For Official Use **EDU/EDPC/ECEC(2014)11/REV1**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **30-Jun-2015**

English - Or. English

DIRECTORATE FOR EDUCATION AND SKILLS
EDUCATION POLICY COMMITTEE

Network on Early Childhood Education and Care

COMPREHENSIVE MEASURES OF CHILD OUTCOMES IN EARLY YEARS: REPORT TO THE OECD

8-9 July 2015
Paris, France - OECD Conference Centre, Paris - Room CC13

This paper was prepared by Dr. Steve Barnett, Dr. Shannon Ayers and Dr. Jessica Francis of the National Institute of Early Education Research, Rutgers, The State University of New Jersey.

Rowena Phair, Project Leader; Tel: +33 (0) 1 85 55 64 10; Email: rowena.phair@oecd.org
Arno Engel, Consultant; Tel: +33 (0) 1 45 24 86 74; Email: arno.engel@oecd.org

JT03379611

EDU/EDPC/ECEC(2014)11/REV1
For Official Use

English - Or. English

COMPREHENSIVE MEASURES OF CHILD OUTCOMES IN EARLY YEARS: REPORT TO THE OECD

STEVE BARNETT, SHANNON AYERS, & JESSICA FRANCIS

1. In this report we provide basic information to inform decision-making regarding the assessment of young children's learning, development, and well-being for national and international data collections designed to inform Early Childhood Education and Care (ECEC) policies. Our primary focus is on the pre-primary years with an emphasis on assessments that are relevant to a broader age range including older children. Given the large number of assessments available, this report begins with a broad overview and then considers specific examples of the various approaches to illustrate strengths and weaknesses rather than conducting an exhaustive review. Several much broader reviews with exhaustive compendia are already available that can be consulted. These include major publications from the U.S. National Academy of Sciences (Snow & Van Hemmel, 2008) and the World Bank (Fernald, Kariger, Engle, & Raikes, 2009).

1. Why is early childhood assessment important?

2. Assessments of young children can provide information about Learning, Development, and Well-Being (LDWB) that is useful to teacher, parents, and others. For teachers assessment can be a tool that informs the care and education they provide to children. Parents often wish to be informed about the progress and wellbeing of their children, less to inform the specifics of their interactions than to be assured that their children are doing well and that the arrangements they have made for their care and education are in the child's best interests. Program administrators can use child assessment data to explore the effectiveness of program design and supports for teachers including professional development. With respect to public policy, there are several valuable uses. Screening, and where indicated, diagnostic assessments conducted on a large scale can identify disabilities and other developmental problems so that children's special needs can be addressed as early as possible. Nationally representative descriptions of children's LDWB and how this varies geographically and with children's family backgrounds as well as with the characteristics of children's ECEC experiences can inform a wide range of public policies to support children and families, keeping in mind that drawing valid causal conclusions regarding public policies and programs is a complex process and imposes strong demands on research design and analysis, not just on assessments.

3. As nations increase their public and private investments to support the care and education of young children, it is to be expected that they will want information about the contributions of these investments to the lives of young children. In particular, there is increased concern about how specific public policies affect children before they enter primary school. This desire to establish cause and effect and to estimate the magnitude of benefits to children's LDWB increases the technical demands on

assessment (discussed below). In addition, causal attributions require more than simply children's describing LDWB over time, it requires rigorous research methodologies that warrant strong causal inferences. Historically, relatively little information of this type has been collected by public agencies prior to age 8, well after entry to primary school in many countries.

1.1 Use and concerns

4. Broadly speaking, the use of assessments can be described as formative or summative. Formative assessment is the use of assessment to inform teaching with some definitions going so far as to equate formative assessment with scaffolding. Formative evaluation is internal and takes place during the educational experience. It looks forward in a process that is responsive to the needs of the learner. Summative assessment is the use of assessment to judge progress or attainment relative to a standard. Summative assessment of the performance of a child looks backwards and may be used to judge the contributions of a teacher or program to child progress. Summative assessment generally is external in its orientation. Summative assessments may be used to inform professional development and other supports for teachers and programs, but they also may be used to make "high stakes" decisions including to sanction or reward teachers, schools, and to inform decisions about public programs and policies. In addition, summative assessments are commonly used to make high stakes about individual children including the provision of additional supports (e.g., special education services and services for immigrant children who have limited proficiency in the local language) and opportunities (e.g., programs for gifted children), as well as to determine whether a child should enter primary school at the typical age or delay entry. The last use is quite controversial and may be viewed as indicating a lack of supports and individualization in the first year of primary school.

5. As it is the use of an assessment that is formative or summative rather than the assessment instrument itself, the same instrument can be used for summative or formative purposes. Confusion can arise because of instruments have been designed so as to be particularly useful for formative or summative purposes, and sometimes the instruments themselves are referred to as formative or summative measures. In addition, there is a tendency to think of qualitative assessments and teacher observations as formative tools. However, the Early Years Foundation Stage Profile (EYFS Profile),discussed in the next section, is an example of an observation-based assessment that is used for primarily summative purposes. The EYFS Profile also provides an example of how the distinction between formative and summative use can become less clear when looked at from a longer term perspective. Even though the assessment is not used to inform teaching in the child's current stage, it is used to inform the child's education at the next stage and to inform changes in policy.

6. Despite the widespread use of assessment, there is widespread concern regarding potential negative consequences. Among the greatest concerns are (1) narrowing of ECEC to focus on what is most easily measured; (2) misuse of assessments for high stakes decisions about children, teachers and programs; and, (3) excessive burdens on children and teachers from time consuming assessments. These concerns have been greatest for direct tests used for summative purposes, but they may arise with any type of assessments regardless of the use for which it was developed. For example, screening tests are sometimes misused to make high stakes decisions about children rather than to refer them for additional assessment. Screening tests also are sometimes used to collect data on a large scale to inform policy because they are quick to administer and so impose minimal costs on everyone involved. However, this should be done in full recognition that screening tests often are designed to err on the side of over identifying problems and may measure better at the lower end than at the higher end of the range of abilities or skills.

7. Concerns about negative impacts of assessment on learning and teaching and misuse by policy makers are lessened when assessment is conducted with broad observational measures embedded in the

educational process for formative purposes. As we discuss in more detail in later sections, teachers may document in detail children's interests, dispositions, learning, development, and well-being as a tool to assist them in providing the best care and education for each child. Yet, even the kinds of data teachers collect for these purposes can be turned to other, summative purposes. Moreover, because the broader and more detailed such assessments become, the greater the time burdens they may impose on teachers.

1.2 Current policy and practice

8. Today, assessment of children's learning and development in the years before the age of 5 is common in OECD countries. Typically teachers conduct such assessments as an integral part of their teaching; most often these are not standardized tests, but ratings, observations and collections of children's work. Less often, teachers formally assess children's well-being, but teachers frequently make judgments about each child's well-being (e.g., happiness, self-actualization, and friendships). Although considered good ECEC practice, it is uncommon for these ECEC assessments by teachers to be required by law. Most often whether, how and when this assessment is conducted is up to the discretion of ECEC providers. As it is good practice, it may be encouraged by public policy guidance. However, some countries (or states in federal systems) require assessment by law, and a few of these specify the assessment to be used and when it is to be used.

9. In a recent survey, OECD countries varied in the extent to which they reported that assessments were used for monitoring purposes in ECEC programs. Many reported that assessments were used as monitoring for summative and formative purposes, with formative use more common. Ireland, Italy, Korea, the Netherlands, and New Zealand reported that they did not use assessments for monitoring purposes. However, to some extent reported differences among countries appear to reflect differences in their interpretation of the questions as well as differences in practice. Generally, the use of assessments by teachers--particularly rating scales, checklists, portfolios, and storytelling--to improve practice is ubiquitous across countries. Much less common is the use of standardized tests for such purposes or the use of standardized tests for any purposes. Standardized tests most often are used to assess language and literacy, motor skills and physical development. Observations and ratings most often assessed children's LDWB very broadly across many domains. The use of assessments for external evaluation of program performance was rarely reported outside of the Americas.

10. Assessments of young children also are collected in national longitudinal and panel studies and, less often, national evaluations of specific ECEC programs. Such studies assess only a sample of young children rather than every child in an age cohort. Typically national samples include children from all ECEC arrangements including those who are only at home with family members. Data from these studies can be used to inform policy makes and the public of the status of children's LDWB and how it is changing over time. Its usefulness for these purposes increases with the frequency with which it is collected.

