WHAT DATA SUPPORT USING TPBA?

Research on the original TPBA

Numerous studies have looked at the advantages and disadvantages of the TPBA model. In the following discussion, studies that have been conducted on TPBA to determine reliability and validity are examined. Reliability, as defined by Gall, Borg, and Gall (2006), refers to the level of internal consistency or stability of a measuring device over time. Reliability of the TPBA process across time and raters has been studied (Al-Balhan, 1998; Cornett & Farmer-Dougan, 1998; Friedli, 1994; Linder, Green, & Friedli, 1996; Myers, McBride, & Peterson, 1996) and was well supported in most test–retest and interrater conditions.

Interrater Reliability

Interrater reliability generally refers to the degree of agreement between two or more raters on the point value of responses to specific test items. Interrater reliability was studied by Myers and colleagues (1996), who noted higher percentages of interrater agreement when using TPBA as compared with standardized assessments. In their study, they compared each rating on every developmental profile item with ratings made by other team members and parents on the same item. Higher percentages of agreement were found using transdisciplinary play-based assessments as compared with standardized assessments. For TPBA, 11 of 15 developmental domains had higher percentages of exact agreement, and 12 domains had higher percentages of agreement within one point. For TPBA, there was a mean agreement of 51.8% for staff and parents' developmental ratings with a range of 28.8%–84.4%. At the same time, the standardized assessment had a mean agreement of 46.5% for staff and parents' developmental ratings, with a range of 38.4%–77.3%.

In a second measure of interrater reliability, staff members' mean ratings were correlated with parents' ratings. Here, higher correlation coefficients were found between parents and professionals for transdisciplinary play-based assessments ($r = 0.70$, $p = 0.001$) than with standardized assessments ($r = 0.67$, $p = 0.001$) when examining developmental profile ratings. A Fisher's z-test indicated a slightly higher, although nonsignificant, difference in correlation between parents and professionals for transdisciplinary play-based assessments.

The interrater reliability of TPBA was also examined by Friedli (1994) using videotapes of TPBA sessions, with independent raters assessing each child. She revealed that the reliability for the language, motor, and combined domains met the same criteria for standardized tests used to make eligibility decisions. She also found that professionals from various disciplines were as accurate in rating young children's developmental competence in other domains as they were in their domains of expertise when they were given specific guidelines for observation.

Interrater reliability was also examined by Al-Balhan (1998). In his study of training and implementation of TPBA in Kuwait's kindergartens, he compared the TPBA results of newly trained kindergarten teams with the results of specialists in each developmental domain who were also newly trained in TPBA. He found that the mean overall percentage of agreements between kindergarten team members and specialists in the field ranged from 68% to 82%. This is especially interesting when one considers the linguistic and cultural adaptations that were necessary to complete the training and implementation within the country of Kuwait. Although some additional adaptations may be necessary, it appears to demonstrate a reasonable level of agreement in most domains for newly trained kindergarten teams and specialists in Kuwait.

Test–Retest Reliability

Test–retest reliability refers to overall stability of a measure across a specified time interval. In 1994, Friedli studied the stability of TPBA assessments for 10 children between the ages of 3 and 6 years. Each child studied was assessed twice using TPBA within a 6-week period of time. Her results showed that chi-square tests of association between the first and second test was significant at $p < 0.0001$ for all domains.

Validity

Validity can take several forms including content validity, criterion-related validity, and construct validity. Validity studies on TPBA have shown strong support for validity across all of these dimensions (Al-Balhan, 1998; Friedli, 1994; Karr, 1998; Linder & Green, 1995; Myers et al., 1996).

Content Validity

Gall et al. (2006) describe content validity as the degree to which the sample of test items represents the content that the test is designed to measure. In a study of content validity, Friedli (1994) found that the TPBA guidelines were supported by early childhood professionals most likely to use the process, such as psychologists, educators, speech-language therapists, and motor specialists. These experts rated the developmental domains and subcategories for relevance, clarity, and comprehensiveness using a Likert scale ranging from 1 to 7. All of the subcategories of the TPBA guidelines were judged favorably (higher than 4), with most ratings between 6 and 7. Linder and Green (1995) surveyed 40 professionals across the country ($N = 10$ for each domain) concerning the content validity of each domain, using a Likert scale of 1–7. All subcategories for all domains were rated between 6 and 7 on the scale, indicating strong support for content validity.

