JoDD
Journal on Developmental Disabilities
Le journal sur les handicaps du développement

Volume 16, Number 2, 2010

Authors

Ksusha Blacklock,
Adrienne Perry

Department of Psychology,
York University,
Toronto, ON

Correspondence

perry@yorku.ca

Keywords

autism,
benchmarks,
intensive behavioural
intervention

Testing the Application of Benchmarks for Children in Ontario's IBI program: Six Case Studies

Abstract

Intensive Behavioural Intervention (IBI) is the treatment of choice for young children with autism. Recently a set of benchmarks was developed for the Ontario IBI program to monitor progress and facilitate clinical decision-making. This paper illustrates the benchmarks process using a case study approach based on six children for whom retrospective file data were available, and addresses questions related to their technical validity. Results indicated that current clinical data contained most of the information needed to evaluate the five steps of the benchmarks, and that these steps were developmentally ordered and demonstrated progress over time. These findings provide preliminary support of the psychometric validity of the benchmarks and their ability to provide a well-structured tool that helps clinicians to make transparent, consistent, and evidence-based decisions.

Autistic Disorder (American Psychiatric Association, 2000) has diverse clinical manifestations, behavioural phenotypes, and developmental dimensions, all of which complicate selecting the appropriate intervention for children. A prominent feature of autism is its variability—some children speak in complete sentences while others will never learn to speak; some children remain aloof while others are affectionate and interested in interacting with others. This great variability is also found in children's response to intervention—some will show limited progress in therapy and others make rapid and remarkable gains (Ben-Itzchak & Zachor, 2007).

Research indicates that Intensive Behavioural Intervention (IBI) may facilitate clinically significant gains in intellectual, social, emotional, and adaptive functioning for children with autism (e.g., Cohen, Amerine-Dickens, & Smith, 2006; Eikeseth, Smith, Jahr, & Eldevik, 2007; Howard, Sparkman, Cohen, Green, & Stanislaw, 2005; Lovaas, 1987; McEachin, Smith, & Lovaas, 1993). Great excitement and controversy has surrounded reports on the effectiveness of early IBI for young children with autism. Most of the outcome studies using intensive behavioural techniques have reported that up to half of the participants made substantial gains on standardized tests, while others made only modest progress, or very little progress in some cases. Unfortunately, in all studies, there were some children who did not appear to benefit from IBI. A recent meta-analysis concluded that "it is imperative children not responding to intervention are identified early so additional and/or different treatments can begin. Therefore, practitioners must continuously monitor the progress being made by all individuals receiving early IBI." (Reichow & Wolery, 2009, p. 39).

In 1999, Ontario launched a province-wide IBI initiative based on research evidence and stakeholder consultation. IBI is funded by the provincial Ministry of Children and Youth Services and provided to children by one of nine regional programs. Outcome studies from the Ontario IBI initiative have demonstrated the effectiveness of this community-based intervention (Flanagan, Perry, & Freeman, under review; Freeman & Perry (2010); Perry et al., 2008), with about 75% of children making measurable progress. However, there has been considerable controversy about who should be eligible for the program, how long children should remain in the program, how discharge decisions should be made, what services children are discharged to, and so on. There is currently a large wait list for the program.

In October 2006, the Ministry of Children and Youth Services established an independent *Expert Clinical Panel* for Ontario's Autism Intervention Program. The Panel's goal was to develop a set of clinical practice guidelines that would enable consistent clinical decision-making in the delivery of IBI services in Ontario (Expert Clinical Panel, 2007). This Panel recommended that benchmarks were necessary in order to monitor each child's progress during IBI, and to ensure transparent clinical decision-making processes regarding the continuation of IBI or the discharge and transition of the child to school or other appropriate services in the community.

Therefore, a second panel was struck to develop the benchmarks. The Benchmark Development Expert Panel based their work on: literature reviews on the benefits of IBI; developmental steps for language and adaptive skill development; norm-referenced standardized measures of language, cognition, and social skills; curriculum-referenced measures of early development; expert survey information; a set of benchmarks in the US known as the Stockton criteria (Region 6 Autism Connection, 2006); and the Early Learning Measure (Smith, Groen, & Wynn, 2000).

The benchmarks have five steps that include skills in six areas: Functional Communication; Receptive & Expressive Language; Nonverbal Cognitive Skills; Readiness; Imitation; and Social & Play Skills (Benchmark Development Expert Panel, 2008).