11. Some public policies regarding assessment are noteworthy as indications of the extent of international variation. France offers teachers the option to use a national test at age 5 for the purpose of better understanding the children they serve and for the teachers' exclusive use. Austria has compulsory language tests 15 months before entry to primary school (Stevens & Dworkin, 2014, p. 71). The purpose of these tests is to ensure that children who do not have adequate German language proficiency receive additional assistance with language development in kindergarten. The tests used differ from one state to another. Germany mandates language tests in kindergarten for similar reasons, but these tests differ by state. Finland has asked children to evaluate their ECEC programs through photos, drawings, and evaluation forms as well as interviews (though these last were conducted only with 48 children).

12. England mandates assessments when a child is between 2- and 3-years-old and at the end of the school year when they turn 5. These assessments are based on teachers' observations. The age 5 assessment is conducted using a rating scale (with a brief narrative), the Early Years Foundation Stage Profile (EYFS Profile). This Profile was recently redesigned, and a new version was introduced in 2012. The Profile provides a broad assessment of child development that is aligned with the EYFS standards. Information is obtained from parents and from teacher observations. The EYFS Profile provides information to parents regarding their child's progress, to the current teacher for use in transition discussions between teachers, and to the teacher who will receive the child in the first year of key stage 1 of primary school for individualized educational planning. In addition, the EYFS Profile is used to construct "an accurate national data set relating to levels of child development at the end of the EYFS which can be used to monitor changes in levels of children's development and their readiness for the next phase of their education both nationally and locally" (Standards and Testing Agency, 2013, p.7). The resulting performance tables are not published at the school-level.

13. Another common use of assessment among OCED and other countries is large scale screening followed by clinical diagnostic assessments to identify disabilities and other developmental problems (including vision and hearing problems) that would benefit from early treatment. However, we did not locate systematic international data on national policies regarding screening and diagnosis of disabilities, delays, and other developmental problems (including hearing and vision limitations). In some countries screening takes place quite early while in others it is not required until well after entry to primary school.

1.3 Illustrative variations within the United States

14. As education policy in the United States varies greatly amongst the 50 states, a brief review of such policies provides insights into the range of different policies that might be adopted. Of the 40 states that offer publicly funded preschool education programs (typically at age 4), the vast majority require the use of some assessments, though not necessarily specifying the assessment or even the type of assessment. Most often this assessment is to be used for formative purposes by teachers, but most states also seek to use this information to inform teacher professional development. A few states require assessments for high-stakes decisions about children (e.g., kindergarten entry) or for summative purposes including the evaluation of teacher and program performance for sanction or reward. Such states may specify a specific assessment to be used with every child enrolled. State policies regarding preschool assessment are summarized in Table 1 below (Schilder and Carolan, 2014).

15. Most states have or will soon adopt Kindergarten Entry Assessments (KEAs) that measure learning and development when children enter kindergarten (the first year of primary school) after turning age 5. The use of these assessments also varies considerably by states. Some states intend these assessments to provide a broad baseline measure that describes children as they enter school. This information would be used by teachers to inform their practice, but also could be aggregated to inform policy makers about the needs of young children and to assess growth between entry at age 5 to kindergarten and the next time the state mandates uniform assessment of every child, typically at the end of third grade. KEAs often have a "whole child" perspective and are not narrowly academic. However, some state KEAs focus primarily on early literacy and, sometimes, a few other academic domains such as mathematics. A few states plan to use these assessments to judge the educational effectiveness of individual ECEC providers and for this purpose the KEA may be aligned with an earlier assessment in the preschool years.

Table 1: U.S.A. state uses of assessments in Pre-K

How Pre-K Assessment Data Are Used By the States	Number of state programs
Guide teacher training, professional development, or technical assistance	35
Track child and program level outcomes over time	34
Make adjustments to curricula	32
Provide a measure of kindergarten readiness	17
Make changes to state policies regarding the preschool program	16
Make decisions regarding a child's enrolment in kindergarten	6
Identify programs for corrective action or sanctions	5
Make funding decisions about programs or grantees	5
Evaluate teacher performance	2

2. Overview: Deciding what and how to assess.

16. From the perspective of obtaining national or international data that can be used to inform policy rather than practice, there are key criteria to be used in deciding what and how to assess. These criteria are as follows:

1. Measure what matters. What aspects of LDWB are important and of concern to policy makers and the public?

2. Measure well. To be useful, measures of what matters must be valid, reliable, fair, and age and developmentally appropriate.

3. Assessments must be practical and affordable. The younger the child, the more difficult it is to accurately assess their LDWB. The broader and deeper the assessment the higher the cost. In addition some aspects of LDWB are more difficult and expensive to assess. Time demands on children, teachers, parents and others can be substantial (opportunity costs such as lost time from teaching), and the costs of professionals specifically hired (and trained) to administer assessments or interviews may be high as well.

4. Results of assessments should be comparable within and across countries and over time.

2.1 Measuring what matters.

17. Children's LDWB encompasses virtually every possible outcome of ECEC including children's happiness and life-satisfaction, habits and dispositions, attitudes and beliefs, cognitive abilities, social abilities, emotional development, physical development, health, and nutritional status. Such a broad view is consistent with the early childhood field's emphasis on attending the needs of the whole child. In addition, one might add measurement of the extent to which a child's rights are respected, for example, the right of children to have a voice or active role in determining the activities in which they are engaged in ECEC. This could be viewed as a means to producing outcomes for the child (for example, life satisfaction and attitudes toward society and schools). However, it could be viewed as an additional category.

18. Both common values and research indicate the importance of comprehensive measures. In most, perhaps all, countries the goals of ECEC are to support the development and well-being of the whole child. This is evident in the Convention on the Rights of the Child (Melton, 2011). From a child's rights perspective we may include "opportunities to express personal agency and creativity, feeling able to contribute, love and care for others, to take on responsibilities and fulfil roles, to identify with personal and community activities, and to share in collective celebrations (Woodhead & Brooker, 2008, p.4). It is also evident in a U.S. National Academy of Sciences report on the science of early childhood development (Shonkoff & Phillips, 2000) which recognized the value of:

> (1) the development of curiosity, self-direction, and persistence in learning situations; (2) the ability to cooperate, demonstrate caring, and resolve conflict with peers; and (3) the capacity to experience the enhanced motivation associated with feeling competent and loved (p.5).

19. Note that we have not described any of these domains or their measures as "outcomes." The use of the term "outcomes" raises the question: Outcomes of what? Children's learning, development, and well-being are affected by all of their experiences at home and with family more generally, in ECEC arrangements, and in the community as well as of their personal attributes. Drawing valid inferences about the specific influence of ECEC experiences and the policies that shape them is much more complex than simply looking at correlations between ECEC and child LDWB measures in a cross-section or longitudinally. One might call for randomized trials, and it is sometimes possible to conduct these with special data collections or in such a way that they can use data that would have been collected anyway. However, randomized trials are not always possible or ideal. It is much more likely that comparisons of the impacts of ECEC and ECEC policies within and across countries will be conducted using complex statistical models that are more successful in producing valid inferences when there are assessments at multiple time points (at least one "pre-test") and when the assessments are accurate and precise. These statistical methods also benefit from linked information on each child's family, home experiences, and ECEC experiences.

20. What should be assessed does depend on the purposes for which an assessment will be used. If policy makers wish to evaluate differences in ECEC quality and services, these may be expected to influence some aspects of learning and development more than others. For example, if the vast majority of young children is healthy and has good motor development, and these are carefully monitored by health professionals then ECEC programs may not much affect these domains. In this case, there may be little reason for educators to assess them. If there is a strong concern that children's rights to engagement and active decision making are not adequately respected, then this aspect of wellbeing may be an important focus of assessment.

2.2 Measuring well: Desirable features of assessments

21. To be useful assessments should be valid and reliable. Assessments also should be fair. In early childhood there is particular concern that assessments be age and developmentally appropriate. This applies equally to all types of assessments, performance assessments as well as tests, qualitative as well as quantitative.

22. Validity is a fundamental criterion for selecting instruments to measure LDBW. The *Standards for Educational and Psychological Testing* state, "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). A valid instrument--whether an observation, interview, questionnaire, or test--should measure what it purports to measure (Williams & Monge, 2001). Validity refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from an instrument so that it is always judged in the context of the purpose for which those inferences are made (Borg, Gall, & Gall, 1989). In assessing validity, what we wish to know is the extent to which interpretations of a measure hold across persons and contexts.

23. In essence, validity is established by producing and evaluating evidence on how well an assessment represents the construct it purports to measure (Messick, 1995). Validity depends on the extent to which an assessment represents the entire construct (i.e., it is not enough that items be from the appropriate domain, they must be fully representative of it). Validity also requires that a measure not include irrelevant items (as, for example, when language demands obscure the demonstration of math or social skills). In other words, an assessment can be invalid because it is too narrow and shallow or because it is too broad. An assessment also can be invalid because accurate representation does not generalize across populations and contexts.

