Running head: Play-based Assessments of Motor and Cognitive Skills for Infants and Young Children

Research Report

Reliability and Validity of Play-based Assessments of Motor and Cognitive Skills for Infants and Young Children: A Systematic Review

Michael G. O’Grady, Stacey C. Dusing

M.G. O’Grady, PT, MS, Rehabilitation and Movement Science Program, Virginia Commonwealth University, Richmond, Virginia.

S.C. Dusing, PT, PhD, PCS, Motor Development Laboratory, Department of Physical Therapy, Virginia Commonwealth University, and Department of Pediatrics, Children’s Hospital of Richmond, Virginia Commonwealth University, PO Box 980224, Richmond, VA 23298-0224 (USA). Address all correspondence to Dr Dusing at: scdusing@vcu.edu.


© 2014 American Physical Therapy Association

Published Ahead of Print: xxxx

Accepted: August 14, 2014
Abstract

**Background:** Play is vital for development. Infants and children learn through play. Traditional standardized developmental tests measure if a child performs individual skills within controlled environments. Play-based assessments can measure skill performance during natural, child-driven play.

**Purpose:** The purpose of this study was to systematically review reliability, validity, and responsiveness of all play-based assessments which quantify motor and/or cognitive skills in children birth to 36 months of age.

**Data Sources:** Studies were identified from a literature search using PubMed, ERIC, CINAHL, and PsycInfo databases, and the reference lists of included papers.

**Study Selection:** Included studies investigated reliability, validity, or responsiveness of play-based assessments that measured motor and/or cognitive skills for children to 36 months.

**Data Extraction:** Two reviewers independently screened 40 studies for eligibility and inclusion. Reviewers independently extracted reliability, validity, and responsiveness data. They examined measurement properties and methodological quality of included studies.

**Data Synthesis:** Four current play-based assessment tools were identified in 8 included studies. Each measured motor and/or cognitive skills in a different way during play. Inter-rater reliability correlations ranged from 0.86-0.98 for motor development and 0.23-0.90 for cognitive development. Test-retest reliability correlations ranged from 0.88-0.95 for motor and 0.45-0.91 for cognitive. Structural validity correlations ranged from 0.62-0.90 for motor and 0.42-0.93 for cognitive. One study assessed responsiveness to change in motor development.
Limitations: Most studies had small and poorly-described samples. Lack of transparency in data management and statistical analysis was common.

Conclusions: Play-based assessments have potential to be reliable and valid tools to assess cognitive and motor skills, but higher quality research is needed. Psychometric properties should be considered for each play-based assessment before it is used in clinical and research practice.
Play provides infants and young children with the ability to practice skills and support all domains of
development: motor, cognitive, social-emotional, communication, and adaptive.\textsuperscript{1-4} Play has been
variably-defined in the literature given different disciplines and reasons for assessing development
through play.\textsuperscript{5} In this systematic review, play is defined as a pleasurable, active, self-motivated
developmental phenomenon\textsuperscript{1,6} by which infants and young children learn about the world through
interactions with objects and people.\textsuperscript{5,7}

Play fosters both motor and cognitive development.\textsuperscript{1,2,7} Play is common to all infants, and it is a primary
arena within which domain-specific and global aspects of development occur.\textsuperscript{1,8,9} Early play helps to
prepare infants and young children to learn in school.\textsuperscript{10} Children learn through the repetition of behaviors
during play within typical environments and routines.\textsuperscript{11}

Play is the basis for many developmental interventions used with children with disabilities.\textsuperscript{12} Play, however, is often not a part of traditional standardized developmental tests used by pediatric physical
therapists and other early intervention providers to determine the need for intervention or the efficacy of
intervention.\textsuperscript{13} Traditional standardized developmental assessments typically involve a child performing
a specific task within a controlled environment that is outside the context of everyday routines.\textsuperscript{13} Some
assessments require the examiner to elicit behaviors by altering the context or moving the child.\textsuperscript{14,15}
Behaviors assessed in this way are not authentic child-directed behaviors, and the child may not perform
optimally.\textsuperscript{16} Furthermore, traditional standardized developmental assessments are designed to determine
if a child can perform a specific skill, not if the child performs the skills in their normal routine.\textsuperscript{15}

Play-based assessments are standardized measures designed to quantify changes in one or more of the
five developmental domains during self-motivated, child-driven play.\textsuperscript{14,17,18} Some literature suggests that
play-based assessments may be effective and efficient means of assessing a child’s developmental level,\textsuperscript{19}
evaluating change over time, and evaluating the efficacy of intervention.\textsuperscript{18,20} Play-based assessments are
often adjuncts to other assessment procedures, though some argue that they can also serve as a basis for discriminative decisions and planning. In this review, play-based assessment is differentiated from an assessment of play, which interprets the type of play in which a child is engaged relative to a hierarchical developmental theory of play. Assessments of play are not discussed in this review.

Play-based assessments focus on child directed activities. During play-based assessment, the child directs the interaction and experience, increasing the likelihood of observing behaviors that the child typically performs. This results in a rich description of a child’s domain-specific strengths and weaknesses. Using the arena of play provides the practitioner not only the ability to assess current skills, but also the added benefit of previewing emerging skills in a functional context. Play-based assessments add authenticity and contextual benefits to the assessment of motor and cognitive development because they measure objective behaviors during child-driven activities within a normal environment. This allows examination of cross-domain relationships by integrating findings.