According to the Panel's report, benchmarks are to be used at each 6-month review period to determine whether or not a child continues in IBI. In addition, standardized assessments are completed at the start of treatment and then every 12 months. Together, these two types of information should be used to make consistent clinical decisions regarding whether children continue in IBI or move to the discharge phase and are transferred to other appropriate services. Children can be discharged for success (meeting benchmark step 5—the final step) or for lack of progress (benchmarks not met). When a child first enters the program, his/her skill level will be used to determine a starting point in terms of the steps in the benchmarks. After 6 months in IBI, it is expected that the child will master at least 75% of the benchmarks in the next step. When a child is discharged from the program, he/she moves to the IBI Reduction and Discharge Phase. This is a 6-month period in which there is a systematic reduction of treatment below 20 hours per week and a corresponding increase in time spent in the typical learning environment such as school, leading to discharge from IBI. A 1-year transition phase follows discharge from IBI, which includes a wide range of services in the community and schools (Benchmark Development Expert Panel, 2008).

There has been speculation that implementation of the proposed benchmarks could cause substantial changes to Ontario's IBI program in terms of improving consistency in clinical decision-making as well as potentially changing discharge processes and reducing waitlist times. However, no empirical evidence is yet available on any of these matters and evidence is urgently needed to help guide our thinking. Detailed discussion of these possible consequences and their policy implications is beyond the scope of the present paper. However, some preliminary data are offered in the present paper to contribute to the discussion.

The current series of retrospective case studies allow us to examine the proposed set of benchmarks, as they would apply to the progress of six children in Ontario's IBI program. The data also allow us to examine several specific questions related to the pragmatics and psychometric validity of the benchmarks:

1. Whether it is possible to find the information in current clinical data (e.g., binders, curriculum-referenced assessments) pertaining to each item in the benchmark steps;

2. Whether it is possible to determine the step that children are at when entering the IBI program (of the 5 steps in the benchmarks); and

3. Whether all the sequenced steps in the benchmarks follow a logical developmental progression and whether the benchmarks confirm a child's progress over time.

Method

This study was approved by York University's Psychology Human Participants Research Committee.

Participants

The case studies were based on a convenience sample of six children (5 boys and 1 girl) for whom data were available to help evaluate the benchmarks. It should be noted that children were not randomly selected and they are not necessarily representative of other children receiving IBI in Ontario. However, they were not selected based on any particular child or family characteristics and were a diverse sample in many respects. They were simply all the available cases with the requisite data.

Children had originally been screened by the Toronto Partnership for Autism Services (TPAS) and deemed eligible for the IBI program, then waitlisted for intervention. The children were subsequently referred to the Perry lab at York University for an updated psychological assessment immediately prior to beginning IBI.

The children were aged 4 years 9 months to 5 years 10 months at the time of the pre-treatment psychological update. This initial evaluation included autism diagnosis, and cognitive and adaptive ability tests. Five children were diagnosed with Autistic Disorder, and one was diagnosed with Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). All children were diagnosed by an experienced psychologist using the established DSM-IV criteria for autism (American Psychiatric Association, 2000). Children were assessed at York University just prior to beginning the IBI program and then again one year later, for another research project (see below for details on measures). As shown in Table 1, children's Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1988) scores at start of IBI ranged from 26 (non-autistic) to 40.5 (severely autistic). Mullen Ratio IQ scores (Mullen, 1995) ranged from 24 (severe/ profound) to 74 (borderline). At start of IBI, children's Vineland-II Adaptive Behavior

Table 1. Participants at start of IBI (n = 6)

Participant	*Sex*	*Age at start of IBI*	*Diagnosis*	*CARS*	*Mullen Ratio IQ*	*Vineland-II ABC Age Equivalent (in months)*
Sean	M	4 years 11 months	Autistic Disorder	33.5	25	18
Leighton	M	4 years 8 months	Autistic Disorder	40.5	24	7
Zahin	M	4 years 9 months	PDD-NOS	28.5	50	27
Arben	M	5 years	Autistic Disorder	38.5	67	30
Myles	M	4 years 8 months	Autistic Disorder	26.0	74	34
Sakari	F	5 years 10 months	Autistic Disorder	32.0	48	29

Note. All names are pseudonyms

Composite (ABC; Sparrow, Cicchetti, & Balla, 2005) score age equivalents ranged from 7 months to 34 months. This heterogeneity is similar to that reported in the larger TPAS outcome study (Freeman & Perry, 2010) and the provincial study (Perry et al., 2008), suggesting these children are not particularly different from others in the program.