24. There are multiple types of evidence that help to establish construct validity. These include assessments of content by experts, structural evaluation, comparison against a criterion, and prediction. The extent to which experts concur that an assessment fully covers the key dimensions of the construct being measured and does not tap irrelevant areas is sometimes referred to as face validity. Validity is also judged based on the structure of the assessment. Do patterns of results across items conform to theoretical expectations regarding the underlying concepts? Criterion validity can be assessed by examining patterns of performance across ages and concurrent correlations with other assessments of the same construct. A high degree of correlation with an instrument that has well-established validity provides evidence supporting the validity of the target assessment. At the same time a valid measure should not be highly correlated with a measure that is believed to measure a completely different construct. Other approaches include estimating the extent to which the assessment predicts current or subsequent performance in "real life" that is contingent on what is measured.

25. Assuring validity for many assessments is not simply a matter of design, but also of assuring that procedures are appropriate for individual children. An obvious issue occurs when a child's home language differs from that of the assessment. Another is when a child has a disability, and this is most easily understood with respect to vision and hearing impairments. With respect to both issues accommodations often must be made to the child in order to maintain the validity of an assessment.

26. Reliability is the extent to which an assessment produces stable or consistent results because it produces little random error in its results (Creswell, 2008). A reliable assessment produces the same or highly similar results for a child on different occasions (assuming only a brief interval between assessments) and with different assessors (e.g., one teacher would not rate the same child differently from another teacher). A reliable assessment is also robust with respect to the circumstances of the assessment.

27. Reliability can be improved through several means. Optimizing the length or detail of an assessment is one way to increase reliability. The more items, or samples, obtained the less random error affects the results, unless, for example, a longer "test" results in fatigue or distraction for the child or assessor. Another is to construct items and their scoring so as to maximize clarity and minimize uncertainty or misunderstandings. Minimizing the influence of incidental factors in the environment or assessment circumstances and subjective (idiosyncratic) interpretation also increase reliability as does guidance and training for the assessors.

28. Multiple approaches are available to evaluate reliability. One of the most common is examining internal consistency, or how the items (or samples) in the assessment relate to one another. Historically, reliability as judged by internal consistency has been assessed using Chronbach's Alpha, though recently this approach has been challenged and others recommended as more appropriate (Yang & Green, 2011). All of these approaches produce reliability coefficients (a measure of correlation among items). In general, tests that have a reliability of .80 or higher can be considered as sufficiently reliable for most research purposes (Borg, Gall, & Gall, 1989). However, reliability coefficients should be judged carefully since the value adequacy depends on the phenomenon studied (Hancock & Mueller, 2010). Values of .90 have been recommended for assessments used for high stakes decisions about individuals (Yang & Green, 2011).

29. Other common measures of reliability are the correlations of repeated assessments of the same child by the same assessor and inter-rater agreement of different assessors. Inter-rater agreement also may be assessed as criterion-related observer reliability, which is the extent to which a trained observer's scores agree with those of an expert observer (Borg, Gall, & Gall, 1989). It is important because it declares that the trained observer understands the variables measured in the instrument with the same efficacy as an expert observer. Again, there are norms with respect to the extent of agreement required and this depends on the use with the highest levels of agreement required when use relates to an individual child. A high level of reliability is important not just when use is summative but also when used to inform individualized education of a child.

30. Fairness refers to the ways in which assessments are used rather than a property of assessments per se. In addition, it is socially defined rather than scientifically defined. In our view, fairness does depend on validity and reliability because for an assessment's use to be considered fair most would agree that the assessment should be free of bias (e.g., with respect to gender, family background, or national origin) and that random error should not be higher for some types of children than others (at least at the same age). However, even a valid and reliable assessment can be applied in ways that are not fair.

31. One concern in the early childhood field is that assessments developed for older children not be pushed down to younger children when they are neither age nor developmentally appropriate. This concern arises, in part, because of the much greater availability of assessments for older children than for younger children. As demands grow to assess young children on a broader set of domains for which fewer assessments are available, for example, creativity and subjective wellbeing this temptation to use inappropriate assessments only increases. The problem can be avoided by limiting assessments to those with substantial evidence of validity and reliability, which depend on instruments being age and developmentally appropriate.

2.3 Practical Issues: Feasibility and cost

32. In addition to meeting the criteria for validity, reliability, and fairness, a desirable assessment or set of assessments is feasible and affordable. Otherwise, it will not be used or, if used, will create unintended negative consequences. Depending on how information is collected assessments impose costs on children, parents and teachers as respondents or observers. More detailed and comprehensive

assessments impose higher costs on respondents and assessors. In addition there are costs for purchasing assessment tools and training those who administer and use them. If assessment results are made available to people other than those collecting the data there are costs of the systems for storing and sharing data, as well. Finally, the younger the child, the more practical difficulty there is in obtaining information without placing unrealistic demands on the capacities of the child or assessor. As discussed earlier, one of the costs of excessive demands is deterioration in the quality (reliability) of the information obtained. In addition, imposing (unreimbursed) costs on teachers, parents, and others will increase nonresponse rates.

33. The costs to purchase assessments or the tools for their use are minor compared to the actual costs of training assessors and administering assessments. Yet, policy makers sometimes ignore the later and act as if there is no cost for administration if teachers (who are already paid) conduct the assessments. This assumption seems especially likely when assessment is formative and integrated into teaching. However, there is always an opportunity cost, and for teachers this can be quite high. The cost of time spent in classrooms collecting, recording, and reviewing assessment information is best measured by the value of the activities that teachers forego as a result--this can be other forms of planning, but is likely to include direct caring and education of children. Similarly, it should be recognized that parents' time is not "free," and while it is desirable to obtain multiple perspectives, requesting that they provide information imposes opportunity costs on them, as well. Ultimately, the time costs imposed on parents and teachers may result in costs to children by decreasing the time they have to interact with children.

34. Different types of measures not only have different costs, but differ in who bear those costs. For example, brief direct tests or parent interviews (to obtain ratings) impose some costs on children and parents, but could have substantial costs for specially trained assessors who administer the instruments. Direct tests administered by teachers might be brief individually, but require substantial time if obtained for every child in a setting. Depending on the nature of the test it might be perceived by children as enjoyable (e.g., a game) or stressful. On the other hand, portfolios or rating scales completed by teachers may be collected unobtrusively without interfering with the children's activities and requiring no parent time or outside staff. However, teacher assessments may require many hours observing children and recording the results rather than interacting children in ways that directly enhance their wellbeing, learning, and development.

2.4 Comparability

35. For OECD and national policy-related purposes the results of assessments must be broadly comparable over populations and time to be useful. Most countries now have substantial numbers of children from different linguistic, cultural and national backgrounds, and such diversity is by design in international comparisons. Policy questions often span substantial periods of time so assessments should be comparable over a lengthy period. In addition, policymakers often have an interest in continuity and change over the life course so useful assessments will be comparable, at least to some extent, across ages. In this last instance, it is not necessarily the case that scores or ratings would be strictly comparable. What it means for a child to be competent at math or to be highly creative or what social behaviour is considered to be appropriate varies with age. An assessment might ask how many times a young child has a physical conflict in a day as a pre-schooler and in a year as a teenager. However, it is desirable that results on an early childhood assessment of a given construct predict results on assessments of the same construct in the primary and secondary years.

36. Differences in language and culture raise concerns regarding comparability. International studies must confront the problem that languages (and cultures) are not comparable in every respect. Of course, this is an issue that is commonly dealt with in international assessments including the IEA Preprimary Study conducted of children's abilities at age 7. Often this is addressed by expert translation and back-translation evaluated by professional opinion and statistical assessment of differential item functioning

supported by qualitative approaches (Ercikan, 2002; Benitez & Padilla, 2014). Even within a single country language issues arise. Variations in language within countries may be longstanding or the result of recent migration. Children may be monolingual or multilingual. With multilingual children, it must be decided whether the goal of language assessment is to measure the child's knowledge and proficiency in a single language or across all of the languages used by the child.

3. Approaches to assessment.

37. Information on children's LDWB can be collected through a variety of methods, both quantitative and qualitative. Assessments vary in the extent to which they are standardized and in the source (or sources) of their information. Information on children can be obtained directly from children or from those who observe them, most often parents and teachers or other adult caregivers. It may even be obtained from other children (nominations of friends or evaluations of peers to assess relationship status and social skills).

