at least 20 minutes of any type of IBI. Assessments of quality are made by viewing three randomly-selected 3-minute segments of at least 10 teaching trials.

The YMQI captures information about nine important characteristics of quality teaching (see Table 1). For each of these nine characteristics, 1 to 8 specific behaviours are assessed, leading to a total of 30 evaluated behaviours. For example, the assessment of appropriate use of reinforcement involves an evaluation of whether: (a) reinforcers are delivered quickly, (b) the child appears to be motivated to obtain reinforcers, (c) the level of reinforcement is varied based on the quality of the child's response, and (d) verbal reinforcement appears to be sincere. Scores for specific behaviours within each category are averaged to create nine category level scores, which can be averaged to yield a total overall score.

Table 1. Descriptions of the Categories of the YMQI

Categories	*Description*
$S^{D}s$	Clarity of verbal instructions, verbal instructions are not repeated, directing $S^{D}s$ while child is attending.
Reinforcement	Differential reinforcement, sincere praise, motivating reinforcement, rapid reinforcement delivery.
Prompting	Effectiveness of prompts, timing of prompt delivery, fading and augmenting of prompts, and following through on requests, and the appropriate number of prompts per S^{D}.
Learning	Appropriate level of task difficulty.
Pacing	Inter-trial intervals are not too long, time with reinforcement is appropriate, and a distinct inter-trial interval exists.
Engagement	The child responds appropriately to requests.
Generalization	Wording of $S^{D}s$ is varied, materials are varied, tasks are mixed, teaching takes place away from the table, teaching is embedded into naturalistic activities, response generalization, embedding reinforcement, responding to child initiations.
Problem Behaviour	Evidence of an plan, problem behaviour is not reinforced, and reinforcement of positive behaviour.
Organization	Materials are well organized and data collection does not interfere with teaching.

The pilot version of the YMQI exists in two parallel forms: an Objective Scale and a Subjective Scale. Both scales include the same 30 items measured in different ways. The Objective Scale measures characteristics using 30-second interval recording and takes 2-4 hours to complete for each tape. For example, an item about whether requests are made when children are attending (item 3) requires coders to mark an X for any intervals in which requests are made when a) the child's head is turned more that 90° away from the therapist or task or b) the child is engaged in an activity unrelated to the task. The same items are evaluated by the Subjective Scale using a Likert scale ranging from 1 to 3 with half points (1, 1.5, 2, 2.5, 3). Verbal anchors range from consistently appropriate/no concerns (given a 3),

to generally appropriate/moderate concerns (given a 2), to little evidence of appropriateness/significant concerns (given a 1). Each tape takes up to 75 minutes to code using the Subjective Scale. For each item, specific guidelines are provided about the type of behaviour that would result in ratings of 1, 2, or 3 and coders can provide ratings of 1.5 or 2.5 if the quality of an item falls between these criteria (e.g., for item 3, if most requests are given while the child is attending, but on one occasion the child's head is turned more than 90° from the task during a request, a rating of 2.5 could be given). Although there is a greater degree of subjectivity in the Subjective Scale than in the Objective Scale, the coding manual accompanying the Subjective Scale outlines specific criteria on how to code each item.

On both scales, coders make item-level ratings for all 30 items (e.g., whether reinforcement delivery is rapid). On the Subjective Scale coders also make broad category level ratings for each of the nine categories (e.g., whether Reinforcement is "appropriate"). Therefore, the Subjective Scale produces two different category-level and total scores, one calculated by averaging individual items (without incorporating coders' broad category level scores), called the Subjective Item Scale, and the other obtained by focusing only on coders' broad category level ratings, called the Subjective Broad Category Scale. Results will be presented for each of the Objective Scale, Subjective Item Scale, and Subjective Broad Category Scale.

Reliability and Validity

The purpose of the present study was to evaluate the reliability and the validity of the YMQI. Inter-rater reliability (IRR) was evaluated with percentage agreement and intra-class correlation coefficients. Percentage agreement above 80% is considered good and above 70% is deemed satisfactory, especially for a complex coding scheme (Miltenberger, 2004; Kazdin, 1977). Intra-class correlations are evaluated as follows: 0 to .1 virtually no IRR, .11 to .4 slight IRR, .41 to .6 fair IRR, .61 to .8 moderate IRR, and .81 to 1.0 substantial IRR (Shrout, 1998). In addition, internal consistency was evaluated using Coefficient alpha (Cronbach, 1951). Coefficients above .80 are considered good.