Myers et al. (1996) also examined staff perceptions of the amount of information obtained from assessments as well as the usefulness of that information. Following each evaluation, each staff member rated the amount of information obtained for the primary developmental domains using a Likert scale (1 = none, 2 = limited, 3 = fair, 4 = moderate, 5 = great). They found that significantly more information was obtained through the use of TPBA in the domains of communication, social, and motor skills. Staff reported that equivalent amounts of information were obtained within the cognitive, sensory, and self-help domains.

Criterion-Related/Concurrent Validity

Gall et al. (2006) state that criterion-related or concurrent validity is determined by relating the test scores of a group of subjects to a criterion measure administered at the same point in time, or within a short interval of time. The concurrent validity of TPBA was measured by comparing the outcomes of play-based assessment to traditional standardized and norm-referenced tests for children with and without disabilities. Friedli (1994) found that TPBA was as accurate as standardized measures for determining whether a child was eligible for services as compared with the results of the Battelle Developmental Inventory (BDI-2; Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984). She also found that the two assessments produced similar profiles regarding the examinee's strengths and needs. In addition, in Friedli's study, TPBA was actually more accurate in identifying one child with social-emotional concerns.

Karr (1998) used a sample of typically developing children and compared results from the administration of the BSID-II and TPBA. After converting data to standard

scores for a comparison, significant correlations between the two approaches were found. A study by Kelly-Vance, Needelman, Troia, and Ryalls (1999) with at-risk 2-year-olds modified the TPBA and compared only the results from the TPBA cognitive section with the Bayley Scales of Infant Development-II (BSID-II; Bayley, 1993). Comparison of age equivalents (AE) by averaging the subdomains for the cognitive domain on TPBA with the AE on the mental scale of the BSID-II, revealed higher cognitive AE on the TPBA than on the BSID-II. It appears that children show their optimal performance on the TPBA. It should be noted that this study did not include other domains of development, and thus was not transdisciplinary, which somewhat defeats the purpose of the process, as each of other domains is related to cognitive development. In addition, none of the previously mentioned studies took into consideration the qualitative observations from the guidelines. Failing to do so also ignores a major purpose of the TPBA, which is to look at *how* the child performs, behaves, and learns, not just what age level is represented by the skills observed. These transdisciplinary and qualitative elements are essential both for determining eligibility, but also for planning intervention.

Social Validity

Social validity refers to the ecological integrity of assessment information, the acceptability of the assessment methods, and the importance of the results to families and professionals (Neisworth, 1990). Myers and colleagues (1996) compared a multidisciplinary standardized model with the transdisciplinary play-based approach through random assignment of 40 children under the age of 3 referred to one of the models. They then examined consumer feedback (parent and professional) on the approaches, time spent on evaluations, and evaluation of written discipline reports.

Through questionnaires with 17 statements describing positive aspects of the assessment and report process, parents were asked to rate their degree of agreement or disagreement with each of the statements using a 5-point Likert scale. Myers et al. (1996) found that the means for 13 of the 17 items, along with the overall total for TPBA, were higher, although not statistically significantly higher, than those for standardized assessments. Specifically, parents appeared to feel more comfortable seeking information from professionals during TPBA, and they also perceived the goals identified as a result of the TPBA as important. Myers and colleagues also found that speech-language pathologists and school psychologists rated TPBA as significantly more useful than standardized assessments for identifying a child's strengths and weaknesses and for developing program planning.

Functional utility, which refers to the clarity, completeness, and usefulness of information, was also studied by Myers et al. (1996). Myers and colleagues examined the functional utility of TPBA assessment reports. Compared with reports written after traditional testing, the reports resulting from TPBA rated higher in their study, particularly for ease of obtaining an overview of the child's abilities, ease of determining which developmental areas were of concern, the number of developmental areas discussed in the report, the report being written in jargon-free language, the integration of discipline-specific information, and the objectives being clearly based on the child's strengths and weaknesses. When all of the items were combined, the mean scores for TPBA assessment reports were significantly higher than the mean for the standardized assessment reports. In another study, Cornett and Farmer-Dougan (1998) tried two different approaches for scoring TPBA and found that more objective scoring procedures were preferred.

In another study, Al-Balhan (1998) examined social validity through staff perceptions of the usefulness of TPBA in identifying strengths and weaknesses as well as generating programming or intervention ideas following a training and implementation

program in Kuwait. In his analysis, 90% of trained professionals found TPBA to be useful for identifying strengths and needs and for generating programming or intervention ideas. His findings have important implications for use of TPBA in multicultural settings.