Children were all receiving IBI from the TPAS program for the duration of the data collection period. Detailed information was not available regarding service intensity, although five of the six children were in the regular program (approximately 25–30 hours/week) and one was transferred to the school stream during the study period (which is somewhat less intensive: half days in IBI; half days in school).

Procedure

All the participating children's parents were contacted by telephone by the second author. She explained the purpose of the current study, using a standard script, and if they agreed to participate, which all did, they were sent a consent form. Parents signed the informed consent form approving the use of the data obtained by the Perry lab at York University during the initial and one-year follow-up assessments, as well as from children's data binders at various TPAS intervention locations. Parents received no monetary compensation for agreeing to participate.

Files on all six children were reviewed with particular attention paid to the standardized assessments performed at York University, Assessment of Basic Language and Learning (ABLLS) or ABLLS-Revised, individual education plans, supervision notes and team meeting notes. Additionally, each child's Senior Therapist was interviewed when necessary to add to or to clarify information not contained in the file.

Measures

Autism severity

The Childhood Autism Rating Scale (CARS; Schopler et al., 1988) is a behavioural observation measure of the severity of autism, based on observations conducted during the course of a psychological assessment through direct interaction with the child. Ratings are supplemented with a parent report, for items that cannot be observed during the assessment. The CARS contains 15 items, each rated on a scale of 1 to 4, with half-points. The sum of the scores on the individual items is used to obtain a Total Score, with higher scores indicating greater severity. Scores fall within 3 classifications; severe autism, mild/moderate autism, and not autism. The CARS has proven to be a very reliable and valid tool, displaying good internal consistency, high inter-rater agreement, agreement with clinical diagnosis, and meaningful differentiation among clinical groups (Perry, Condillac, Freeman, Dunn-Geier, & Belair, 2005; Schopler, et al., 1988).

Cognitive Ability

At start of IBI, cognitive level was measured for all children using the Mullen Scales of Early Learning, a standardized, norm-referenced measure of children's level of cognitive functioning (Mullen, 1995). A Mental Age (MA) score was obtained, which was based on the median of the Fine Motor, Visual Reception, Expressive Language, and Receptive Language subscales. This score was used to calculate a Ratio IQ (MA/CA×100). For one child, the scores were so different on the subscales that two medians were used: one for nonverbal IQ (Fine Motor and Visual Reception); and another for verbal IQ (Expressive and Receptive Language). At the one-year follow-up, four children were administered the Mullen, while two children were administered the Wechsler Preschool and Primary Scale of Intelligence (3rd ed.), a standardized norm-referenced measure of cognitive functioning (WPPSI-III; Wechsler, 2002). For the WPPSI-III, a Full Scale IQ and MA were calculated. Unfortunately, to most appropriately determine cognitive ability at different points in time, scores from different tests had to be used for some children when starting in the IBI programs and at the one-year follow-up, which is a common issue with autism research (Perry et al., 2008).

Adaptive Behaviour Levels

To assess the adaptive levels of children at the beginning of IBI and at the one-year follow up, a trained interviewer conducted a semi-structured interview with parents based on the Vineland Adaptive Behavior Scales-II, a norm-referenced parent-interview measure of adaptive behaviour, which evaluates the skills displayed in everyday situations (Sparrow et al., 2005). The results from the Vineland-II provided an age equivalent score, which was used as part of the standardized assessment. Specific items were also used in our analyses of whether specific benchmark items were met.

Specific Goal Attainment in IBI

Specific curriculum-referenced measures were on file at TPAS, specifically the Assessment of Basic Language and Learning Skills (ABLLS; Partington & Sundberg, 1998) or the Assessment of Basic Language and Learning Skills-Revised (ABLLS-R; Partington, 2006). Each child in the program had either an ABLLS or an ABLLS-R done at the centre where they were receiving the IBI. These would have been completed by Instructor Therapists (staff who work with the child daily, who have a college or university education and specific IBI training, who receive regular supervision by an MA-level Senior Therapist). Detailed information on these individuals was not available to the researchers. The assessment was performed within 3 months of beginning IBI, around 6 months after starting in the program and 12 months after starting in IBI. The 6-month assessment results, along with any program data in the programs' data binder from that time period helped the researchers determine whether specific items of the benchmark steps were met.