This study also evaluates the criterion-related validity and construct validity of the YMQI. Prior to this study, the content validity of the YMQI was supported through a number of steps taken during measure development, such as carefully specifying behaviours and response dimensions of interest and using multiple sources during instrument development. After the measure was complete, it was systematically compared to five existing staff

evaluation tools (Davis et al., 2002; Koegel et al., 1977; Hundert et al., 2000; Leaf, & McEachin, 1999; Provincial Regional Trainers Network, 2004). This comparison strongly supported the content validity of the YMQI, revealing that 24 of the 30 items were present in at least one of the existing scales and that few items were consistently present in other scales which were not in the YMQI (Penn, 2005).

In order to evaluate criterion-related validity, the selection of relevant criteria for comparison with the YMQI required careful consideration. Scores on our measure cannot be compared with scores from another measure of IBI quality as no empirically-supported tools have been developed that evaluate a wide range of characteristics of IBI in a range of contexts. In our opinion, the most well accepted indicator of quality IBI teaching currently available is expert clinical judgment. Therefore, concurrent validity was assessed by comparing YMQI scores to expert clinical judgments of quality based on the same segments of teaching using Pearson's correlation co-efficient. Correlations between .1 and .3 are considered small, correlations between .3 and .5 are considered moderate, and correlations above .5 are considered strong (see Cohen, 1988). Construct validity was assessed by evaluating intercorrelations of categories, items, and scales. High intercorrelations across items and categories suggest that different components of a scale measure the same underlying construct (in this case, quality). Similarly, high correlations of items within categories support construct validity at the category level.

Method

We obtained videotapes of child-therapist dyads that represented a variety of behavioural approaches, such as discrete trail training (e.g., Lovaas, 2003), incidental teaching (e.g., McGee, Daly, Izeman, Mann, & Risely, 1991), and teaching based on Skinner's analysis of verbal behaviour (Sundberg & Micheal, 2001). Most videotapes came from one of two agencies providing IBI in the Toronto region; some were obtained privately or by word of mouth. There was a 48% response rate from therapists and a 60% response rate from parents of children receiving IBI. Consent was obtained from both parents and therapists of children depicted in the videotapes.

Obtaining Videotapes of IBI

In total we obtained 28 tapes depicting different child-therapist pairings (therapists could be in up to two pairings and children could be in up to three pairings). Tapes depicted 16 children (12 boys, 4 girls) who ranged

in age from 2 to 11 years old. Each child had been previously diagnosed with Autism or Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) independent of this study. There were 22 therapists, 2 of whom were male. Their experience with IBI ranged from 1 to 120 months with a mean of 29 months (*SD=23.5*).

After tapes were obtained, three 3-minute segments depicting at least 10 teaching trials (i.e., learning opportunities) were selected randomly from each videotaped session. In order to ensure viewing consistency across coders, segments were converted into digital format to be viewed on a computer.

YMQI Training and Coding

There were six coders who evaluated tapes using the YMQI. Two coders had designed the measure (primary coders) and the remaining four were recruited for this study (trainees had either no or some previous experience with IBI). The training procedure involved about 30 hours of class time, independent reading, and the rating of 10 tapes. Training for the Objective Scale was longer than for the Subjective Scale and took place first. At the end of the training, a 50-item paper and pencil test was given to the trainees to assess their newly acquired knowledge and all trainees met or surpassed the previously defined criterion of 80% correct. After training was complete, each of the 28 tapes was evaluated by two to four coders (a primary coder and at least one trainee), producing 73 ratings for reliability and validity analysis.