3.1 Tests

38. In education, the first type of assessment that comes to mind for many people is standardized tests. They are widely used to assess cognitive abilities, particularly to assess academic achievement in specific content areas. However, this approach has been used to assess a wide range of cognitive abilities. They were developed to increase the reliability, validity, and especially to increase the fairness of assessments by reducing assessor (particularly teacher) bias. Standardization refers not just to the instrument itself, but also to the process of its administration. It aims to reduce random fluctuations in the circumstances and procedures, and to eliminate systematic biases by the assessor through variations in procedures as well as subjective judgment. Tests may be group or individually administered. As our focus is only assessment prior to primary school we review only individually administered tests; group administered tests are not recommended for children in this age range due to inadequate reliability and validity.

39. Games, including digital games, may be viewed as a type of test. They may be explicitly designed to assess specific knowledge and skills. Their administration and scoring can be more or less standardized. They can be administered "on demand" as are tests generally or children may play them in the ordinary course of their activities. Thus, games as assessments, can share some of the characteristics of authentic or performance assessments, which are discussed in the next section, below.

3.2 Performance assessments and qualitative interviews

40. Another broad type of assessment with which most early childhood professionals are familiar is performance, or authentic, assessment for which observation of children in their everyday activities is the primary basis for data collection (Dunphy, 2008). These assessments typically are embedded in teaching and data are collected continuously during the year and as part of ordinary activities. Documentation can include notes, and observation records, artifacts, art, dictation and children's writing, photographs, and video and audio recordings. Conversations with children and clinical interviews (in-depth, open-ended, and highly sensitive to individual interviewed) are related qualitative methods that may be used to collect information when the phenomena of interest are difficult to observe.

41. The documentation obtained can be collected and organized in portfolios for each child. As developed by Reggio Emilia and other constructivist approaches, representation is a means to involve children in self-assessment as well as for sharing information among teachers and parents (Dunphy, 2008). The results of such assessments may be communicated in highly qualitative form in learning stories and other narratives or quantified in ratings or scores. Narrative approaches seek to maintain the

whole child perspective and recognize the inter-relatedness of children's dispositions, habits, skills, and knowledge as well as the importance of context for understanding children's LDWB.

42. The procedures for conducting and reporting or scoring performance assessments vary from highly standardized to completely unstandardized. Clinical interviews are by design highly individualized and unstandardized, though their methodology is standardized. Such interviews could be considered a separate type of data collection on their own, and have been used to assess children's perceptions of their own decision-making and influence in ECEC (Sheridan, & Pramling Samuelsson, 2001). However, this seems to have been done more often to characterize classrooms or programs than to represent the wellbeing of an individual child.

43. Some performance assessment systems are linked to specific curricula and provide tools and detailed procedures for data collection and scoring based on rubrics. Specific training in the assessment system is part of the professional development for learning to use the curriculum. Some other performance assessments are more general while others are highly specific, but so much a part of an emergent and developing approach to curriculum that they are no more standardized than the curriculum.

3.3 Checklists and rating scales.

44. A third type of assessment frequently used is the checklist or rating scale. Performance assessments can be scored using a checklist or rating scale (and accompanying rubric) either at one point in time or recorded periodically over a year. However, in this section we refer to measures do not necessarily require continuous data collection over time (and are summative in use). Instead, parents, teachers, or other adults rely on their general knowledge of the child or a brief current observation to answer questions about the child's capabilities, personality, dispositions, behaviour, or other characteristics. Such assessments may be standardized in the sense that the precise form and order of the questions has been devised based on research and are not be varied.

3.4 Time diaries

45. Time diaries collect data about children's activities including information about the types of activity, duration of each activity, the place of each activity, and who else was engaged with child as well as what else may have been going on at the same time. They provide a unique and very detailed, approach to assessing children's engaged capacities and wellbeing. Multiple techniques are available including: "beeper studies" where an activity is recorded when prompted, observation, written short recall, and telephone interview short recall. For example, a telephone survey might obtain a 24-hour record by asking a parent: "beginning at midnight yesterday what did your child do?" Basically, this is one long open-ended question with many prompts. Such methods typically are not used with young children. However, parents have been asked to report in telephone surveys for infants and young children, and teachers have been asked to complete written diaries for children in ECEC centers (Barnett & Boyce, 1995; Hofferth & Sandberg, 2001; Rossbach, 1988).

3.5 The different types of informants

46. Information on children's LDWB is obtained not only with different types of techniques, but also from different types of informants. These include parents and other (informal) caregivers, pre-primary and primary teachers (we include here all those responsible for the care and education of children in formal settings including some family home care), and health professionals. In addition, children themselves are key informants and can be active participants in their assessment. There are advantages and disadvantages for each informant when assessing young children's LDWB. Informants may provide

information directly or professionals specifically trained to administer an assessment may be employed to obtain information, typically from children and parents and other caregivers.

47. Parent and other caregivers are valuable informants because of the intimate knowledge they acquire of a child due to their relationship and the time they spend with the child. However, if caregivers are asked to provide ratings relative to an implicit standard or expectation (for example regarding learning, development, relationship quality, life satisfaction or happiness) they may differ greatly from one socio-economic environment and culture to another regarding what is typical or normative (Ertem et al., 2008). Caregivers also tend to provide socially desirable answers. Despite these disadvantages, caregivers' information about children can be valuable and nationally representative information can be readily obtained through household surveys. Some checklists and rating scales are designed to be relatively robust with respect to variations among parents.

48. Teachers in preschool or school settings often provide valuable insights into children's LDWB, thought they can only report on those children who attend ECEC programs. Teachers make good informants because they tend to spend a great deal of time with the children and have working knowledge of and/or training in learning and development. However, teachers vary considerably in their preparation and training. This can be expected to greatly affect their ability to evaluate children's LDWB, especially with performance assessments. The less standardized and more qualitative an assessment, the more the quality of the results (validity, reliability and fairness) depends on the teacher's knowledge and skills regarding both LDBW and assessment. For many instruments, specialized training of the teacher (or other assessor) may be required.

49. The ratio of students to teachers varies considerably and can be fairly high in some countries. Differences in ratios can affect how well teachers know each child and how much time is required of the teachers if asked to assess all of the children for which they are responsible. For these reasons, ratios may be expected to affect the quality of assessment and reduce the reliability (though not necessarily the validity) of assessments in some countries or subpopulations within a country compared to others.

50. Health professionals have advantages as informants of children's development because of their understanding of how children progress through development and in some instances the health services may be the only professional services available to young children. However, for some health professionals, monitoring child development can be a new concept (Ertem et al., 2008). Also, the familiarity with the child and the level expertise possessed by health professionals can vary by socio-economic context. As with teachers, this can affect the reliability and, perhaps, validity, of the assessment.

51. Children are always, in a sense, the basic source of the information on their LDWB. Often, this is indirect and mediated by others. However, young children can provide direct responses in tests, other direct assessments, and interviews. They can be asked to provide ratings. The younger the child, the greater the difficulty of obtaining direct information that is valid and reliable.

4. Critical review of exemplars for comprehensive assessment young children

52. The vast number of options available to assess the LDWB of young children presents a challenge to any review. The number of domains, approaches, and purposes has called forth many different assessments. Fortunately, others have provided exhaustive reviews of the available assessments. Prominent among them are efforts by the World Bank and U.S. National Academy of Sciences (Fernald, Kariger, Engle, & Raikes, 2009; Snow & Van Hemmel, 2008). Additional exhaustive compendia have been developed for the U.S. Department of Health and Human Services (Berry, Bridges, & Zaslow, n.d.)

and the state of Washington (Slentz, Early, & McKenna, 2008). Also, useful is a more focused critical review that addresses key issues in both tests and authentic assessment (Atkins-Burnett, 2007). Currently, UNESCO is developing a Holistic Early Childhood Development Index and has conducted a review of early childhood development and wellbeing indicators to support that project (Tinajero & Loizillon, 2012). A very broad review of early childhood development indicators from an international perspective is provided by Frongillo, Tofail, Hamadani, Warren, and Mehrin (2014).

53. The existing compendia describe the instruments with respect the domains covered, ages at which they are appropriate, methods of administration, strengths and weaknesses, time for administration, and cost. For any specific assessment one may wish to consider they provide a key resource to which readers of this paper can refer. Our purpose here is to consider key illustrations of different approaches that are better known and may be considered guides to the most appropriate possibilities for an international assessment in OECD countries.

54. A list of the instruments reviewed here and the domains that they cover is presented in Table 2, below. In addition, a summary of the information collected on each assessment is presented in a separate set of 3 matrices. Detailed narrative descriptions and evaluations of each follow.