Benchmark Steps Checklists

The researchers developed five checklists, one for each of the five steps of the benchmarks, that included each item of that step as well as the three time points we were interested in (initial assessment, after 6 months in IBI, after 12 months in IBI). All five checklists were completed for each participant, to record all the specific items in all the benchmarks at all the steps over the three time points in order to help answer the research questions of this study.

The child's ABLLS or ABLLS-R results at start of IBI and then every 6 months provided much of the information need to complete the Benchmarks Steps Checklists, along with the Vineland-II and Mullen or WPPSI-III performed at York University when children were starting IBI and at the 1-year follow-up. For example, to find out whether a child achieved item 8 (imitation of 10 familiar motor actions with objects) on the first step of the benchmarks, the researchers looked at the child's results on item D1 of the ABLLS-R. This item asks whether a child can imitate a motor action using an item/object when asked "do this," and the possible answers are 2 actions, 5 actions and 10 actions. If it was noted on the ABLLS-R that a child could imitate 10 motor actions using an item/object when asked to "do this," the researchers scored that the child had mastered item 8 of step 1 of the benchmarks. Often times, the decision of awarding mastery on a specific item was corroborated with other sources. For example, item 4 on step 3 of the benchmarks looks to see whether a child has mastered 100 unprompted labels. The researchers used the Vineland-II Expressive Communication Domain item number 26 in order to find the answer. This item asks whether a child says at least 100 recognizable words. This item can also be corroborated by information from the ABLLS-R assessment. ABLLS-R item G2 asks whether the student labels at least 100 objects which are commonly found in his/her environment and item G4 asks whether the student will label at least 100 pictures of items which are commonly found in his/her environment.

Results

Case Studies (All Names are Pseudonyms)

1. Sean is a boy diagnosed with Autistic Disorder and a moderate intellectual disability. He is 4 years 11 months at the start of IBI. His assessment at entry to the program indicates that his skills fall below step 1 of the benchmarks. Therefore, he would be expected to master at least 75% of the step 1 items after 6 months in the IBI program. In fact, at 6 months duration, he only meets 44% (4 of 9) of the step 1 items and thus (if the benchmarks were in effect) he would be moved to the 6-month discharge phase focusing on transition to an appropriate school placement.

2. Leighton's experience with IBI is very similar to Sean's. Leighton is a 4-year-8-month-old boy diagnosed with Autistic Disorder. At his assessment at start of IBI, it is determined that he has a severe intellectual disability.

He starts IBI below step 1 and is expected to master 75% of step 1 items after 6 months in the IBI program. However, at 6 months, Leighton only meets 33% (3 of 9) of step 1 items and would therefore move to the 6-month discharge phase.

3. Zahin is a boy diagnosed with Pervasive Developmental Disorder - Not Otherwise Specified (PDD-NOS). He is 4 years 8 months old and has a mild intellectual disability. At his assessment just before treatment starts, he has mastered 100% (9 of 9) of step 1 items. At 6 months duration, Zahin has mastered 78% (7 of 9) of step 2 items and would therefore continue in the IBI program. At 12 months duration, Zahin masters 75% (9 of 12) of the step 3 items, shows meaningful gain in age equivalent scores on adaptive behaviour (score of 27 to score of 41 months on Vineland-II) and, although his verbal IQ remains relatively similar, he gains over 20 points on standardized measures of nonverbal IQ (Performance Ratio IQ of 50 to 73). He would therefore continue in IBI.

4. Arben is a 5-year-old boy diagnosed with Autistic Disorder. At his assessment at the start of IBI, it is determined that he has mild developmental delays. It is clear that he has mastered 100% (9 of 9) of step 1 items, but only 67% (8 of 12) of step 2 items, therefore he starts at step 1. At 6 months duration, he masters 100% (12 of 12) of step 2 items and would therefore continue in IBI. At 12 months duration, he masters 93% (13 of 14) of step 3 items, gains at least 10 points on standardized measures of IQ (Ratio IQ of 67 to IQ of 83) and displays meaningful gains in age equivalent scores on adaptive behaviour (from a score of 30 months to 50 months on the Vineland-II). Therefore he would continue in the IBI program.