Another way of looking at social validity is what professionals choose to use when standardized testing fail to meet their needs. A survey of school psychologists by Bagnato and Neisworth (1994) found that when traditional tests found the child to be "untestable," the most frequently used alternative assessment approaches included parent interviews (58%), play-based assessments with toys and people (44%), and parent–child observations in natural settings (30%). These finding demonstrate that professionals look to alternative measures for more valid assessment results. The revised TPBA process incorporates a multidimensional approach that includes these most frequently used alternatives. A survey of professionals in the state of Colorado (University of Colorado, 2003) revealed that 97% of Part C and 82% of Part B assessments to determine eligibility for services involved play-based assessment as a major piece of the process. This was compared with 21% using play-based assessment in 1999 for Part C and 18% using play-based assessment for Part B. More than 50% of both Part C and Part B used a transdisciplinary approach. This increase in both play-based assessment and the transdisciplinary approach is indicative of the trend in the field.

Research on TPBA2

Construct and content validity of the revised *Transdisciplinary Play-Based Assessment* (TPBA2) was analyzed in a study with national and international experts (Linder, Goldberg, & Goldberg, 2007). The first aspect of the study was to look at the validity of the transdisciplinary concept. Twelve experts, three for each domain, were provided a survey related to all of the content. Each expert was asked to rate how frequently a problem in one subcategory directly contributed to problems in other subcategories. These questions were posed in the format "How often do problems in (Subcategory A) directly contribute to problems in (Subcategory B)?" with the answer choices of *never, rarely, sometimes, frequently,* and *always.* These frequencies were later converted into numerical scores using the experts' rating of the frequencies on a scale of 0 (never) to 10 (always). An algorithm was then used to look at the relationships within and across domains. Across the 28 subcategories, the total number of possible bidirectional transdisciplinary influences is 588. The number of transdisciplinary influences that exceed the frequency value of 5 was 185 (31.4%). Furthermore, the number of cases for which the experts rated no influence was only 61 cases, or 10.4% of the total possible influences, meaning that in nearly 90% of the cases, the experts indicted that some degree of transdisciplinary influence was present.

Findings revealed that strong interdependent relationships exist both within and across developmental domains, thus supporting the construct of transdisciplinary development and the need for an integrated approach to assessment and intervention (see Table 1.1). Because some of the cross-disciplinary influences were stronger than others, future studies will examine the potential of developing "influence maps" based on these interrelationships to guide the intervention process.

The 12 experts in this study were also asked about the content of each domain and the definitions of content for each subcategory for their area of expertise. Each was asked to comment on the breadth of the subcategories, the clarity of the definitions, and the completeness with which the subcategories covered their associated developmental domains. The results of the survey indicated that the subcategories were conceptually solid, and changes in the subcategories were made according to the offered feedback.

Table 1.1. Consolidated influences between developmental domains

	Influenced				
Influencer	Cognitive	Sensorimotor	Language and communication	Emotional and social	Mean influence
Cognitive	Mean 6.53 SD 1.78	Mean all 1.68 Mean Cog 2.08 Mean SM 1.29 SD all 1.51 SD Cog 1.39 SD SM 1.35	Mean all 3.03 Mean Cog 3.45 Mean Comm 2.61 SD all 1.69 SD Cog 1.48 SD Comm 1.41	Mean all 4.03 Mean Cog 4.86 Mean ES 3.21 SD all 2.23 SD Cog 2.24 SD ES 1.63	3.82
Sensorimotor	Mean all 3.79 Mean SM 2.86 Mean Cog 4.73 SD all 2.10 SD SM 1.77 SD Cog 1.79	Mean 3.75 SD 2.02	Mean all 2.09 Mean SM 2.00 Mean Comm 2.17 SD all 1.86 SD SM 1.43 SD Comm 1.85	Mean all 3.31 Mean SM 3.01 Mean ES 3.61 SD all 2.05 SD SM 1.85 SD ES 1.77	3.23
Communication	Mean all 3.69 Mean Comm 3.16 Mean Cog 4.22 SD all 1.81 SD Comm 1.41 SD Cog 1.88	Mean all 0.44 Mean Comm 0.55 Mean SM 0.32 SD all 0.64 SD Comm 0.65 SD SM 0.61	Mean 3.13 SD 1.68	Mean all 3.74 Mean Comm 3.18 Mean ES 4.29 SD all 2.17 SD Comm 1.99 SD ES 1.92	2.75
Emotional and social	Mean all 4.64 Mean ES 3.57 Mean Cog 5.71 SD all 2.15 SD ES 1.69 SD Cog 1.60	Mean all 1.59 Mean ES 1.63 Mean SM 1.55 SD all 1.55 SD ES 1.25 SD SM 1.51	Mean all 2.52 Mean ES 2.41 Mean Comm 2.63 SD all 1.77 SD ES 1.64 SD Comm 1.46	Mean 5.34 SD 2.17	3.52
Mean influence	4.66	1.86	2.69	4.10	