Table 2. Domains covered by exemplar assessments (domains not represented on the table are absent because they were not included in any of the assessments reviewed)

	Physical	**Social/ Emotional**	**Cognitive**	**Communication and Language**	**Executive Function**	**Approaches to Learning**	**Arts/ Creativity**
Zambian Child Assessment Test	✓(fine motor)	✓(parent report)	✓(informatn processing, nonverbal reasoning)	✓	✓		
NIH Measures	✓	✓(psychological well-being, stress, social relationships, negative affect)	✓(EF, attention, memory, language, processing)	✓	✓		
Brigance Early Childhood Screens	✓	✓	✓	✓			
Denver II	✓	✓	✓	✓			
Griffiths Mental Development Scales Extended revised	✓	✓	✓	✓			
Mullen Scales of early learning	✓	✓	✓ (visual reception)	✓			
Schedule of Growing Skills	✓	✓	✓	✓			
Hong Kong Early Development Scale	✓ (gross & fine motor, physical fitness)	✓	✓	✓			
ELS	✓	✓	✓(math, science, literacy)	✓	✓ (self-regulation)		
International Performance Indicators in Primary Schools (iPIPS)	✓ (parent survey)	✓ (teacher rating)	✓ (math, literacy)	✓	✓		
Work Sampling System	✓	✓	✓(literacy, math, science)	✓			✓

	Physical	Social/ Emotional	Cognitive	Communication and Language	Executive Function	Approaches to Learning	Arts/ Creativity
Teaching Strategies GOLD	✓	✓	✓	✓	✓ (self-regulation)	✓(attends & engages, persists, solves problems, curiosity, motivation, flexibility & inventive thinking)	✓
High Scope Child Observation Record	✓	✓	✓ (math, literacy, science, social studies)	✓	□	✓	✓
EDI	✓	✓(social competence & emotional maturity)	✓ (literacy & numeracy)	✓	✓ (aggressive behavior, hyperactivity, inattentive behavior)	✓ (independence & adjustment)	
Kindergarten Entrance Inventory for Connecticut	✓	✓	✓ (numeracy, literacy)	✓			✓
Ages & Stages Questionnaire	✓ (gross & fine motor)	✓		✓		✓ (problem solving)	
Parents' Evaluation of Developmental Status	✓	✓	✓	✓			
Battelle Developmental Inventory	✓	✓	✓	✓	✓ (attention and memory)		
Child Development Inventory	✓	✓	✓	✓			
Early Years Foundation Stage (EYFS)	✓	✓	✓	✓	✓ (attention)	✓ (engagement motivation thinking critically)	✓

4.1 The Zambian Child Assessment Test (ZamCAT; 2012) provides an example of a broad assessment that was constructed by adapting a range of existing instruments each of which is designed to measure specific domains and using variety of methods, but primarily from a one to one direct assessment (testing) perspective.

Purpose: The ZamCAT is a population measure administered to preschool children along with the standard population-based household survey. The ZamCAT is available at http://developingchild.harvard.edu/activities/global_initiative/zambian_project/

Age: Preschool

Format and administration: The ZamCAT followed a mixed approach in development of the tasks on the assessment. Several of the tasks included in the ZamCAT are existing assessments with some adaptations where appropriate; while other tasks are newly developed. The ZamCAT is administered in partnership with the population-based household survey to preschool children by a trained examiner.

The ZamCAT assesses 7 domains of child development by blending existing measures with newly developed tasks. Each component of the assessment is described here. First, ZamCAT evaluates **fine**

motor skills through two tasks. In the first task, the child is asked to copy letters, numbers, and a triangle using a pencil (taken from the Development Assessment in Zambia; Ettliing, et al., 2006). The second is a newly developed timed activity where the child is asked to string beads on a shoelace, place beans into a cup, unbutton and button a shirt, and play a variation of the traditional game nsolo.

Both receptive and expressive **language development** are assessed in the ZamCAT. *Receptive vocabulary* refers to words that a child can comprehend and respond to, even if the child cannot produce those words. The ZamCAT examines receptive vocabulary with 30 items heavily adapted from the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2006). These were modified to be culturally and linguistically appropriate for Zambian children. The authors of the assessment note that scores on this component of the ZamCAT cannot be used to compare to the PPVT (Fink, Matafwali, Moucheraud, and Zuilkowski, 2010).

Expressive vocabulary refers to words that a child can express or produce. The ZamCAT examines this by posing two questions to the child 1) Can you tell me about something exciting that happened to you? 2) Can you tell me about the people you live with at home? These two questions were taken from previous research by Matafwali (2010). The responses are scored 0 (non-responsive) to 5 (multiple-sentence answer using correct grammar). The authors note that this sub-test was used particularly well across languages (Fink, Matafwali, Moucheraud, and Zuilkowski, 2010).

Nonverbal Reasoning is assessed on the ZamCAT with two tests. The first is a newly developed Object Pattern Reasoning (OPR) which uses patterns with concrete items. Here the child is asked to complete patterning sequences. The second test of nonverbal reasoning is taken from the NESPSY (Kirkman, et al., 1998) which is a series of neuropsychological tests. The NESPSY Block Test is used in the ZamCAT as an assessment that measures the child's ability to capture, analyze, and replicate abstract forms. In this test the child is asked to assemble blocks in reproduction of a pictured design.

Information Processing is assessed on the ZamCAT with the Rapid Automatized Naming (RAN; Dencla & Rudel, 1976) task. This task has the child look at pictures, colors, letters, or numbers and then name them as quickly as possible. However, for the ZamCAT only the picture subtest of this assessment was used. Photos for this test include chair, tree, bicycle, duck, and scissors.

Letter naming was included in the ZamCAT to examine children's preparedness for early literacy. Children were given two minutes to name letters shown in random order on a piece of paper.

Executive Functioning is assessed on the ZamCAT through two tasks. Attention is examined through the Pencil Tapping test (Brooker, Okello, et al., 2010). This is where children need to remember and apply the "rules" of the game (when to tap with the pencil) and the task is made more difficult by also providing the child another small task to divide his or her attention. Executive functioning is also examined by assessing a child's delayed gratification (impulse control). Previous assessments used either candy or a wrapped gift to measure this component of executive functioning. The ZamCAT offers one candy immediately or two candies if the child waits until the assessor is done talking to parents to receive his or her treat. Several issues arose with this task. For instance, some parents didn't allow candy, children were reluctant to take candy from strangers, or they lost candy to older siblings so assessors needed to give candy to all of the family.

Socio-emotional Development is measured by parent report on the ZamCAT. This is a series of 20 questions to capture parents' overall perceptions of development. The responses to the questions regarding if the child displayed the behavior are never, sometimes, usually, always.

Task Orientation is rated by the child evaluator. The evaluator rates children on their attitude and performance during the child assessment tasks. The rating scale measures executive function, compliance, and attention as rated by the child evaluator.

Developmental domains covered: This assessment examines nonverbal reasoning, receptive and expressive language, fine motor skills, information processing, socio-emotional development, task orientation, and executive function through direct assessments of young children. Additionally, this survey instrument includes an extensive questionnaire regarding the mother's health and health care during pregnancy and the child's health during the first few years of life.

Time Required: The total battery - including the child assessment and the questions asked to caregivers - takes between 90 and 120 minutes; the child assessment itself takes between 30 and 45 minutes on average, but varies quite a bit depending on how easily the child manages to do the tasks (G. Fink, personal communication, August 12, 2014).

Training and materials: Training for Administration is extensive. The assessors in previous studies have participated in training for 5 days, which could be considered on the short side for this type of assessment. The researchers tried to give feedback through supervisors on a daily basis during field work. It is recommended that 2 weeks of training with extensive field trials be implemented as a more rigorous approach to training. Training usually is divided into 5 parts: 1) Getting familiar with the tool: objectives, concepts, procedures; 2) Introduction to tool administration: rules, guidelines, and practical issues, including a mock assessment by trainers; 3) Within group practicing - 2-3 full assessments of other trainers; 4) Translation: group interviewers grouped by language to review the translation of all instructions and items; 5) Supervised field tests which are full assessment of 3-5 children, partially supervised (G. Fink, personal communication, August 12, 2014).

Technical (psychometric) properties: The ZamCAT reports reliability information using the Chronbach's Alpha. The Cronbach's Alpha coefficients reported for fine motor, receptive language, pattern reasoning, pencil tap test, and task orientation are between .75 and .91. The lowest at .75 is Object Pattern Reasoning and the highest at .91 is Task Orientation. This shows that the internal consistency of these tasks is within acceptable range. It is not surprising that Task Orientation demonstrated the highest internal consistency. This is a rating scale completed by the evaluator and often perceptions of development reported by evaluators or observers tend to demonstrate high correlations between items or domains.

No evidence of validity was reported for the ZamCAT. Although some of the tasks would be shown to have validity because they are already established measures.