5. Myles is a 4-year-8-month-old boy diagnosed with Autistic Disorder. He has a mild developmental delay. At his assessment at start of IBI, Myles is determined to start at step 1 of the benchmarks as he has mastered 89% (8 of 9) of the step 1 items. Six months later, Myles masters 64% (7 of 11) of step 2 items, which is close, but not quite the 75% required, and therefore would move to the 6-month discharge phase.

6. Sakari is a girl diagnosed with Autistic Disorder. She is 5 years 10 months old and has a moderate intellectual disability. At her first assessment just before she starts treatment, she is evaluated as having mastered 100% (9 of 9) of step 1 items, and 60% (6 of 10) of step 2 items. She therefore starts at step 1. At 6 months duration she only meets 58% (7 of 12) of step 2 items, indicating essentially no progress. She would move to the 6-month discharge phase and be placed in an appropriate school placement. However, as seen in Table 3 shown later in this paper, she does meet 92% (11 of 12) of the step 2 criteria after 12 months of IBI. In addition, her IQ improves from 53 at start of IBI to 63 after 12 months of IBI.

Table 2 shows the decisions that would have been made about each of the six children at each 6-month interval had the benchmarks been in place.

Table 2. Progress of children in IBI program based on Benchmarks

Participants	*Step at Start of IBI*	*Step at 6 months*	*Continue after 6 months?*	*Step at 12 months*	*Continue after 12 months?*
Sean	Below 1	Step 1 not met	NO		
Leighton	Below 1	Step 1 not met	NO		
Zahin	Step 1	Step 2	YES	Step 3	YES
Arben	Step 1	Step 2	YES	Step 3	YES
Myles	Step 1	Step 2 not met	NO?		
Sakari	Step 1	Step 2 not met	NO?		

Was it Possible to Find Information Pertaining to Each Item in the Benchmark Steps in Current Clinical Data?

We could not always find all the necessary information in order to evaluate each and every step of the benchmarks for each child (in fact, there was a seventh possible participant who had to be dropped due to too much missing data). However, for the remaining six participants included in this study, we were able to find most of the information necessary from a careful review of each participant's files. Due to differences in children's programming at the various centres, as well as current data keeping practices, all children had several items for which the researchers could not find an answer. Clearly when benchmarks are implemented, data keeping practices could be adjusted easily to meet the data requirements. Some items we could not evaluate were common among all participants:

1) All 6-month items were harder to assess, as there was no standardized assessment performed at this time point. As such, the researchers had to rely on data from the child's ABLLS or ABLLS-R assessment as well as any program data from the child's binder, without having the standardized measures for corroboration.

2) All items pertaining to Readiness, such as attending, reinforcement, and group learning skills. We could not tell the amount of time for which children could attend 1:1 or in a group. None of the children had specific enough data regarding their reinforcement schedules in the data examined. The ABLLS and ABLLS-R explore group learning skills in its Group Instruction subscale, however, this information does not match exactly what the benchmark items specify.

3) Item 2 of the additional criteria based on a standardized assessment—Meaningful gains in age equivalent scores on a standardized measure of language. None of the children had any standardized measures of language during their time in IBI.

4) Limited available data made it hard to distinguish some items on step 4 with similar items on step 5 of the benchmarks. Unfortunately the same items on the ABLLS and ABLLS-R were used in order to assess Pretend Play on step 4 (emerging functional pretend play with at least 3 sets of objects) and Pretend Play on step 5 (simple pretend play with adult or peer). There were no data in children's files that would distinguish the two forms of pretend play, so children either mastered both pretend play items or neither item.

Because not all the data for every benchmark item at every step was available, we calculated a percentage of criteria met, out of those with available data, at each step for each child and these percentages were used in subsequent analyses.

Was it Possible to Determine the Step That Children Were at When Entering the IBI Program?

Our next goal was to determine what step of the benchmarks the children were at when they entered the program. We found that determining the step that children started at as they entered the program was very straightforward. As shown in Table 3, two of the participants (Sean and Leighton) started below step 1. This was determined because at the start of IBI they had mastered only 11% of the step 1 benchmark items. This means that by their 6-month assessment they would be expected to master 75% of the step 1 items. Three participants (Zahin, Arben and Sakari) started at step 1, as evidenced by mastering all items of step 1, but less than 75% of items of step 2. One final participant (Myles) started at step 1 as well since he mastered 89% items of the step 1 benchmarks, but less than 75% of the step 2 items. This supports the internal consistency of the items within the steps.