Expert Coding Using the Expert Judgment Scale

To allow for an assessment of criterion-related validity, tapes were also evaluated by four experts who were licensed psychologists and/or Board Certified Behavior Analysts (BCBA) using an Expert Judgment Scale designed for the purpose of this study. Experts had between 8 and 20 years of experience in the autism field and had experience supervising IBI programs and carrying out IBI training. Similar to approaches used in other behavioural research (e.g., Hagopian et al., 1997), expert ratings were made based on the consensus judgment of two experts who viewed tapes concurrently. Experts assessed the overall quality of each tape using a 7-point Likert scale and rated the quality of nine specific characteristics of teaching that corresponded to the nine category level scores in the YMQI using a 3-point scale with half points. Each 9-minute tape took between 15 and 30 minutes to view and code. Experts reported that consensus was easy to achieve.

Results

Inter-rater Reliability

The IRR of the Objective Scale was assessed in two ways. First, we examined the reliability across each pair of raters (*N=13*), which produced percentage agreement scores ranging from 67.8 to 79.4, with all but one pair of raters being above 70% and a mean of 75.5%. Next, the percentage agreement indices across tapes were calculated, yielding coefficients ranging from 63.8 to 82.9 with a mean of 76.3%. Finally, a comprehensive reliability index was computed by averaging across both tapes and coders (consequently weighing the more frequently occurring pairs of coders and the tapes coded by more pairs more heavily), which yielded a coefficient of 75.8%. These results indicate that the Objective Scale has moderate inter-rater reliability . To calculate inter-rater reliability for the Subjective scale, percentage agreement was used and we allowed for a half-point difference between coders (e.g., a rating of 1.5 was considered to agree with a rating of 1 or 2). Across the 13 pairs of raters, agreement ranged from 71.4% to 83.0% with a mean of 77.1%. The 28 child-therapist tapes were assessed, resulting in percentage agreement scores ranging from 64.3 to 90.3 with a mean of 78.0. Finally, the comprehensive reliability index across raters and tapes was 77.9. Therefore, both the Objective Scale and the Subjective Item Scale have moderate IRR across items.

To evaluate reliability further, the IRR of category and total scores was also examined using intra-class correlations (results are presented in Table 2). At the total score level, the Objective Scale was most reliable with an intra-class correlation of .80. The Subjective Scales yielded fair to moderate correlations (.61 for the Subjective Item Scale and .55 for the Subjective Broad Category Scale). At the category level, the two Subjective Scales were slightly more reliable, producing fewer low correlations. Categories that were rated moderately to substantially reliably across all three scales are *Engagement* and *Generalization*. Therefore, when reporting scores at the category level, it appears that these categories have sufficient IRR, no matter which scale is used. In addition, all categories except Reinforcement were moderately reliable (.61 to .80) on at least one of the three scales. IRR analyses at the item level were also conducted and were satisfactory (see Prichard, 2005).

Further IRR analyses for each pair of coders suggested that all raters reached satisfactory levels of agreement at all levels of analysis on all scales. Moreover, levels of agreement did not vary systematically based on coders' previous experience with IBI.

Table 2. Intra-class correlations for Categories and Total Scores on the Objective and Subjective Scales

Category name[a]	*Objective Scale*	*Subjective Scale*	
		Item	*Broad Category*
SDs	.70	.67	.53
Reinforcement	.54	.53	.40
Prompting	.47	.72	.54
Learning	.61	.45	.52
Pacing	.34	.55	.64
Engagement	.77	.76	.71
Generalization	.86	.73	.61
Problem behaviour	.59	.91	.87
Organization	.61	.70	—
Total score	.80	.61	.55

[a] *Descriptions of categories can be found in Table 1.*
The dash indicates that calculation was not possible as Broad Category level ratings of Organization were not obtained due to changes in the design of the YMQI during training.

Internal Consistency

Cronbach's coefficient alpha was used to assess the internal consistency of the scales (six items were removed from the analysis since they were not applicable to more than 50% of the original cases). Scores for each rater were averaged to yield one set of data for each of the 28 tapes. The coefficient alpha for the Objective Scale was .70, indicating moderate to low internal consistency. In contrast, the Subjective Scales had high internal consistency, with coefficient alphas of .89 for the Subjective Item Scale and .91 for the Subjective Broad Category Scale.