Play-based assessments can be contrasted with traditional standardized assessments. First, play-based assessment takes place within a naturalistic environment and context while traditional standardized developmental tests require specific responses to an examiner-provided stimulus. Second, play-based assessments typically quantify if and how often a child performs specific types of skills during a naturalistic observation, rather than just assessing if the child can perform the skill. Third, these assessments are child-driven rather than examiner-driven, giving the practitioner insight into the child’s ability to explore and learn. Fourth, play-based assessments can document limitations commonly seen in children with developmental delays such as decreased attention to toys, using fewer toys and less variety of active play skills, and being more passive during play.

While the theoretical value of play-based assessments is clear, the reliability and validity of play-based assessments need to be considered before they are used in clinical practice or research. The first aim of
this systematic review is to determine the inter-rater and test-retest reliability of play-based assessments of motor and/or cognitive skills for infants and children 0-36 months. The second aim is to identify the content and structural validity of play-based assessments of motor and/or cognitive skills for infants and children 0-36 months, as well as the responsiveness of these measures.

This paper focuses on the assessment of infants and toddlers, birth to 36 months who, based on their age, could be eligible in the United States for Early Intervention (EI) services under Part C of the Individuals with Disabilities Education Improvement Act (IDEIA).\textsuperscript{27} Play-based assessments allow for assessment in a variety of cultures, countries, and at a variety of ages. As in many countries, the goal of providing intervention in young children in the United States is to support early development and improve readiness to learn in children with or at-risk for developmental delays. Intervention programs with similar goals around the world may find play based assessments an option for assessing the needs and progress of children if these tools are reliable and valid.

The results of this study will provide information on the reliability, validity, and responsiveness of play-based assessments. This information may help to determine if play-based assessments can be used for research and clinical purposes. In addition, this information will help clinicians to determine which play-based assessments are best to supplement traditional standardized developmental tests that are currently used to evaluate the need for and efficacy of early developmental intervention services.
Methods

Search criteria were developed to identify studies which met inclusion and exclusion criteria specified prior to the study. Studies were required to evaluate one or more of the following measurement properties of a play-based assessment of motor and/or cognitive skills: inter-rater reliability, test-retest reliability, structural validity, content validity, and responsiveness to change over time. Subject ages were fully or partially 0 to 36 months. Subjects could have a diagnosed disability, delay, or be developing typically. Studies which did not include play-based assessment of motor or cognitive skills, children birth to 36 months, were not available in English, or were a review of previous research or theory without new data were excluded.

Data Sources and Searches

A literature search was performed using the PubMed interface from Medline (late 1940s - May 2013), ERIC (1966 - May 2013), CINAHL (1937 - May 2013), and PsycInfo (1894 - May 2013). Search terms were developed with the help of a research librarian using MeSH headings, key words, and phrases. Terms were purposefully broad to capture all publications which met the inclusion criteria for this systematic review. The full search strategy is described in the Appendix.

Study Selection

Consistent with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, results from the literature search were reviewed for duplicates prior to screening for inclusion. The title and abstract of all identified publications were screened using the inclusion and exclusion criteria. Any publication which clearly did not meet the exclusion criteria was moved to the screening process. During screening, two reviewers independently reviewed the full text publication to determine eligibility for the systematic review. Any question of inclusion between the two reviewers during eligibility was resolved through discussion. The bibliographies of included papers were also reviewed by both authors to determine if additional studies warranted inclusion.
Data Extraction and Quality Assessment

The inter-rater and test-retest reliability, content and structural validity, and responsiveness data for each included study were extracted independently by each reviewer using data collection forms developed for this systematic review. Any discrepancy in the extracted reliability or validity data was discussed between reviewers, and a consensus was reached. No statistical or meta-analysis was conducted given the limited number of included studies. A priori, a correlation ranging from 0.00 to 0.50 was considered weak, 0.50 to 0.75 was considered moderate, and 0.75 – 1.00 was considered to be strong. The strength of the correlation presented as a measure of reliability or validity was used to categorize the degree of reliability and validity documented for each play-based assessment and for the group of play-based assessments. Therefore, the terms weak, moderate, and strong reliability and validity are used to describe the results of each included paper.

The COConsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) was used as a measure of methodological quality of the measurement properties. The five measurement properties assessed for this study were defined by the COSMIN. Inter-rater reliability is a measure of whether different raters can score the same testing occasion and obtain the same score. Test-retest reliability is the extent to which the scores for patients who have not changed are the same for repeated measurement over time. Structural validity is the degree to which scores of a health-related instrument adequately reflect the same construct as a validated assessment. Content validity is a judgment about whether the content of a test adequately reflects the construct to be measured. Responsiveness is the ability of the measurement tool to measure change over time in the focal construct.