Do All the Sequenced Steps in the Benchmarks Follow a Logical Developmental Progression and Do the Benchmarks Confirm a Child's Progress over Time?

As evidenced by Table 3, the benchmarks do confirm a child's progress over time, since they show each child getting a higher percentage of benchmark items met on a specific step as time goes on. In this way, the benchmarks show

Table 3. Mastered items at each time point for all participants

	Below Step 1	*Step 1*	*Step 2*	*Step 3*	*Step 4*	*Step 5*
Sean						
Start of IBI	*___	1/9 (11%)	0/12 (0%)	1/12 (8%)	0/12 (0%)	0/12 (0%)
6 months		4/9 (44%)	2/12 (17%)	2/11 (18%)	0/12 (0%)	0/12 (0%)
12 months		6/9 (67%)	4/12 (33%)	3/12 (25%)	0/12 (0%)	0/12 (0%)
Leighton						
Start of IBI	*___	1/9 (11%)	0/11 (0%)	0/12 (0%)	0/12 (0%)	0/12 (0%)
6 months		3/9 (33%)	3/13 (23%)	1/11 (9%)	0/12 (0%)	1/12 (8%)
12 months		6/9 (67%)	3/14 (21%)	1/11 (9%)	0/12 (0%)	1/12 (8%)
Zahin						
Start of IBI		9/9 (100%) *	4/12 (33%)	2/12 (17%)	0/7 (0%)	0/7 (0%)
6 months		9/9 (100%)	7/9 (78%)	?	?	?
12 months		9/9 (100%)	7/9 (78%)	9/12 (75%)	3/8 (38%)	2/10 (20%)
Arben						
Start of IBI		9/9 (100%)*	8/12 (67%)	9/12 (75%)	6/12 (50%)	4/12 (33%)
6 months		9/9 (100%)	12/12 (100%)	9/12 (75%)	6/12 (50%)	6/12 (50%)
12 months		9/9 (100%)	12/12 (100%)	13/14 (93%)	6/12 (50%)	7/12 (58%)
Myles						
Start of IBI		8/9 (89%)*	4/11 (36%)	3/12 (25%)	1/12 (8%)	1/12 (8%)
6 months		8/9 (89%)	7/11 (64%)	4/11 (36%)	2/12 (17%)	1/12 (8%)
12 months		9/9 (100%)	9/11 (82%)	8/13 (62%)	5/12 (42%)	1/9 (11%)
Sakari						
Start of IBI		9/9 (100%)*	6/10 (60%)	5/8 (63%)	2/9 (22%)	0/8 (0%)
6 months		9/9 (100%)	7/12 (58%)	7/11 (64%)	2/12 (17%)	3/12 (25%)
12 months		9/9 (100%)	11/12 (92%)	7/13 (54%)	4/12 (33%)	3/12 (25%)

* = *initial step*

a child's progress since children get higher mastered percentages over time. In addition, the benchmarks are developmentally ordered since, in most cases, at the same time point, children mastered higher percentages of step 1 items than step 2 items, higher percentages of step 2 items than step 3 items, and so on.

The only exception lies with some children getting a slightly higher percentage of mastered items on step 5 than on step 4. It is believed that this is due to an improper match of the generalization item the benchmarks are describing and the item that the ABLLS or ABLLS-R is measuring. It is probable that item 8 of step 5 on the benchmarks (good generalization of skills to novel situations and people) implies a much more advanced skill than is measured by the ABLLS-R P2 (generalizes across instructors) and P3 (generalizes across environments) items. This is likely as step 5 is the last and most advanced step of the benchmarks, a step that measures children's ability to function independently in a school environment. Unfortunately, there were no other data available to measure the generalization item (item 8) of step 5 of the benchmarks.

Discussion

Six case studies, based on a convenience sample, were used to illustrate the process and outcomes of the proposed benchmarks for the Ontario IBI program, as well as to investigate some specific questions related to the technical validity of the benchmarks. Overall, the results of this study provide preliminary evidence to suggest that the benchmarks are a well-structured tool that could help clinicians make transparent, consistent, and evidence-based decisions regarding children's progress in Ontario's IBI program.