Use: This assessment was developed as a part of the larger collaboration between the Zambian Ministry of Education, the Examination Council of Zambia, UNICEF, the University of Zambia, and the Center on the Developing Child at Harvard University which launched the Zambian Early Childhood Development Project (ZECDP) in 2009. The ZECDP is an effort to measure the effects of an ongoing anti-malaria initiative on children's development in Zambia. The intention was to develop a tool that could 1) provide internationally comparable measures of child development across domains; 2) be sensitive to local culture and linguistic differences; 3) be adapted for other developing countries (Fink, Matafwali, Moucheraud, and Zuilkowski, 2010).

Strengths:

1. ZamCAT covers a broad range of developmental domains.

2. Several of the domains assessed are less-commonly evaluated on large scale measures such as executive functioning and socio-emotional development.

3. The ZamCAT demonstrates that child development measures can accompany standard population-based household surveys.

Limitations:

1. The ZamCAT is limited in age use to preschool-age children.

2. A validity evaluation is needed. With so many alterations to the published measures and the development of new measures an examination of the validity is warranted.

3. The ZamCAT has only been used with children in Zambia (although it was used across several local languages).

4.2 The Ages and Stages Questionnaires (ASQ-3; 2009) is an example of a single parent rating scale that provides a very broad assessment of children's learning and development.

Purpose: The ASQ is primarily used to screen for developmental delays, but it has been used in research, as well.

Age: 1 month to 66 months (5 ½ years).

Format and administration: ASQ-3 is a developmental screening system comprising 21 age specific questionnaires for children between ages 1 month and 5 ½ years. Each questionnaire is completed by parents and includes a short demographic section and then 30 questions about the child's development. The child development questions are divided into five domains. Parents respond using the options of *'yes', 'sometimes' 'not yet'.* Questions are phrased at a reading level for 4^{th}-5^{th} US school grade, which is roughly equivalent to a reading age of 9-10 years.

Developmental domains covered: The ASQ reports on communication, gross motor, fine motor, problem solving, and personal-social domains.

Time required: The ASQ takes approximately 10 to 15 minutes for a parent to complete and 2-3 minutes for professionals to score.

Training and materials: Little training is required for paraprofessionals or office staff to score the questionnaires. A User's Guide and training materials are available. Questionnaires, forms, letters, and activity sheets in the user's guides can be reproduced as many times as needed by a single site. Questionnaires are available in English or Spanish.

Scoring: The ASQ-3 results in a score (out of 60) for each area (communication, gross motor, fine motor, problem solving and personal-social) and these are compared to cut-off points on the scoring sheet. Scores beneath the cut-off points indicate a need for further assessment; scores near the cut-off points call for discussion and monitoring; and scores above the cut-off suggest the child is on track developmentally.

Technical properties: The ASQ-3 was standardized on 15,138 children in the United States whose parents completed 18,232 questionnaires. Families were educationally and economically diverse, and their ethnicities roughly matched estimates from the 2007 U.S. Census. Sensitivity (proportion of positives for developmental delay correctly identified) was .86 and specificity (proportion of negatives for developmental delay correctly identified) was .85 overall. Figures for sensitivity and specificity at key ages between 24-30 months are given below:

At 24 months: sensitivity 91.2%, specificity 71.9%

At 27 months: sensitivity 77.8%, specificity 86.4%

At 30 months: sensitivity 86.7%, specificity 93.3%

The ASQ has been validated using the Bayley Scales of Infant Development II (BSID-II) and found to have a sensitivity of 100% and specificity of 87% at 24 months for severely delayed status.

Use: The ASQ-3 has been translated and used in a number of settings (e.g., France, Norway, Finland, Spain, the Netherlands, Turkey, North America, South America, Asia, and Australia). It has been used in studies with the general pediatric population and with children at increased risk for disability. Parents report that they find the questionnaires easy and quick to complete and they have been found to complete the questionnaire with reasonable accuracy.

Strengths:

1. ASQ-3 covers a broad range of developmental domains.

2. ASQ-3 produces scores (out of 60) for each domain and an overall score, which may allow measurement of small changes longitudinally.

3. Its format allows flexibility in administration. For example, for parents who may have difficulties with literacy or with language barriers, another individual could go through the items with the parent at the time of the administration. This would be a useful way of increasing access.

4. The authors comment that an important difference between this and other screening tools is that it is designed to show what children can do, not just what they cannot do.

5. The ASQ reports acceptable sensitivity and specificity.

6. The ASQ has been used among children at high risk of developmental problems.

7. It is quick and easy to complete and to score.

8. The ASQ is cost efficient as a one-off purchase with questionnaires and other materials being photocopied as required.

Limitations:

1. It has only been standardized in the USA so there is a lack of standardized norms for other populations.

2. In only a few studies has its psychometric properties been examined in their own cultural setting after translation.

3. ASQ-3 covers a broad range of developmental domains, but does not include social-emotional development, thus issues such as relationships are less well covered. However, ASQ-Social-Emotional focuses solely on social and emotional development, and could be used in conjunction with ASQ-3.

4. It is not clear whether it is valid to combine the scores from age specific questionnaires into one overall score.

5. Some of the language used in ASQ is 'Americanized'. Parents' understanding of this needs to be assessed and it possibly needs to be adapted for use in other settings.

4.3 Early Development Instrument (EDI; 1998). The EDI is an example of a rating scale completed by teachers or parents that has been very widely used internationally.

Purpose: The EDI is used as a screening tool for Kindergarten readiness.

Age: 4 to 7 years

Format and administration: The Early Development Instrument (EDI) is an assessment tool that provides a standard measurement that can help assess where children are and what areas need to be addressed to ensure that children start kindergarten ready to learn. Teachers complete a 104-item questionnaire on each student, for which they check whether or not students have met specific developmental milestones across five domains.

Developmental domains covered: The EDI reports on physical well-being, social competence, emotional maturity, language and cognitive development, communication, and general knowledge.

Training and materials: Training is a necessary preliminary step to EDI implementation. A copy of the EDI Guide should be provided to each teacher respondent. In addition, a training/information session will ensure accurate, consistent interpretation of items, as well as inform respondents about the purpose of data collection, how results will be used, and the logistics of the data collection process. Respondents with some training in the early childhood area will likely require only minimal training on the use of the EDI. Questionnaires are available in English or French.

Scoring: Average scores are calculated for the 5 domains and 16 sub-domains. Scores are used to identify percentile ranks. Scores also allow for an estimate of the overall percentage of children vulnerable in school readiness. Scores are categorized as follows:

- **On track (Very Ready)** - The total group of children who score in the best 25% of the site's distribution.

- **On track (Ready)** - The total group of children who score between the 75th and 25th percentiles of the site's distribution.

- **Not on track (At risk)** - The total group of children who score between the lowest 10th and 25th percentile of the site's distribution.

Technical properties: Since 1999, EDI data have been collected for more than 300,000 children ages 4–5 years in Canada and several other countries. A subset of the database, consisting of data collected from 2000 and later, has been analyzed to establish normative values for the EDI domains. The subset comprises 116,860 kindergarten children.

The EDI has also been tested to ensure its reliability and validity psychometrically (Janus and Offord, 2007). Internal reliability and test-retest reliability is high for each domain (ranging from .82 to .96). However, parent-teacher correlations were low (ranging from .36 to .64). Concurrent, external, and predictive validity have also been reported, and there are a wide range of correlations depending upon the type of validity and specific comparison. For example, emotional maturity domain of the EDI has a

correlation of only 0.11 with *PPVT* scores and 0.73 with the social-emotional subscale of *First Step*, a comprehensive screener used for identifying developmental delays in preschool children.

Use: The EDI has been translated and used in a number of settings (e.g., some regions in the United States, Canada, Australia, Chile, Egypt, England, Holland, New Zealand, and implementation in Jamaica, Kosovo, Moldova, and Mexico). It is to be completed by teachers in kindergarten classes after several months of observations. Primary uses are the following:

- Serves as a population-level measure for interpreting outcomes for groups of children.
- Yields results that could be used by communities to identify weak and strong sectors.
- Encourages communities to mobilize and make plans to improve children's outcomes.

Strengths:

1. EDI covers a broad range of developmental domains.
2. EDI is an effective tool to assist decision-makers at various levels with resource planning for children.
3. EDI maintains consistent core concepts but it is culturally adaptable to local communities.
4. EDI is malleable to various populations.
5. The EDI is a helpful tool for determining school readiness.
6. Repeating data collection over time using the EDI in the same communities or regions makes it feasible to assess change.

Limitations of EDI:

The technical adequacy of the EDI is unclear. Particularly concerning are the strong differences between teacher and parent ratings. Hopefully, there will be additional information available in the near future.