The COSMIN can be used to measure methodological quality with a 3-step process. First, the measurement properties assessed in the paper are identified. Second, reviewers score each measurement
property. Each measurement property on the COSMIN has a rating box containing 5-18 individual items specific to that measurement property. Each item within the rating box is scored based upon specific scoring criteria: from 1-4 possible answers representing excellent, good, fair, and poor quality for that item. An item is scored as excellent when there is adequate evidence provided for a given item. When information is not provided, but it is reasonable to assume information regarding that item, the item is rated as good. Fair indicates that methodological quality for the item is doubtful, while poor is scored when there is evidence that the methodological quality pertaining to a specific item is inadequate. For example, clear evidence of patient stability between test and retest in a study receives a score of excellent on that item in the rating box for the measurement property test-retest reliability. If it was unclear if patients were stable during the time between test and retest, however, that item is marked as fair. The third step to score each rating box of the COSMIN involves determining the overall rating of methodological quality of each measurement property. This is determined by the lowest score for all items in the rating box for that measurement property. For example, if all responses in the test-retest rating box are excellent except for one judged to be fair, then the quality of the test-retest reliability measurement of that paper is considered to be fair.

Each author of this systematic review independently rated the methodological quality of each included study using the COSMIN. Any discrepancies in scoring between the raters which resulted in different measurement properties being scored, a change in the overall rating of methodological quality of any measurement property, or a discrepancy of two or more ordinal levels for any single item within a measurement property rating box was discussed and a consensus was reached. The overall methodological quality for each measurement property included in a paper was recorded and is reported in this systematic review.

**Results**
The title and abstract of 2133 studies were screened for possible inclusion. Forty studies could not be excluded during screening and were reviewed in full text for eligibility. Eight of these studies matched the inclusion criteria and were included in the systematic review, while 32 studies were excluded (Figure).

Studies including four separate play-based assessments currently available for commercial use were included in this systematic review: Play in Early Childhood Evaluation System (PIECES), Transdisciplinary Play-Based Assessment, 2nd edition (TPBA-2), Assessment, Evaluation, and Programming System, 2nd edition (AEPS), and the Individual Growth and Development Indicators (IGDI). Related assessments or precursors to these play based assessments were also identified in the literature. The Play-Based Assessment (PBA)\(^\text{16}\) was the initial form of the PIECES.\(^\text{21}\) The Play-Based Assessment was also the cognitive portion of the Transdisciplinary Play-Based Assessment,\(^\text{14}\) which was never tested for reliability or validity with a young population. The Transdisciplinary Play-Based Assessment (TPBA) is the previous version of the TPBA-2.\(^\text{19}\) Psychometric properties of the Evaluation and Programming System for Infants and Young Children (EPS-I)\(^\text{33}\) were presented in Bailey and Bricker\(^\text{34}\) and Bricker et al.\(^\text{35}\) The EPS-I is the predecessor of the AEPS.\(^\text{36}\) Part of the AEPS was used for the experimental Assessment, Evaluation, and Programming System for Eligibility (AEPS:E) as reviewed herein.\(^\text{20}\) Two other play-based assessments, a general outcome measure of growth in movement for infants and toddlers\(^\text{21}\) and the Early Problem Solving Indicator (EPSI)\(^\text{26}\) met inclusion criteria. These last two assessments are predecessors to the movement and cognitive sections of the (IGDI):\(^\text{37}\) the Early Movement Indicator (EMI-IGDI) and the Early Problem Solving Indicator (EPSI-IGDI), respectively (Table 1).

**Inter-rater Reliability**

Inter-rater reliability was measured in five studies.\(^\text{20,21,26,34,35}\) The Pearson correlation coefficient for inter-rater reliability of motor assessments ranged from 0.86 - 0.98 (Table 2).\(^\text{20,21,34,35}\) Inter-rater reliability of
cognitive assessments ranged from 0.23 - 0.90 (Table 3). One study of cognition had inter-rater reliability coefficients for individual skills, but not for the aggregate of cognitive behaviors displayed. The methodological quality in each of these studies was rated fair except for Bailey and Bricker which was poor (Table 4). Main reasons for fair ratings included use of a Pearson correlation coefficient rather than an intra-class correlation coefficient (ICC) and missing items from the sample. Bailey and Bricker was rated poor based upon small sample size and major flaws in their study design including that items not observed by one or both observers were omitted from analysis. It is possible that one observer missed several test items that actually occurred. This would have reduced the variability of the sample and artificially inflated the correlation coefficient.

**Test-retest reliability**

Test-retest reliability was assessed in five studies. Test-retest reliability of motor assessments had a range of Pearson correlation coefficients from 0.88 - 0.95 (Table 2), while the cognitive test-retest reliability Pearson correlation range was 0.45-0.91 (Table 3).

Three of the studies had fair methodological quality ratings while two were rated poor (Table 4). Areas in which these studies received low ratings included using a Pearson correlation coefficient rather than an ICC and small sample size and important flaws in the study design including different observers for the first and second test observations, different situations between each observation, and items not scored by either observer being omitted from the test.

**Structural Validity**

Structural validity was assessed in seven studies by comparing the scores on a play-based assessment to scores on traditional standardized developmental tests. Three of the studies had both motor and cognitive components (Tables 2 and 3). Two studies assessed solely motor skills during play (Table 2), while the other two were solely cognitive assessments (Table 3). The Pearson’s
correlation between the play-based assessments of motor skills and validated traditional standardized developmental tests of motor skills ranged from 0.62 - 0.90. Correlation between the play-based assessments of cognitive skills and validated traditional standardized developmental tests ranged from 0.42 to 0.93. Several different validated traditional standardized developmental tests of motor and cognition were used as comparisons (Tables 2 and 3).