Key: SD = standard deviation; Cog = Cognitive domain; SM = Sensorimotor domain; Comm = Language and communication domain; ES = Emotional and social domain

Interrater Reliability

Several interrater reliability studies were conducted with TPBA2 (Linder, 2005) to determine interrater agreement with regard to individual domains. These studies included participants from numerous disciplines involved in 1) a 2-day initial training on TPBA2; 2) an initial 2-day training, time to practice, and then a 2-day follow-up training; 3) graduate students in Child, Family, and School Psychology in a 10-week course on TPBA2; and 4) professionals who have used TPBA2 during the past several years. Several of the studies involved watching videotapes of four different children. Cuts from four tapes were shown of children in TPBA sessions, and professionals were asked to rate each tape in relation to one area of development. Observers were given no background information about the child other than his or her age. Children in the tapes ranged from having typical development to moderate delays. Professionals and students rated all domains, regardless of their professional discipline, using the TPBA2 Observation Guidelines and Age Tables to make a judgment of whether the child 1) was above average, 2) was typically developing, 3) should be monitored, or 4) had a concern that would make the child eligible for services. Level of agreement for each child was then obtained for each sample group. All disciplines rated each child observed to determine whether transdisciplinary observations were effective. In addition, after they had rated the children individually, 10 teams also rated the tapes and determined a team rating.

Table 1.2. Reliability of observations of TPBA with video recordings

Level of training	Child 1[a] Sensorimotor Proportion agreement (mild to moderate)	Child 2[b] Communication Proportion agreement (moderate)	Child 3[c] Emotional/social Proportion agreement (mild)	Child 4[d] Cognitive Proportion agreement (typical-at-risk)	All areas
2-day training (professionals), State A	*N = 9 0.88	N = 10 0.90	N = 11 1.00	N = 10 0.80	0.89
2-day training (professionals), State B	*N = 8 0.75	N = 8 1.00	N = 8 0.75	N = 8 0.875	0.843
2-day follow-up training (professionals)	N = 23 0.95	N = 23 0.95	N = 23 0.95	N = 23 0.95	0.95
20-hour training (students)	N = 9 1.00	N = 9 1.00	N = 9 1.00	N = 9 1.00	1.00
Experts	N = 4 1.00	N = 4 1.00	N = 4 1.00	N = 4 1.00	1.00
Teams	N = 10 1.00	N = 10 1.00	N = 10 0.90	N = 10 1.00	0.975

[a]Child 1: Previously assessed and determined to have mild to moderate concerns with muscle tone, motor planning, and sensory modulation.

[b]Child 2: Previously assessed and determined to have moderate language delays and communication disorders.

[c]Child 3: Previously assessed and determined to have attention deficit disorder and regulatory concerns, with mild concerns related to language and social interaction.

[d]Child 4: Previously assessed and determined to be at-risk due to environmental factors, but cognitively within the typical range of development.

*N in each category may vary if people left the training early or forms were incomplete.

Table 1.2 shows the findings across the four domains for the individuals and teams using videos to rate children. It should be noted that only segments of each full evaluation were shown. The videos also do not always convey nuances that are picked up in actual in vivo observations. Given these limitations, the results demonstrate strong reliability.

It should be noted that the level of agreement was greatest for children who were typically developing or had moderate delays. When concerns were mild, observers were more cautious and ratings were spread across Watch and Concern. In the case of the child at risk due to environmental factors, the ratings varied slightly across Typical and Watch, revealing that some subtle issues can be seen in TPBA. The observers tended to use the category "watch" when they were unsure. The value of the team discussion can be seen clearly, as participants were able to discuss their observations and compare their opinions. Teams, versus individuals, agreed on all children with 90%–100% agreement. It appears that, particularly for children who are in the "gray" area, a team discussion can be particularly beneficial in determining whether a concern is of sufficient severity to warrant services.