4.4 GOLD (Teaching Strategies) illustrates a performance assessment that is widely used for formative purposes, but which also has been used as a summative assessment.

Purpose: Teaching Strategies GOLD tracks children's efforts, achievements, and progress. It is designed to inform instruction and enhancing learning outcomes.

Age: Birth through Kindergarten

Format and Administration: Teachers rate children's skills, knowledge and behaviours along a progression of development and learning. Scoring for GOLD varies slightly on the objectives. Objectives 1 through 23 are scored on a 0 to 9 point scale and objectives 24 through 36 are scored on a 0 to 2 point scale. The scale has descriptions or "indicator levels" at scores 2, 4, 6, and 8 that describe the developmental continuum. The scores without "indicator levels" are used to document that the skill may be emerging but not yet fully established. This assessment is available in English and Spanish.

Developmental Domains Covered: GOLD evaluates social–emotional, physical, language, cognitive, literacy, mathematics, science and technology, the arts, and English language acquisition (where appropriate)

Time Required: Teachers collect evidence (anecdotal notes, work samples, photographs, etc.) of children's development over a period of between 4 and 12 weeks. Then, use this evidence to score the children on the rating scale. In one survey of kindergarten teachers, the teachers reported using 1.5-3.0 hours of documentation time per child over an 11 week period (Williford, Downer, & Hamre, 2013).

Training and Materials: Training for the assessment ranges from one day in-person training to several days. Two days of in-person training is typical. There is also an online training option. Teachers using the instrument are offered optional participation in an inter-rater reliability certification. Here teachers analyze portfolios, score the data, and those scores are then compared with those of Teaching Strategies GOLD developers, with an agreement goal of 80% or better agreement.

Technical properties: Norms were calculated using a nationally representative norm sample of 18,000 children from 50 states, Puerto Rico, and Washington, DC and spans across age cohorts. GOLD provides norm tables across all six areas of development (Teaching Strategies, 2013). Each norm table includes expected scores for children across 24 different 3-month age bands from 0-71 months. There are norms for fall, winter, and spring. A study by Kim and Smith (2010) with infants through children aged 2 years showed high internal consistency reliability with a coefficient of .95-.99. This study also showed moderately high reliability. Teaching Strategies (Teaching Strategies, 2013) reports on GOLD's concurrent validity with a study looking at preschool children. First, teacher rating scales of children's social functioning and their learning behaviours related to the GOLD scale scores (r = .426-.541). Second, the related GOLD sub-scale scores for preschool children correlated low to moderately to standardized test scores on appropriate standardized assessment (r = .307-.522). These standardized assessments included the Peabody Picture Vocabulary Test, Pre-Language Assessment Scales, Woodcock-Johnson III Tests of Achievement, The Pencil Tapping portion of the Preschool Self-Regulation Assessment, and the Head-Toes-Knees-Shoulders Task. In an evaluation of GOLD in kindergarten (Williford, Downer, & Hamre, 2013), all teachers began the reliability process, 50% completed the tasks and 28% achieved reliability certification for all domains. Additionally, this research showed that the GOLD assessment was similar in scores to direct assessments in mathematics (r = .64) and literacy (r = .53), but children's language and cognition were not as similar to the direct assessments of language (r = .36) and self-regulation skills (r = .27 and .31). Further, this study concluded that GOLD appears to measure children from different backgrounds equitably.

Use: Teaching Strategies GOLD and its predecessor, The Developmental Continuum, has been used extensively in the US in early education classrooms. Recently, GOLD has been used more consistently as a kindergarten entry assessment. In addition, Teaching Strategies is in development of a first through third grade assessment tool.

Strengths:

1. GOLD covers a broad range of developmental domains.

2. GOLD is a teacher-administered tool based on the child's performance in his or her natural learning environment.

3. This assessment will help inform instruction for children.

4. GOLD is used widely in the United States.

5. GOLD has evidence of effective use with children with disabilities and dual language learners and appears to measure children from different backgrounds equitably.

6. GOLD provides normative data.

7. GOLD covers a large age span (currently birth through kindergarten with up to third grade in development).

Limitations:

1. Ongoing support of use in the classroom is generally needed when implementing an observation-based assessment system of this magnitude.

2. This systematic approach to assessment in the classroom environment may be new to many teachers and can be cumbersome at first.

3. This type of assessment should not necessarily be used alone for high-stakes decisions on individuals or programs.

4. The GOLD is normed on US students only.

5. GOLD reports varying psychometric properties depending on the study and age group of the children.

6. GOLD may more accurately assess math and literacy skills than other skills on the instrument.

4.5 The Hong Kong Early Child Development Scale (HKECDS; 2012) provides an example of a broad direct assessment developed outside North America designed to give a holistic assessment of child development.

Purpose: The HKECDS is used to assess the holistic development of preschool children as well as incorporating current expectations of early child development in Hong Kong. It can be used to evaluate the efficacy of targeted interventions and broader child-related public policies in early child development in Hong Kong.

Age: Preschool (ages 3 to 6).

Format and administration: The HKECDS relies on direct assessment with preschool-age children by a trained examiner. It is a developmental scale such that older children achieve higher scores in each learning domain.

Developmental Domains Covered: The HKECDS examines the following 8 learning domains with 95 test items: personal, social and self-care (9 items), language development (14 items), pre-academic learning (29 items), cognitive development (10 items), gross motor (12 items) fine motor (10 items), physical fitness, health and safety (9 items), and self and society (10 items).

Time Required: The total battery takes 30 to 45-minutes for all 95 items. The original version consisted of 190 items and required two testing sessions of 30 to 45-minutes each.

Training and Materials: Assessors in the validation study were undergraduate and graduate students majoring in early childhood education. Prior to formal data collection, they were required to go through all test items, instructions and materials with the second author. In addition, they had to achieve an inter-rater reliability of about 90% agreement before starting formal data collection. For each item, assessors use standardized stimuli and follow standardized instructions, procedures, and scoring rules. Gross motor activities are conducted either inside or outside the room, depending upon the space in the room. Specialized training is required for future assessors who wish to administer the HKECDS.

Technical properties: The HKECDS reports reliability information using the Chronbach's Alpha. This coefficient assesses the reliability of a test by examining the internal consistency. The following Cronbach's Alpha coefficients are reported for the domains: Personal, Social and Self-care .63, Language development .80, Pre-academic learning .95, Cognitive Development .70, Gross Motor .78, Fine Motor .75, Physical Fitness, Health and Safety .61, Self and Society .64. However, there were also moderate inter-correlations among subscales that assess theoretically different constructs.

Use: This instrument is used to assess the holistic development of preschool children as well as incorporating current expectations of early child development in Hong Kong.

Strengths:

1. The HKECDS covers a broad range of developmental domains.

2. Results from the validation study indicate that the HKECDS is a psychometrically robust, culturally and contextually appropriate measure of holistic child development for children ages 3 through 6.

3. Items in the HKECDS tap culturally sensitive expectations in each domain (for example, the measure examines young children's development of finger coordination with chopsticks, which is a unique activity in the Chinese culture).

4. Domains that are represented are translatable to children in different countries.

Limitations:

1. While this tool is valuable for its culturally sensitivity in Hong Kong, it is not adaptive to alternative populations.

2. The validation sample was not a representative sample of children in Hong Kong.

4.6 International Performance Indicators in Primary Schools (iPIPS) was explicitly developed as a broad assessment for use in international studies.

Purpose: The iPIPS was developed based on the Performance Indicators in Primary Schools (PIPS). It is used to assess what children know and can do, and how that changes during the first year of school. It also includes Personal Social and Emotional Development (BSED), Behaviour and Physical Development. It collects information about prior educational experiences from parents or guardians and information about school from teachers. The intention is to use iPIPS in an international study to examine how one country compares to another at the start of school and after one year of schooling. Collecting data at this early

stage also provides information to the extent that differences in later international studies (e.g., PISA) can be explained by differences in the early years.

Age: Used in the first year of formal schooling; at ages 4-7.

Format and administration: The direct assessment part of iPIPS is administered one-on-one either using a computer adaptive test or using a booklet with an application on a smart phone. Teachers complete a questionnaire rating children's Personal Social and Emotional development. Parental information can be collected using paper based questionnaires or over the internet.

Developmental Domains Covered: iPIPS directly assesses early reading, phonological awareness, early mathematics and short term memory. To be more specific, it includes: name writing (hand writing), picture vocabulary, ideas about reading, concepts of print, phonics, phonological awareness, letter identification, reading and word attack skills, word recognition and decoding skills, comprehension, early math, ideas about math, size and location, counting ability, simple number problems, digit identification using single, double and triple digits, and shapes. Optional items include: short-term memory, behaviour and attitudes. Teacher completed rating scales collect data on 12 aspects of Personal, Social and Emotional development and also can provide 18 items on behaviour. A parent survey collects basic data on children's physical development (height, weight, fine and gross motor coordination).