Criterion-related Validity

To evaluate criterion-related validity, YMQI scores for each tape were averaged across coders and correlated with expert consensus ratings. Total YMQI scores were compared to experts' evaluations of the overall quality of tapes, and category level scores on the YMQI were compared to expert category level ratings (Pearson's correlation coefficients were used for all analyses). Results for all three scales are depicted in Table 3. As correlations in the present study are based on only 28 data points, smaller correlations could not attain statistical significance and individual results should be considered preliminary.

Criterion-related validity varied across scales and levels of analysis and was generally adequate, especially when a subjective approach was used. Total scores on all three scales had strong criterion-related validity (all coefficients above .48). At the category level, scores for one category on the Objective Scale, three categories on the Subjective Item Scale, and six categories on the Subjective Broad Category Scale were strongly correlated with expert ratings (coefficients above .5). The category *Generalization* had strong criterion-related validity across the three scales, and all other categories aside from *Prompting* and *Problem Behaviour* had moderate or strong criterion-related validity on at least one of the three scales. Criterion-related validity for total scores and most categories was strongest when a subjective approach was used. For four categories (*SDs, Pacing, Engagement*, and *Problem behaviour*), criterion-related validity was higher for the Subjective Broad Category Scale than for the Subjective Item Scale. Additional analyses suggested that the criterion-related validity of total scores varied somewhat across coders, but did not vary systematically based on coders' previous experience with IBI.

It is possible that expert ratings were more strongly correlated with Subjective YMQI ratings because expert ratings were made subjectively. However, our goal was to create a measure which approximated expert judgment as closely as possible. Although expert judgment is necessarily subjective, we feel that it provides the best estimate of true underlying quality, and set out to evaluate whether an objective or subjective tool would best approximate this criterion.

Table 3. Correlation between the YMQI and Expert Ratings for Category and Total Scores

		Subjective Scale	
Category name[a]	*Objective Scale*	*Item*	*Broad Category*
SDs	.34	.33	.53**
Reinforcement	.44*	.53**	.51**
Prompting	.08	.28	.27
Learning	.37	.70**	.73**
Pacing	-.18	.21	.52**
Engagement	.26	.29	.48*
Generalization	.56**	.58**	.56**
Problem behaviour	.29	.02	.24
Organization	.29	.39*	.—
Total score	.48**	.56**	.61**

continued

Table 3. (cont'd)

Note. n = 28 for all categories aside from Problem behaviour (n = 14 for Objective; n = 11 for Subjective) and Learning (n = 27).
The dash indicates that calculation was not possible as Broad Category level ratings of Organization were not obtained due to changes in the design of the YMQI during training.
[a] Descriptions of categories can be found in Table 1.
** p< .05. ** p< .01.*

Construct Validity

Construct validity was assessed by evaluating intercorrelations across items, categories and scales (YMQI scores were averaged across coders and Pearson's correlation coefficients were used). Subjective and Objective Scales were strongly correlated with one another, with only correlations for Pacing lower than expected (.31 and .12 for the Subjective Item and Subjective Broad Category Scales respectively; correlations for all other categories were between .45 and .87). Strong correlations between the Objective and Subjective Scales are not surprising as many items in the Subjective Scale have operational definitions identical to those in the Objective Scale. However, these correlations suggest that a well-specified Likert scale methodology can produce scores which correspond closely to information gathered through very detailed observational recordings taking three to four times longer to complete.

Further assessment of construct validity was done by correlating items and categories in the YMQI with one another. Only intercorrelations for subjective ratings are presented here in order to simplify analyses and because they possess better psychometric properties than objective ratings (e.g., better criterion-related validity, better internal consistency). Table 4 presents intercorrelations for both the Subjective Item Scale and Subjective Broad Category Scale scores. In most cases, categories were moderately to strongly correlated with one another, suggesting that they measure the same underlying construct of quality. On both scales, correlations for the category $S^{D}s$ were relatively low, which suggests that this category may tap a construct that differs somewhat from that measured by other categories.