Another interesting finding is the impact of training and practice. As can be seen from Table 1.3, in almost all instances having time to practice and work with TPBA increased reliability. This may also relate to the transdisciplinary learning and teaching that occurs over time. As teams work together they become more competent in assessing the whole child. The experts, who came from a team that had been working together with TPBA for several years, had 100% agreement, which also may demonstrate that teams working together over time come to understand each others' perspective.

Table 1.3. Impact of training on reliability

Level of training	*N*	Child 5* Sensorimotor % agreement	Child 5* Communication % agreement	Child 5* Emotional/Social % agreement	Child 5* Cognitive % agreement	All areas
20-hour training with live observation	10	100	100	90	100	97.5%
Experts with live child	4	100	100	100	100	100%

*Child 5 was previously assessed and determined to have mild to moderate delays in all areas.

In another study an expert TPBA team evaluated a child and each team member, plus 10 student observers rated the child across all four domains. In this instance, the child had mild to moderate delays in all areas. The only disagreement was in the area of emotional development, which was a relative strength for the child.

Validity

Conducting concurrent validity with TPBA2 and other instruments is difficult, because there are characteristics of TPBA (i.e., informed clinical opinion, discussion with team members about cross-disciplinary influences, inclusion of parental perceptions) that are unique and even could be said to be at odds with traditional testing. This can be seen in the findings of a study on the concurrent validity and social validity of TPBA as compared with the BDI-2 (DeBruin, 2005). Both types of validity were addressed simultaneously, because the social validity could address whether parents and professionals felt the assessment was comprehensive and revealed an accurate picture of the child.

DeBruin (2005) conducted a study to determine whether TPBA demonstrated a significant association between categories of eligibility and noneligibility for Part B services compared with the BDI-2. The population studied consisted of 45 children from 2 to 5 years of age recruited from a large school district in Texas who were being assessed as they made the transition from the Part C program to the district preschool program or who were referred to the district due to developmental concerns by caregivers, doctors, and early childhood education programs. Forty-five children were assessed both on the TPBA and the BDI-2. Per modified random assignment, half of the children received TPBA first and the other half received the BDI-2 first.

Analysis of whether there was a statistically significant association between TPBA and the BDI-2 on categories of eligibility and noneligibility for Part B services was done by performing a test of association in category membership using a chi-square and a ϕ coefficient. Team members conducting both assessments rated the child as "eligible for services" or "not eligible for services" using the procedures defined for each instrument. To be eligible, in accordance with state policies, children needed to demonstrate the equivalent of a 25% or more developmental delay in one or more areas. To make the processes more comparable, the results were looked at if 1) informed clinical opinion was omitted (as it was not considered as part of the BDI-2 process), 2) the Watch category on TPBA was omitted (as it is not a consideration on the BDI-2), 3) and the developmental history was omitted (not a consideration in scoring the BDI-2). When these modifications to the TPBA process were done, significant agreement of 82.2% ($p < .001$) was found. However, because this is not the way TPBA is meant to be used, results also were examined when the other processes were included. When informed clinical opinion, the Watch category, and developmental history were included, agree-

ment dropped to 57.8% with the Battelle cutoff of 70 and 62.2% when the cutoff was 75. However, when the data from the social validity ratings were compared for children when there was disagreement, in most cases both the psychologist and parents felt the BDI-2 was not accurately addressing the child's strengths and needs.

These findings demonstrate that there is a strong correspondence between TPBA and the BDI-2 when absolute cutoffs are used, but there is a weak correspondence when professional opinion, a Watch category, and developmental risk factors are taken into consideration on TPBA2. When looked at through the traditional lens of how concurrent validity is measured, the addition of the social validity measure reveals that TPBA may, in fact, be more accurate in the eyes of professionals and parents, as described below in the second part of DeBruin's study.

DeBruin (2005) also investigated whether TPBA and the BDI-2 demonstrated social validity in assessing children eligible for Part B by examining primary caregivers' as well as assessment team members' perceptions of the process as measured by questionnaires completed following each assessment. Each questionnaire asked the primary caregiver to use a 5-point Likert scale, with "1" meaning they strongly disagreed, and "5" meaning they strongly agreed, to rate their perceptions about their experiences during their child's assessment. Ratings of 4 or above were considered to be a demonstration of high social validity. Individual questions were also evaluated to determine the percentage of caregivers who scored each question at 4 or above. Mean scores and standard deviations for each question for both TPBA and the BDI-2 were analyzed and compared. A paired *t*-test also was performed to evaluate the significance of the differences between the means for each question (see Table 1.4).