Methodological quality for structural validity was fair to poor (Table 4). Four studies earned fair ratings. All of these studies had missing items, and the methods employed for handling missing items were unclear.16,21,26,35 Two of these also had methodological flaws.16,35 In one of these studies,16 the age equivalents on the play-based assessment were converted into standard scores for comparison with the standard scores on the Bayley Scales of Infant Development, second edition (Bayley-2).38 This comparison has not been validated as statistically sound. The other study35 was validated with two traditional standardized developmental tests, the Gesell and the Bayley Scales of Infant Development (Bayley).39,40 The disability levels and ages of each of the samples was unclear. The sample compared to the Bayley was also small. The three other studies that assessed structural validity were marked as poor due to a small sample size.19,20,34

Content Validity
Content validity of motor skills was assessed using the EMI-IGDI.21 It was found that there was a significant increase in the frequency of movements between each of three age cohorts (3-12 months, 13-24 months, and 25-36 months) during the 45-week study. Methodological quality was good in this study based upon a moderate sample size (Table 4).21

Responsiveness
One study21 measured responsiveness of a play-based assessment indirectly. This study assessed only motor skills in a sample of mostly typically-developing children 3-36 months old twice on a play-based assessment, the EMI-IGDI, and the Peabody Developmental Motor Scales -2 (PDMS-2). The PDMS-2
Locomotor and stationary subtests were responsive to change, with a statistically significant increase in raw score between two time points that were 45 weeks apart. A similar comparison was not made for the data from the EMI-IGDI. The movement rate on the EMI-IGDI, however, was correlated with the PDMS-2 locomotion scale at each of the two time points. Pearson’s correlation coefficient was 0.77 at time 1 and 0.90 at a time point 45 weeks later. Methodological quality was poor due to a small sample size (Table 4).

Discussion
The results of this systematic review indicate that inter-rater and test-retest reliability of play-based assessments range from $r = 0.23 – 0.98$, and structural validity of play-based assessments ranges from 0.42 – 0.93. As a group, both reliability and validity of play based assessments are inconsistent. The methodological quality of measurement properties among the studies contained in this systematic review is generally poor to fair, with only one study having a good quality rating. With only one or two studies of reliability or validity on each play-based assessment tool, and the poor to fair methodological quality of the studies, it was difficult to draw conclusions about any individual assessment or the group of play-based assessments as a whole. Therefore, reliability and validity for each play-based assessment need to be considered carefully before research or clinical application.

Inter-rater reliability of play-based assessments for both motor and cognitive skills was generally strong. One of 8 studies had a weak inter-rater reliability correlation, but the majority was $r \geq 0.86$. These inter-rater reliability findings indicate that the definitions of terms and scoring used in the assessments are clear to raters. Of the tests reviewed for both motor and cognitive skills, inter-rater reliability findings are highest for the AEPS:E$^{20}$ and its predecessor, the EPS-I.$^{34,35}$ The assessment with the best inter-rater reliability for motor skills is the EMI-IGDI.$^{21}$ Traditional standardized developmental tests of motor and cognitive development, such as the Peabody Developmental Motor Scales-2 (PDMS-2)$^{42}$ and the Bayley-2$^{38}$ have similar strong inter-rater reliability.$^{43}$
Test-retest reliability scores vary by construct measured. Test-retest reliability was strong for all three studies of motor development. Two of the studies used the same test (EPS-I) with different age groups. The EPS-I measured motor tasks using a criterion-referenced, curriculum-based assessment with a modified developmental checklist. The fact that environments, objects, and checklist questions were controlled may have improved reliability. The other play-based assessment of motor skills, the EMI-IGDI, measured skills in a longitudinal fashion with 3 to 8 measures per child, each measure separated by at least three weeks. The researchers used a split-half reliability method in which they averaged the odd trials and the even trials before comparing the average of the odd and even trials. While this is an acceptable way of measuring test-retest reliability, it reduced variability which otherwise may have resulted in lower test-retest reliability. Age of subjects did not affect the reliability of either the EPS-I or the EMI-IGDI. The Bayley Scales of Infant and Toddler Development, 3rd edition (Bayley-3) and the PDMS-2 have similar strong test-retest reliability for motor skills.