Lastly, items in the Subjective scale were intercorrelated (producing a correlation matrix too large to reproduce here). Most items were positively correlated with one another, with an average correlation of 0.25. There were four weaker items which were negatively correlated or not correlated with a number of other items in the scale (items about the clarity of verbal instructions, whether verbal instructions were repeated, whether there was a

Table 4. Intercorrelation of Categories on the Subjective Scale

Category name	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*Total*
A. $S^D s$	–	.33	.56**	.19	.12	.31	.17	.22	–	.45*
B. Reinforcement	.26	–	.67**	.74**	.42**	.79**	.79**	.57	–	.86**
C. Prompting	.41*	.56**	–	.60**	.52**	.72**	.59**	.59	–	.83**
D. Learning	-.11	.67**	.32	–	.72**	.72**	.85**	.39	–	.88**
E. Pacing	-.03	.18	.31	.39*	–	.37	.68**	.43	–	.72**
F. Engagement	.27	.54**	.67**	.27	-.07	–	.68**	.61*	–	.84**
G. Generalization	-.05	.63**	.41*	.70**	.19	.33	–	.49	–	.87**
H. Problem behaviour	-.32	.33	.49	.07	.13	.36	.40	–	–	.70*
I. Organization	-.14	.33	.23	.57**	.40*	.05	.46*	.42	–	–
Total score	.24	.81**	.77**	.77**	.43*	.63**	.75**	.49	.61**	–

Note. Intercorrelations of categories on the Subjective Item Scale fall below the diagonal; intercorrelations of categories on the Subjective Broad Category Scale fall above the diagonal.

The dashes under column I indicate that calculation was not possible as Broad Category level ratings of Organization were not obtained due to changes in the design of the YMQI during training.

n = 28 for all categories aside from Problem behaviour (n = 11).

**p < .05. **p < .01.*

distinct intertrial interval, and whether children spent an appropriate amount of time using reinforcers). Comparisons which did not include the above items almost always produced positive correlations, supporting the construct validity of the Subjective Scale as a whole. Supporting construct validity at the category level, items within categories tended to be more correlated with one another than with items in other categories. The mean of correlations within categories was .51, while the mean of correlations outside categories was .22. All items within the categories *S^Ds, Prompting, Organization, Generalization*, and *Problem behaviour* were moderately to strongly correlated with one another (aside from two correlations within the category Generalization which were slight). However, the categories *Reinforcement* and *Pacing* had weaker construct validity (some items within these categories were not correlated with one another and some items correlated strongly with items from other categories).

Discussion

No previous empirically supported measure of IBI quality has been published that is applicable in a range of contexts. Although previous measures have been developed and used for specific purposes, they are either not empirically supported or were designed to evaluate prescribed teaching and assess a limited number of characteristics of IBI.

This pilot study provides information about the reliability and the validity of the YMQI. Results indicate that reliability varied from one scale to another and from one level of analysis to another. Generally, IRR, as measured by percentage agreement and intra-class correlations, ranged from fair to good on both scales at all levels of analysis. The Subjective Scales were more reliable than the Objective Scale at the category level. However, the IRR of the total score for the Objective scale was higher. All categories aside from *Reinforcement* are moderately reliable on at least one scale. Internal consistency was moderate for the Objective Scale and strong for both Subjective Scales.

Analyses carried out prior to this study support the content validity of the YMQI, indicating that it measures a large number of aspects of IBI quality that experts consider to be important. This study demonstrated that criterion-related validity was adequate for total scores on all three scales. Furthermore, criterion-related validity was adequate for most categories when a subjective approach was used. However, the validity of YMQI ratings for *Prompting* and *Problem behaviour* was weak across all scales. Construct validity was adequate, with Subjective and Objective scales strongly correlated with one

another and positive intercorrelations for most items and categories. Items within categories tended to correlate more with one another than with items from other categories, supporting construct validity at the category level (this was true for all categories other than Reinforcement and Pacing).

Results suggest that a subjective approach to evaluating the quality of IBI may be more appropriate than an objective approach. Subjective ratings were similarly or more highly correlated with expert judgment, more internally consistent, and possessed better IRR across items and across categories. Furthermore, subjective ratings took substantially less time to complete. Objective and subjective ratings were strongly correlated with one another, suggesting that a well-defined ratings recording approach (though "subjective") can capture information similar to that measured through detailed observational techniques typically used in behavioural research.