In order to evaluate the benchmarks for this study the authors had to find assessment items and specific behavioural program data that they could match to benchmark items. Some items of the benchmarks could not be assessed since a matching item could not be found in any of the standardized assessments of current program binders of the participating children. If and when the benchmarks are implemented, it should be much more straightforward to evaluate the benchmarks, as data collection will likely become tailored to the specifications of the benchmark items. This would eradicate the problem of not being able to find the necessary information. At this point standardized measures of language (though done at screening) are not required in the IBI program. However, if the benchmarks were to be implemented, this would become mandatory and would allow for the assessment of children's gains on standardized measures of language. Since teaching and data collection would be based on the benchmarks, it would be more feasible to distinguish mastery of some similar items of the step 4 and 5 benchmarks. This would allow for the evaluation of benchmarks achieved to be more explicit and objective. Fortunately, one aspect that was easy to determine was the step of the benchmarks at which children started. This was extremely beneficial, as from this point one can conclude what step the child should meet after 6 months in the program and after 12 months in the program.

The results support the technical validity of the proposed benchmarks in that the items within a step seem to converge, follow a logical developmental progression, and demonstrate sequential progress over time. At any one time point, children have fewer mastered items as the benchmark steps get higher. This is an important validation of the benchmarks, showing that they are well designed to measure a child's natural progress through the IBI program. Although there was one exception to these results, it appeared to be due to unavailability of the precise data required and was not due to a fault in the benchmarks. Again, if the benchmarks are implemented and the program starts to reflect the benchmark goals this will no longer be a problem.

This study showed that two of the six children in the case studies would clearly have continued in the IBI program had the benchmarks been in place. Two would have been moved to the discharge phase after a 6-month trial of IBI (whereas, in reality, they continued in the program). The data show that these children did not, in fact, make much progress between the 6- and 12-month assessments, so it appears that the decision to move these children to the discharge phase would have been an appropriate clinical decision. Finally, two children did not quite meet enough criteria to have continued in IBI but this was unclear from the present study. Once again, had the researchers had all the necessary data for the benchmark item specifications and had the children been taught to the benchmark specifications, it is very possible that these children would have continued in the IBI program. Although these two children did not meet step 2 of the benchmarks after 6 months in the program, based on our data collection, we can see that they did meet step 2 of the benchmarks after 12 months in the program (Table 3). Data from these two cases suggest that a certain degree of clinical judgment would be advisable in implementing the benchmarks. It might be important to consider issues related to intensity of IBI and other significant clinical factors, which were not collected as part of this evaluation, but that may well be at play.

Study Limitations and Future Directions

A significant limitation of the current study is certainly the small sample size. Clearly, no claims can be made that these six case studies are representative of other children in the program, although they were also not systematically biased in any apparent way and the children were similarly variable in developmental and diagnostic parameters to the sample of 332 children in the provincial outcome study (Perry et al., 2008). It was simply a sample of convenience that presented the opportunity to examine, in a descriptive way, several questions related to the implementation of the proposed benchmarks. We hoped to use these data to provide some empirical evidence to contribute to the debate surrounding these issues.

Another limitation was that a certain amount of clinical judgment was required of the researchers in some circumstances when deciding whether a child had mastered a specific item of the benchmarks, because of the data availability issues mentioned. In the future, if the benchmarks are implemented, much of this judgment would be eliminated as children in the program would likely be taught to the benchmark specifications, and therefore data for all the specific benchmark items being taught and mastered would be available.

Further, there were limited data at the 6-month time point. This is because there was no standardized assessment performed at this time, so the researchers had to rely on children's binder data and ABLLS or ABLLS-R data in order to see whether specific items were mastered by children after 6 months in IBI. This problem would be alleviated if the benchmarks are implemented, as the Benchmark Development Expert Panel (2008) suggests that a Vineland-II be administered at the 6-month assessment. It would be very useful to have the Vineland-II results at 6 months in order to corroborate the results of the 6-month assessment based on the ABLLS or ABLLS-R and binder data and to include the parents' perspective in the data.

In future research, it would be helpful to address some of the same questions we have examined with a much larger sample size, using children from all across Ontario. Once the benchmarks are implemented, it would be helpful to continue research on their psychometric properties, including the developmental ordering and whether they continue to show a logical progression of children through the program. It would also be helpful to test the interrater reliability of the assessment of mastery of benchmark items, to determine whether different assessors would get the same results, when assessing whether children have mastered specific benchmark items. This will be extremely important clinically if the benchmarks are implemented, as they would be used to make very important decisions about whether children continue in the IBI program or whether they are discharged and moved to a different service that is more appropriate to their needs.