Test-retest reliability for cognition were lower than motor. This was evident with two tests, the PIECES and the EPS-I for children with and without disabilities. The reliability of the PIECES was assessed using a strict test-retest method (1-3 weeks between assessments) of the child’s most advanced cognitive skill level used during play. Test-retest reliability correlations of the PIECES were similar to those of the initial study with the EPS-I. The test-retest reliability of the EPS-I, however, varied substantially between each of the two studies which evaluated this measure. Two different methodologies were used to assess test-retest reliability, which might account for this discrepancy. Bailey and Bricker used different observers and different situations during test-retest which expanded the opportunity for error. In the study during which test-retest reliability was stronger, several items were not tested due to issues of privacy during videotaping (self-care, dressing), and several items from the gross motor scale were omitted due to constraints of videotaping. This creates a smaller pool of scored behaviors which may not represent the reliability of the test as a whole. In light of these contrasting findings, we suggest that
the complete EPS-I test-retest reliability cannot be identified. The EPS-IGDI\textsuperscript{26} had strong test-retest reliability using the split-half reliability method. Given the moderate or unclear results of the other play-based assessments of cognition, the EPS-IGDI has the best test-retest reliability, though it must be noted that single session test-retest studies of the EPS-IGDI (no split-half reliability) may show lower reliability due to changes in play behaviors that commonly occur from one session to the next.

Structural validity of play-based assessments that measure motor skills ranged from moderate to strong when compared to traditional standardized developmental tests. The lowest structural validity was the AEPS:E\textsuperscript{20} compared with the Gesell gross motor portion.\textsuperscript{39} All other play-based assessments had strong structural validity (correlations greater than 0.76). The EPS-I, an earlier version of the AEPS:E, had a strong correlation with the Gesell gross motor test.\textsuperscript{34} EPS-I items are arranged in a hierarchical-developmental progression, similar to the neuromaturational construct which the Gesell tests. The AEPS:E\textsuperscript{20} uses different standardization procedures, activities, and materials appropriate for toddlers which are not as strongly aligned with the hierarchical model of development. While this makes the AEPS:E more relevant to current theoretical approaches, it reduces the relationship with the Gesell. Comparison with a different traditional standardized developmental test may have increased the structural validity. The TPBA-2 motor section had a strong correlation when compared to the Bayley-3,\textsuperscript{19} and the EMI-IGDI likewise had a strong correlation when compared to the PDMS-2.\textsuperscript{21} The Bayley-3 and PDMS-2 assess developmental constructs similar to play-based measures, while the Gesell does not. There is no published study regarding the structural validity of the AEPS:E motor portion compared to a more current standardized test of motor development than the Gesell. Therefore, we suggest that the TPBA-2 and EMI-IGDI are the best play-based assessments of motor skills to assess a construct similar to traditional standardized tests of motor development.

Structural validity of play-based assessments of cognitive skills was generally lower than play-based assessments of motor skills. Structural validity of the AEPS:E was moderate when compared with the
Battelle Developmental Inventory (BDI).\textsuperscript{20} It is interesting to note that the correlation between the AEPS:E and the BDI was lower in subjects older than 24 months. We hypothesize this may be due to the fact that the older children did not display the full range of their cognitive skills during a play session with a limited set of toys and space in contrast to a traditional standardized developmental assessment which tests the child on specific skills of all difficulty levels. Structural validity was weak to moderate for the EPSI-IGDI\textsuperscript{26} and the PBA.\textsuperscript{16} when compared to the Bayley-2. Child-driven free play tends to decrease validity because the child may or may not show her full repertoire of cognitive skills during a given test session. Validity was strong when more structure was part of the play-based assessment such as in the EPS-I.\textsuperscript{35}

Content validity of the EMI-IGDI, measuring change with increasing age, was significant (p < 0.01) when using a hierarchical linear modeling level 2 design across three age cohorts.\textsuperscript{21} The results indicate that age affects movements during this play-based assessment, which supports the validity of the EMI-IGDI for measuring movement. Having only one study, however, limits generalization of the findings to other play-based assessments.

The methodological quality of measurement properties as defined by COSMIN\textsuperscript{32} in each of the studies in this systematic review was poor to fair with the exception of content validity in one study that was rated as good. One measurement property for three studies was downgraded due to small sample size.\textsuperscript{19,23,34} Other methodological issues for these studies included no evidence of patient stability on test-retest,\textsuperscript{23} methodological flaws in study design,\textsuperscript{19} and unclear handling of missing data.\textsuperscript{34} Use of a Pearson’s correlation coefficient instead of an intraclass correlation coefficient (ICC) for reliability reduced the methodological quality rating for all included studies. Only one study, however, had a reliability score that would have been upgraded had the ICC been used.\textsuperscript{20} While no one methodological issue impacted all the studies, each included study had some methodological problems. Future studies that adhere to
rigorous methodology will provide more detailed information about using play-based assessments for research and clinical measurement.

Play-based assessment tools are designed to assess a child’s ability to use motor and cognitive skills during self-motivated play within contextually-relevant environments. While play-based assessment allows the child to select the activities, this systematic review demonstrates that the overall reliability of these measures is similar to more traditional standardized developmental tests. The slightly lower test-retest reliability is likely the result of the varied responses children display when given toys in the naturalistic environment during play but not specifically prompted to react to the toy in a certain manner, as during a traditional standardized developmental test. The validity results of this review of play-based assessments suggest that as a group, play-based assessments measure a similar, but not identical, construct to traditional standardized developmental tests. Similar to reliability findings, we suggest that slightly lower validity findings of play-based assessments may be acceptable because of the naturalistic context of activities during the assessment. Studies of individual play-based assessments, however, indicate varied structural validity correlations. As a result of varied validity and poor to fair methodological quality, individual tests need additional research to document reliability and validity in high quality studies. At present, results using play-based assessments should be interpreted with caution.