However, additional research is necessary for the YMQI to be implemented as a reliable and valid clinical and research tool. Currently, the psychometric properties of the YMQI are not adequate and changes are necessary prior to its implementation. Certain items should be removed or adapted and training should be altered in an attempt to improve the reliability and validity of weaker categories. A future version of the YMQI would likely combine categories from the two existing Subjective Scales, as categories on the Subjective Item Scale tended to be more reliable but categories on the Subjective Broad Category Scale tended to be more valid. Although the new version of the YMQI would be more subjective in its approach, training would employ a protocol that incorporates a highly detailed observational coding manual and well specified operational definitions. Certain behaviours measured by the YMQI did not occur in all tapes in this study (e.g., items that tapped *Problem behaviour*), which limited our ability to evaluate their reliability and validity. Therefore, when the psychometric properties of a new version of the YMQI are assessed, a larger number of tapes will be used. This would allow for additional analyses, such as an assessment of intra-rater reliability and a further evaluation of construct validity through a factor analysis.

Although modifications are needed to improve the psychometric properties of some categories in the YMQI, the pilot version provides a useful model about important characteristics of IBI and could be helpful to service providers in the meantime. For example, service providers could consult the list of characteristics depicted in Table 1 as a guide when developing training and evaluating staff or programs.

The YMQI holds promise as a meaningful tool to operationalize and promote quality intervention for children with autism. An empirically supported measure of IBI quality would be useful in a range of clinical and research contexts. Clinically, it could be used for performance and program evaluation. In a research context, it would enable the construct of quality to be taken into consideration when evaluating the effectiveness of IBI. It is likely that the quality of IBI that children with autism receive contributes to differential outcomes, and future studies should evaluate quality in addition to other variables such as quantity and child factors. Furthermore, specific aspects of quality (measured using category level scores on the YMQI) could be linked to child outcomes in future longitudinal studies, which would provide much needed information about core components of treatment effectiveness. Although it is clear from this research that the construct of quality is difficult to measure, it is our hope that this study is an important step towards promoting quality assurance in the IBI field.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bibby, P., Eikeseth, S., Martin, N.T., Mudford, O.C., & Reeves, D. (2001). Progress and outcomes for children with autism receiving parent-managed interventions. *Research in Developmental Disabilities, 22*, 425-447.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Davis, B.J., Smith, T., & Donahoe, P. (2002). Evaluating supervisors in the UCLA treatment model for children with autism: Validation of an assessment procedure. *Behavior Therapy, 31*, 601-614.

Eikeseth, S. (2001). Recent critiques of the UCLA young autism project. *Behavioral Interventions, 16*, 249-264.

Eikeseth, S., Smith, T. Jahr, E., & Eldevik, S. (2002). Intensive behavioral treatment at school for 4- to 7- year-old children with autism: A 1-year comparison controlled study. *Behavior Modification, 26*, 49-68.

Green, G. (1996). Early behavioural intervention for autism: What does the research tell us? In C. Maurice, G. Green, & S. Luce (Eds.), *Behavioral intervention for young children with autism: A manual for parents and professionals* (pp. 29-44). Austin, TX: Pro-Ed.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30*, 313-326.

Harris, S.L., & Handleman, J.S. (2000). Age and IQ at intake as predictors of placement for young children with autism: A 4- to 6-year follow-up. *Journal of Autism and Developmental Disorders, 30*, 137-142.

Howard, J.S., Sparkman, C.R., Cohen, H.G., Green, G., & Stanislaw, H. (2005). A comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Research in Developmental Disabilities, 26*, 359-383.

Hundert, J., Walton-Allen, N., Earle-Williams, K., Sim, M., & Cope-Scott, K. (2000). *Intensive Behavioural intervention: A manual for instructor therapists.* Hamilton, ON: Behaviour Institute.

Jacobson, J. (2000). Early intensive behavior intervention: Emergence of a consumer-driven service model. *The Behavior Analyst, 23*, 149-168.

Kasari, C. (2002). Assessing change in early intervention programs for children with autism. *Journal of Autism and Developmental Disorders, 32*, 447-461.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

Koegel, R.L., Russo, D.C., & Rincover, A. (1977). Assessing and training teachers in the generalized use of behavior modification with autistic children. *Journal of Applied Behavior Analysis, 10*, 197-205.