Paired *t*-tests showed that TPBA was rated significantly higher than the BDI-2 on all of the questions as perceived by the caregivers. The results of this analysis revealed that there was a significant difference between parents' perceptions of TPBA2 and the BDI-2 on questions relating to 1) the comfort of the child ($p < 0.002$), 2) allowing the child to demonstrate his or her highest level of ability ($p < 0.000$), 3) accurately showing the child's needs and concerns ($p < 0.001$), 4) demonstrating skills and behaviors seen at home ($p < .037$), 5) parents' comfort with the assessment approach ($p < 0.004$), 6) understanding what skills were being evaluated ($p < 0.002$), and 7) their feeling like a valued member of the team ($p < 0.001$).

Table 1.4. Parents' perceptions of TPBA and BDI-2

Statements from caregiver questionnaires	Tool	*N*	Mean	SD	% of scores at 4 or above	*p* value
My child appeared comfortable during the assessment process.	TPBA2	44	4.86	0.35	100.0%	0.002**
	BDI-2	44	4.50	0.73	90.9%	
The assessment allowed my child to demonstrate his/her highest level of ability.	TPBA2	45	4.60	0.54	97.8%	0.000***
	BDI-2	45	3.82	1.13	66.7%	
The assessment accurately showed areas in which my child has needs or concerns.	TPBA2	45	4.40	0.94	88.9%	0.009**
	BDI-2	45	3.87	1.10	62.2%	
My child was able to demonstrate skills and/or behaviors I typically see at home or in the community.	TPBA2	45	4.31	0.87	86.7%	0.037*
	BDI-2	45	3.96	1.07	71.1%	
I felt comfortable with the approach that was used to assess my child.	TPBA2	45	4.84	0.37	100.0%	0.004**
	BDI-2	45	4.44	0.87	84.4%	
I had a good understanding of what skills were being evaluated during the assessment process.	TPBA2	43	4.72	0.45	100.0%	0.002**
	BDI-2	43	4.30	0.86	84.5%	
I felt like a valued member of the assessment team.	TPBA2	42	4.74	0.54	95.3%	0.001**
	BDI-2	42	4.31	0.81	84.5%	

Note: Responses were based on a 5-point Likert scale with 1 = parents strongly disagree, and 5 = parents strongly agree.
Key: *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

DeBruin's study (2005) also looked at the social validity of TPBA and the BDI-2 as perceived by team members. Following both the TPBA and the BDI-2 assessments, team members involved in each assessment were asked to complete a questionnaire using a 5-point Likert scale to rate their perceptions about the social validity aspects of the assessment. Items rated included the assessment's 1) ability to make the child comfortable, 2) accuracy in measuring the child's strengths, 3) accuracy in measuring the child's needs, 4) ability to measure highest level skills, 5) usefulness in providing information for program planning, and 6) ability to get an integrated holistic view of the child. Mean scores and standard deviations were again computed, along with *t*-tests to look for significant differences. As with the parent questionnaire, ratings of 4 or above were considered to be a demonstration of high social validity. An analysis of team member responses for TPBA assessments showed that all of the questions had averages of 4 or above except for one TPBA team's response to the ability of the assessment to provide useful information for program planning, which had a mean rating of 3.98.

An analysis of the psychologist's responses for the BDI-2 assessments showed that all six questions had averages below 4. The psychologist's ratings ranged from the highest mean rating of 3.29 on the perception of the child's comfort, to the lowest mean rating of 2.13 on the assessment's ability to accurately measure the child's needs.

Overall, the social validity results reveal that TPBA is socially valid as perceived by both parents and professionals. When DeBruin's concurrent validity results are combined with the social validity results, it becomes clear that TPBA is perceived by parents and professionals to be more accurate, holistic, and useful. Concurrent validity results demonstrated that although absolute cutoff scores provided a significant level of agreement, a more accurate picture of the child may require inclusion of background and history and professional judgment. As a result of this study, the Observation Summary forms were changed to enable professional opinion to be considered separately from the level of delay.