Limitations
This systematic review has several limitations. The broad nature of the search terms resulted in a very large number of titles and abstract which required screening of title and abstract. While only a single author reviewed each title and abstract, the criteria to eliminate a study based on title and abstract was designed to retain any study which might meet the inclusion criteria. Two reviewers completed all other eligibility determination and data extraction. These search criteria could have resulted in missing other potentially relevant studies of the reliability or validity of play-based assessment tools. Also, COSMIN
was originally developed for assessing the quality of health-related patient-reported outcome measures in order to assess complex subjective health changes over time. While play-based assessments do not fit the type of studies usually assessed using the COSMIN, the majority of measurement quality assessment tables could be completed without difficulty.

**Conclusion**

A standardized assessment of skills used during play is critical to determining the need for and efficacy of developmental intervention, yet therapists are challenged to find tools to meet this objective. While the challenge continues, the results of this systematic review demonstrate that play-based assessments have the potential to be reliable and valid tools. Researchers must continue to assess reliability and validity of specific play-based assessment tools and re-assess psychometrics as adaptations are made to the tools. Before play-based assessments can be used as evaluative measures, responsiveness to change must be evaluated. Changes in skills in response to therapeutic intervention, as measured on play-based assessments, would provide evidence that therapy not only can teach a child a new skill, but also that the child can spontaneously use their new skills in daily activity. Determining adequate responsiveness of play-based assessments would give early developmental intervention therapists an opportunity to use play as not just a process of intervention, but also as a reliable and valid method of assessing development. A primary therapeutic goal in all cultures is to enhance a child’s use of functional skills for participation in age-appropriate activities such as play. Play-based assessments improve the ability of clinicians and researchers to measure the impact of therapeutic interventions during these age-appropriate activities for children.
Acknowledgments

Both authors provided concept/idea/research design, writing, data collection and analysis, and project management. Mr O’Grady provided consultation (including review of the manuscript before submission). The authors acknowledge the contributions of Virginia Commonwealth University Health Sciences librarian Jennifer McDaniel for her assistance with defining search terms.

This data was presented as a poster at the Virginia Commonwealth University Graduate Student Symposium; April 22, 2014; Richmond, Virginia. This data has also been submitted as an abstract for the American Physical Therapy Association’s Combined Sections Meeting; February 4-7, 2015; Indianapolis, Indiana.

References


25. Calhoon JM. Comparison of assessment results between a formal standardized measure and a play-based format. *Infant-Toddler Intervention.*


Table and Figure Legends

Tables 1-4:

**PBA**: Play-Based Assessment - cognitive portion of the Transdisciplinary Play-Based Assessment

**PIECES**: Play in Early Childhood Evaluation System

**TPBA-2**: Transdisciplinary Play-Based Assessment, 2nd edition

**EPS-I**: Evaluation and Programming System for Infants and Young Children

**AEPS:E**: Assessment, Evaluation, and Programming System for Eligibility
**EMI-IGDI:** general outcome measure of movement/ Early Movement indicator

**EPSI-IGDI:** Early Problem Solving Indicator

- The EMI-IGDI and EPSI-IGDI are each independent parts of the larger assessment tool, Individual Growth and Development Indicators (**IGDI**)

r: Pearson’s product moment correlation coefficient

**Table 1:** Description of play-based assessments included in this systematic review

**Table 2:** Play-based assessments and studies with motor psychometric properties

**Table 3:** Play-based assessments and studies with cognitive psychometric properties

**Table 4:** Methodological Quality of Measurement Properties using the COSMIN for all included studies

**FIGURE:** PRISMA diagram
Table 1: Description of play-based assessments included in this systematic review
(Note: Some of the tests have two references as 2 papers were published documenting the measurement properties of interest.