Leaf, R., & McEachin, J. (1999). *A work in progress: Behavior management strategies and a curriculum for intensive behavioral treatment of autism.* New York: DRL Books.

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

Lovaas, O. I. (2003). *Teaching individuals with developmental delays: Basic intervention techniques.* Austin, TX: Pro-Ed.

McGee, G. G., Daly, T., Izeman, S. G., Mann, L., & Risely, T. R. (1991). Use of classroom materials to promote preschool engagement. *Teaching Exceptional Children, 23,* 44-47.

Miltenberger, R. G. (2004). *Behavior modification: Principles and procedures* (3rd ed.). Belmont, CA: Wadsworth/Thomson Learning Inc.

National Research Council. (2001). *Educating children with autism.* Committee on Education and Interventions for Children with Autism. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

New York State Department of Health. (1999). *Autism/Pervasive Developmental Disorders: Clinical practice guidelines technical report.* New York: Author.

Penn, H. (2005). *The validity of the York Measure of Quality of Intensive Behavioural Intervention.* Unpublished Master's thesis, York University, Toronto, ON.

Perry, A., & Condillac, R.A. (2003). *Evidence-based practices for children and adolescents with Autism Spectrum Disorders: Review of the literature and practice guide.* Toronto: Children's Mental Health Ontario.

Perry, A., Prichard, E. A., & Penn, H. (in press). Indicators of quality teaching in Intensive Behavioral Intervention: A survey of parents and professionals. *Behavioral Interventions.*

Prichard, E. A. (2005). *The reliability of the York Measure of Quality of Intensive Behavioural Intervention.* Unpublished Master's thesis, York University, Toronto, ON.

Provincial Regional Trainers Network, Ministry of Community, Family and Children's Services (MCFCS), (2004). *Provincial IBI competencies evaluation form: Draft.*

Schreibman, L. (2000). Intensive behavioural/psychoeducational treatments for autism: Research needs and future directions. *Journal of Autism and Developmental Disorders, 30*, 373-378.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7,* 301-317.

Smith, T., Groen, A.D., & Wynn, J.W. (2000). Randomized trial of intensive early intervention for children with pervasive developmental disorder. *American Journal of Mental Retardation, 105*, 269- 285.

Stokes, T. F., & and Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Anlysis, 10*, 349-367.

Sundberg, M., & Partington, J. (1998). *Teaching language to children with Autism or other developmental disabilities*. Danville, CA: Behavior Analysts.

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.

Authors' Notes

The authors are very grateful for the collaboration of the Toronto Preschool Autism Service at Surrey Place Centre and the Behaviour Institute, who kindly permitted us to recruit therapists and families. We are also indebted to all the staff and families who consented to participate. We extend our thanks to: the coders: Alissa Levy, Sarah Mitchell, Jennifer Snider, and Abbie Solish for their many hours of hard work; the research assistants who entered the data: Ashley Fawcett, Andrea Kapaleris, Jackie Ragolia, and Aliya Rahim; our expert raters: Dr. Rosemary Condillac, Dr. Leslie Cohen, and Shiri Bartman; and Dr. Nancy Freeman for her support and feedback on the two theses. In addition, many other colleagues and friends provided suggestions, feedback, technical assistance, editorial feedback, and emotional support, for which we are most grateful.

Financial support for this work was provided by the Social Sciences and Humanities Research Council (SSHRC), Autism Society of Ontario, and the Canadian Institutes of Health Research via a Strategic Training Initiative in Health Research Grant in Autism Research (STIHR Holden PI).

This work is based on the Master's theses of the first two authors. Portions of this research were presented in poster format at the Ontario Association on Developmental Disabilities' Research Special Interest Group Conference, Barrie, ON, in April, 2005; and in a symposium at the Ontario Association on Behaviour Analysis Conference in Toronto, ON, in November, 2005.

Correspondence

Helen E. Penn
York University
4700 Keele Street
Toronto, ON
M3J 1P3

hpenn@yorku.ca