Conclusion

Despite all its current limitations, this line of research is very important to Ontario's IBI program. The implementation of the proposed benchmarks is a highly political and emotional endeavor for many stakeholders. Research is needed in order to generate better evidence which can be used to address stakeholders' concerns, questions, and suggestions. Any research examining the reliability, validity, and helpfulness of these proposed benchmarks to clinicians, parents, and policy makers can only broaden our knowledge and aid in fair and sound decision-making surrounding this very important issue.

Disclosure

Dr. Adrienne Perry was a member of the Benchmark Development Expert Panel.

Acknowledgements

We would like to thank all the members of the Benchmark Development Expert Panel, especially Dr. Nancy Freeman, and all the parents who consented to their children's participation in our study.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*, (4th ed., text revision). Washington, DC: American Psychiatric Association.

Benchmark Development Expert Panel. (2008, September). *The Development of Benchmarks for the Delivery of Intensive Behavioural Intervention for Children with Autism*

Spectrum Disorders in Ontario. Unpublished report to the Ministry of Children and Youth Services.

Ben-Itzchak, E., & Zachor, D. (2007). The effects of intellectual functioning and autism severity on outcome of early behavioral intervention for children with autism. *Research in Developmental Disabilities, 28,* 287–303.

Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Journal of Developmental and Behavioral Pediatrics, 27*(2), S145–S155.

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2007). Outcome for Children With Autism Who Began Intensive Behavioral Treatment Between Ages 4 and 7: A Comparison Controlled Study. *Behavior Modification, 31,* 264–278.

Expert Clinical Panel. (2007, September). *The Development of Clinical Practice Guidelines for the Delivery of Intensive Behavioural Intervention for Children with Autism Spectrum Disorders in Ontario.* Unpublished report to the Ministry of Children and Youth Services.

Flanagan, H. E., Perry, A., & Freeman, N. L. (under review). The effectiveness of community-based intensive behavioural intervention for children with autism: A waitlist comparison study exploring outcomes and predictors. Manuscript submitted for publication.

Freeman, N., & Perry, A. (2010). Outcomes of intensive behavioural intervention in the Toronto Preschool Autism Service. *Journal on Developmental Disabilities. 76*(2), 17–32

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

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

McEachin, J. J., Smith, T., & Lovaas, O. I. (1993). Long-term outcome for children with autism who received early intensive behavioral treatment. *American Journal on Mental Retardation*, *97,* 359–372.

Mullen, E. M. (1995). *Mullen Scales of Early Learning.* Circle Pines, MN: American Guidance Service.

Partington, J. W. (2006). *The Assessment of Basic Language and Learning Skills-Revised.* Pleasant Hill, CA: Behavior Analysts.

Partington, J. W., & Sundberg, M. L. (1998). *Assessment of basic language and learning skills (The ABLLS): An assessment for language delayed students*. Pleasant Hill, CA: Behavior Analysts, Inc.

Perry, A., Condillac, R. A., Freeman, N. L., Dunn-Geier, J., & Belair, J. (2005). Multi-site Study of the Childhood Autism Rating Scale (CARS) in Five Clinical Groups of Young Children. *Journal of Autism and Developmental Disorders, 35*, 625–634.

Perry, A., Cummings, A., Dunn Geier, J., Freeman, N. L., Hughes, S., LaRose, L., et al. (2008). Effectiveness of Intensive Behavioral Intervention in a large, community-based program. *Research in Autism Spectrum Disorders, 2,* 621–642.

Region 6 Autism Connection. (2006). *Early intensive behavioral treatment program procedures and guidelines.* [Electronic version] Retrieved December, 9, 2008, from http://vmrc.net

Reichow, B., & Wolery, M. (2009). Comprehensive synthesis of early intensive behavioral interventions for young children with autism based on the UCLA Young Autism Project model. *Journal of Autism and Developmental Disorders, 39*, 23–41.

Schopler, E. C., Reichler, R., & Renner, B. (1988). *The Childhood Autism Rating Scale (CARS).* Los Angeles, CA: Western Psychological Services.

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

Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland-II: Vineland Adaptive Behavior Scales, Second Edition*. Minneapolis, MN: Pearson.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence-III.* San Antonio, TX: Psychological Corporation.