<table>
<thead>
<tr>
<th>Play-based Assessment</th>
<th>Motor/ cognitive</th>
<th>Number of items</th>
<th>Item types</th>
<th>Place(s) and items for assessment</th>
<th>Purpose of the assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBA\textsuperscript{16} prototype</td>
<td>cognitive</td>
<td>Not available</td>
<td>Cognitive domain of TPBA (similar in structure to TPBA-2)</td>
<td>neonatal intensive care unit follow-up clinic at a large hospital</td>
<td>Assessing cognition to determine eligibility for early intervention services</td>
</tr>
<tr>
<td>PIECES\textsuperscript{23,47}</td>
<td>cognitive</td>
<td>13 core play behaviors - 86 total items; coded via observation</td>
<td>Cognitive items grouped into subdomains along a developmental continuum consisting of exploratory and pretend play behaviors</td>
<td>Can be conducted any place in which the child feels comfortable; Behaviors during spontaneous unstructured play without adult guidance for a minimum of 30 minutes; toys arranged in the testing room according to general themes such as kitchen, blocks, etc.</td>
<td>Determine highest level of play behavior so that educators can develop interventions to facilitate higher levels of play</td>
</tr>
<tr>
<td>TPBA-2&lt;sup&gt;19,48&lt;/sup&gt;</td>
<td>Motor and cognitive</td>
<td>118 items across all domains and subcategories</td>
<td>4 developmental domains: cognitive, motor, communication, social-emotional; play skills listed in a developmental sequence to attain age equivalence scores for age equivalent comparisons to norm-referenced measures</td>
<td>Informal play setting with manipulatives, representational toys, art materials, construction and play objects, gross motor equipment. Parent and professionals from 3 or more different backgrounds – usually speech-language pathologist, occupational therapist, physical therapist, teacher, psychologist</td>
<td>Assess developmental skills and process, as well as interaction patterns and learning styles</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>EPS-1&lt;sup&gt;34,35&lt;/sup&gt; prototype</td>
<td>Motor and cognitive</td>
<td>164 items, criterion-referenced and curriculum-based</td>
<td>6 domains (including cognitive and 2 motor domains: gross and fine motor); each item scored as pass, inconsistent, or fail; functional skills</td>
<td>Classroom observation during routine activities - including play, activity groups, and snack</td>
<td>Develop systematic methods to plan and evaluate early intervention practices in order to monitor and demonstrate the efficacy of intervention</td>
</tr>
<tr>
<td>AEPS:E 20,49 experimental - part of currently-available AEPS</td>
<td>Motor and cognitive</td>
<td>AEPS contains 249 items across six developmental areas</td>
<td>Curriculum-based; 5 activities using AEPS items with scripted procedures and standard materials (with some flexibility to accommodate individual child routines); each item scored as does not pass, inconsistent performance, or passes consistently</td>
<td>Home, community-based setting with familiar activities and materials and people; parent and caregiver involvement</td>
<td>Authentic assessment of observed behaviors/skills that links assessment outcomes to goal development and planning; AEPS is appropriate for a broad range of needs and diagnoses in children 1 month – 3 years</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>EMI-IGDI</strong> 21,37</td>
<td>motor</td>
<td>5 key skill elements: position transition, grounded locomotion, vertical locomotion, throw/roll, and catch/trap.</td>
<td>Key skill elements represent postural control, locomotion, and object control. Each skill is coded for frequency; Pre-specified toys for exactly 6 minutes.</td>
<td>Typical environment; administered by any early intervention professional trained with IGDI</td>
<td>Monitoring individual growth and making intervention decisions</td>
</tr>
<tr>
<td><strong>EPSI-IGDI</strong> 26,37</td>
<td>cognitive</td>
<td>4 key skill elements: look, explore, function, solution.</td>
<td>Key skill elements represent visual, object exploration, and problem solving. Each skill is coded for frequency; 3 pre-specified toys presented - each for exactly 2 minutes</td>
<td>Typical environment; administered by any early intervention professional trained with IGDI</td>
<td>Monitoring individual growth and making intervention decisions</td>
</tr>
</tbody>
</table>
Table 2: Play-based assessments and studies with motor psychometric properties

<table>
<thead>
<tr>
<th>Play-based Assessment</th>
<th>Study</th>
<th>Sample Characteristics:</th>
<th>Inter-rater reliability</th>
<th>Test-Retest reliability</th>
<th>Structural validity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sample size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Typical/ atypical development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age at beginning of study: range in months, mean, (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPBA-2</td>
<td>Linas\textsuperscript{39}</td>
<td>n = 19</td>
<td>NA</td>
<td>NA</td>
<td>Bayley-3 motor, r = 0.825</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 typical, 7 atypical development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ages: 0-36, 23.05, (10.36)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPS-I</td>
<td>Bailey and Bricker\textsuperscript{34}</td>
<td>n = 32</td>
<td>r = 0.95 (total group)</td>
<td>r = 0.93 (total group)</td>
<td>Gesell gross motor (developmental quotient): r = 0.79 (total group) r = 0.89 (atypical development)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 typical; 22 atypical development</td>
<td>r = 0.95 (atypical development)</td>
<td>r = 0.94 (atypical development)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ages: typical: 20-39, 29.7, (7.5); atypical: 24 -40, 30.7, (4.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bricker et al\textsuperscript{35}</td>
<td>n = 335</td>
<td>r = 0.96</td>
<td>r = 0.95</td>
<td>Bayley motor age r = 0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90 typical, 245 atypical development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ages: 2-72 (majority less than 48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEPS:E</td>
<td>Macy et al\textsuperscript{20}</td>
<td>n = 68</td>
<td>(gross motor) r = 0.86</td>
<td>NA</td>
<td>Gesell gross motor r = 0.62 BDI r = 0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35 typical, 33 atypical development (receiving EI services)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ages: 18-36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMI-IGDI</td>
<td>Greenwood et al\textsuperscript{21}</td>
<td>n = 29</td>
<td>r = 0.98</td>
<td>r = 0.88</td>
<td>PDMS-2 locomotor r = 0.90, 0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24 typical, 5 atypical development</td>
<td></td>
<td></td>
<td>Stationary r = 0.80, 0.77 CAMS-GM r =0.85, 0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age: 1-34, 15.3, (9.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranges</td>
<td></td>
<td>n = 19-335</td>
<td>r = 0.86- 0.98</td>
<td>r = 0.88-.95</td>
<td>r = 0.62 (Gesell) - 0.90 (PDMS-2)</td>
</tr>
</tbody>
</table>
Table 3: Play-based assessments and studies with cognitive psychometric properties

<table>
<thead>
<tr>
<th>Play-based Assessment</th>
<th>Study</th>
<th>Sample Characteristics</th>
<th>Inter-rater reliability</th>
<th>Test-Retest reliability</th>
<th>Structural validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBA</td>
<td>Kelly-Vance et al(^{16})</td>
<td>Sample size: n = 38; 31 typical, 7 atypical development; Age: 23 months 10 days – 27 months 26 days, 24 months 15 days</td>
<td>NA</td>
<td>NA</td>
<td>Bayley-2 MDI: (r = 0.746)</td>
</tr>
<tr>
<td>PIECES</td>
<td>Kelly-Vance and Ryalls(^{23})</td>
<td>Sample size: n = 32; 25 typical, 7 atypical development; Ages: (typical: 19-46, 32.44; atypical: 22-52, 37.57);</td>
<td>NA</td>
<td>Typically developing: (r = 0.48); atypically developing: (r = 0.58)</td>
<td>NA</td>
</tr>
<tr>
<td>TPBA-2</td>
<td>Linas(^{19})</td>
<td>Sample size: n = 19; 12 typical, 7 atypical development; Ages: 0-36, 23.05, (10.36)</td>
<td>NA</td>
<td>NA</td>
<td>Bayley-3 Cognitive, (r = 0.91)</td>
</tr>
<tr>
<td>Measure</td>
<td>Authors</td>
<td>Sample Size</td>
<td>Ages (Typical, Atypical)</td>
<td>Correlation (Overall, Atypical)</td>
<td>Correlation (Overall, Atypical)</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------</td>
<td>-------------</td>
<td>--------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>EPS-I</td>
<td>Bailey and Bricker $^{34}$</td>
<td>$n = 32$</td>
<td>10 typical; 22 atypical development</td>
<td>$r = 0.23$ (total group)</td>
<td>$r = 0.46$ (total group)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20-39, 29.7, (7.5);</td>
<td>$r = 0.32$ (atypical development)</td>
<td>$r = 0.45$ (atypical development)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>24 -40, 30.7, (4.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bricker et al $^{35}$</td>
<td>$n = 335$</td>
<td>90 typical, 245 atypical development</td>
<td>$r = 0.90$</td>
<td>$r = 0.91$</td>
<td>Bayley mental age $r = 0.93$</td>
</tr>
<tr>
<td>AEPS-E</td>
<td>Macy et al $^{20}$</td>
<td>$n = 68$</td>
<td>35 typical, 33 atypical development (receiving EI services)</td>
<td>$r = 0.88$</td>
<td>NA</td>
</tr>
<tr>
<td>EPSI-IGDI</td>
<td>Greenwood et al $^{26}$</td>
<td>$n = 28$</td>
<td>23 typical, 5 atypical development</td>
<td>$r = 0.70$ - 0.99 (four individual skills)</td>
<td>$r = 0.88$</td>
</tr>
<tr>
<td>Ranges</td>
<td></td>
<td>$n = 19-68$</td>
<td>typical and atypical development</td>
<td>$r = 0.23$ - 0.90 (overall inter-rater reliability cannot be determined for Greenwood et al $^{26}$)</td>
<td>$r = 0.45-0.91$</td>
</tr>
</tbody>
</table>
Table 4: Methodological Quality of Measurement Properties using the COSMIN for all included studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Play-Based Assessment</th>
<th>Inter-rater reliability</th>
<th>Test-Retest reliability</th>
<th>Structural validity</th>
<th>Content validity</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenwood et al(^2^1)</td>
<td>EMI-IGDI</td>
<td>fair</td>
<td>poor</td>
<td>fair</td>
<td>good</td>
<td>poor</td>
</tr>
<tr>
<td>Greenwood et al(^2^6)</td>
<td>EPSI-IGDI</td>
<td>fair</td>
<td>fair</td>
<td>fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macy et al(^2^6)</td>
<td>AEPS:E</td>
<td>fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bricker et al(^3^5)</td>
<td>EPS-I</td>
<td>fair</td>
<td>fair</td>
<td>fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bailey and Bricker(^3^4)</td>
<td>EPS-I</td>
<td>poor</td>
<td>poor</td>
<td>poor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kelly-Vance et al(^1^6)</td>
<td>PBA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fair</td>
</tr>
<tr>
<td>Kelly-Vance and Ryalls(^2^3)</td>
<td>PIECES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fair</td>
</tr>
<tr>
<td>Linas(^1^9)</td>
<td>TPBA-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>poor</td>
</tr>
</tbody>
</table>
Records identified through database searching (n = 2250)

Additional records identified through other sources (n = 29)

Records after duplicates removed (n = 2133)

Records excluded (n = 2093)
- Not a play-based assessment (n = 2088)
- Sample not ≤ 36 months of age (n = 4)
- Measurement properties not included (n = 1)

Records screened (n = 2133)

Full-text articles assessed for eligibility (n = 40)

Full-text articles excluded, (n = 32)
- Not a play-based assessment (n = 21)
- Sample not ≤ 36 months of age (n = 7)
- Measurement properties not included (n = 4)

Studies included in qualitative synthesis (n = 8)