Time required: Direct assessment (computer or booklet) takes approximately 20 minutes. Additional time is required for the parent questionnaire/interview and supplementary surveys.

Training and materials: Time is required to become familiar with the user guide to get to know the computer system for the PIPS.

Scoring: Reports are generated by each country for schools using iPIPS. For the PIPS system the data are available online together with software to allow teachers to use the data. It is reported that it takes 30 to 60 minutes to access and interpret reports for the PIPS.

Technical properties: The Technical Report for PIPS version on a CD ROM from 2001 provides information about the reliability and validity of the instrument. It demonstrates test-retest reliability on 29 students who were re-assessed was 0.98 for the instrument. The subtests ranged from .34 to .99. Others have reported good reliability with different populations (Godfrey & Galloway, 2004). Predictive validity of the PIPS is demonstrated through the correlations ranging from .48 to .66 on assessments given up to 6 years later. The PIPS baseline assessment has been standardized to have a mean of 50 and a standard deviation of 10. The assessment provides conversion charts that offer age-corrected standardized test scores. Reliability and predictive validity data on the PSED and Behavioural part have also been published (Merrill & Tymms, 2001).

Use*:* The PIPS has been used in Australia, Netherlands, Scotland, New Zealand, Abu Dhabi, Germany, and South Africa. In England, PIPS has been widely used as on-entry assessment. To date, versions have been created and used in Dutch (where it is known as OBIS), German (where it is known as FIPS), Russian, Spanish, French, Slovenian and Chinese (both Cantonese and Mandarin), Afrikaans and Sepedi (another Southern African language).

Strengths:

1. Teachers and students reported to enjoy the computer delivery.

2. Teachers report that the program is easy to use.

3. The PIPS has been used wide-scale in several countries for 20 years across several languages.

Limitations:

1. Technology and IT support may be necessary to use the computer-based version.

2. Psychometric properties were found only for PIPS, not iPIPS, though it may be reasonable to extrapolate as they are highly similar. Psychometric properties must be established for each country.

3. The vocabulary and phonological awareness scales have proved the most challenging in terms of generating equivalent versions for the different languages and cultures, and it has been noted that these two scales should not be used for international comparisons (Tymms, Merrell, Hawker, & Nicholson, 2014).

4.7 Brigance Screens III (2013) is an example of a broad screening test relies on direct assessment by an educator or other expert in addition to observation.

Purpose: The Brigance Screens help educators identify potential developmental delays and giftedness, reduce over-referrals with at-risk cut-offs, determine each child's specific strengths and needs, and assess school readiness.

Age: 0–35 months includes Screens for Infants, Toddlers, and 2-Year-Olds

3–5 years includes Screens for 3-, 4-, and 5-Year-Olds

K & 1 includes Screens for 5- and 6-Year-Olds

Format and administration: Brigance Screens are administered by an educator (teacher, data collector, etc.). Educators spend only 10-15 minutes with each child in order to assess the first three domains (physical development, language, and academic/cognitive). These data are paired with parent and teacher observation of self-help and social-emotional skills to provide a quick snapshot of a child's skill mastery.

Developmental Domains Covered: The Brigance Screens assess physical development, language, academic/cognitive, self-help, and social-emotional skills.

Time required: Teachers spend approximately 10-15 minutes with each child, and then parents and teachers complete the assessment through observation.

Training and materials: Free online training is available on the publisher's website. Brigance Screens require very few resources to implement. Educators need the Screen Manual, a Data Sheet, and, for very young children, Screen accessories. For those sites that wish to enter the data on the online system, internet access is required.

Technical properties: Earlier versions of the Brigance Screens have demonstrated acceptable reliability and validity (Hamilton, 2006; Glascoe, 2002). Brigance Screens III (2013) also reports acceptable reliability and validity. The standardization of the assessment was conducted on a sample of children that was nationally representative in the United States in terms of geographic, demographic, and socioeconomic characteristics. Reliability is reported within acceptable ranges. Specifically, internal

consistency is reported as .90 or higher, inter-rater reliability at .80 or higher and test-retest results were stable when tested at multiple points in time. Construct validity is demonstrated by the domain score structure of the assessment validated by confirmatory factor analysis. Differential item functioning analysis was used to examine for bias of gender and race along with a review panel these two methods showed no biased items. Content validity was reported by researchers and educators that the items on the assessment test the important developmental and early academic skills. The Brigance Screens III is reported by the publisher to correlate with other achievement, intelligence, and language tests such as the Vineland II and Woodcock Johnson III. However, exact correlations were not reported. Lastly, the publisher reports that the assessment correctly identifies the children with true developmental delays or disabilities demonstrating accuracy for sensitivity.

Use: The Brigance Screens are used widely in the US mostly in educational settings.

Strengths:

1. The Brigance Screens cover a broad range of developmental domains.
2. This screening assessment can be administered quickly.
3. This assessment spans a wide age range.
4. The Brigance Screens includes parents and/or teachers in rating.

Limitations:

1. It is available in English, Spanish, Laotian*, Vietnamese*, Cambodian*, Taglog* (*For K&1 Screen, kindergarten level only.)
2. This assessment is generally administered by educators.
3. It was standardized on US children only.

4.8 The Kindergarten Entrance Inventory for Connecticut (KEI) is illustrative of teacher ratings of a type widely used in the United Sates for children entering primary school that offers relatively broad coverage of early learning and development.

Purpose: Kindergarten Entry Assessment to measure children's preparedness for kindergarten. Gives a state-wide snapshot of the skills and behaviours students demonstrate.

Age: 5-6

Format and Administration: Based on teachers' observations at the beginning of the kindergarten year. Teachers assign ratings on 6 domains that are defined by 3-5 indicators each.

Developmental Domains Covered: The KEI assesses language skills, literacy skills, numeracy skills, physical/motor skills, creative/aesthetic skills, and personal/social skills.

Time Required: Administration of this assessment requires time to observe the students to get to know them well enough for the teacher to complete this rating scale.

Training and Materials: The rating scale is the only material needed. It does not appear that much training is required for the teacher to rate the children (but the consequences for reliability and validity are unknown).

Scoring: The teacher rates each indicator in the domains on a scale of 1 to 3. Students at a score of 1, "demonstrate emerging skills in the specified domain and require a large degree of instructional support." Students at a score of 2, "inconsistently demonstrate the skills in the specified domain and require some instructional support." Students at a score of 3 "consistently demonstrate the skills in the specified domain and require minimal instructional support."

Technical properties: The validity of the KEI was evaluated by comparing the content to the state preschool framework and curriculum. This comparison was reviewed by teachers in preschool and kindergarten. This instrument demonstrates a relationship to later grade 3 reading proficiency as assessed by the standardized state test.

Uses: The KEI is used in one state in the US as a comprehensive evaluation of children entering kindergarten.

Strengths:

1. The KEI covers a broad range of developmental domains.
2. Materials and training are not required.

Limitations:

1. The KEI I used only small-scale in the U.S.
2. This assessment is only designed for children at age 5.
3. Reliability and validity are largely unknown.

4.9 Evaluation of Potential for Creativity (EPoC; 2011) provides an example of an assessment designed to specifically measure important dimensions missing from many broad assessments.

Purpose: This assessment is used to measure two main modes of creative thinking.

Age: Elementary-middle school students (grades K-6)

Format and administration: EPoC includes two forms (A and B) to assess progress (pre- and post-test). Each form consists of 8 subtests which cover two domains of expression (verbal and graphic) as well as two modes of thinking [divergent-exploratory (D-E) and convergent-integrative (C-I) thinking]. For instance, divergent-exploratory verbal-type tasks, children generate ideas in response to one stimulus or problem (e.g., A DE verbal domain task is to propose as many story endings to a story beginning as possible within 10 minutes). In C-I graphic-type tasks, children are asked to produce an integrated, elaborate and finalized composition (e.g., A CI graphic domain task is to generate an original drawing which combines a set of heterogeneous elements presented on a photo within 15 minutes).

Developmental domains covered: The EPoC examines creativity, cognition, and problem-solving.

Time required: Ten minutes are required for each divergent-exploratory task, and 15 minutes for each convergent-integrative task. There is also an allotted time warm-up activity before verbal tasks. This is a total of nearly 2 hours to complete this assessment.

Training and materials: Electronic version requires a computer (with recording for verbal tasks, and drawing program and a mouse for graphic tasks). The paper and pencil version requires only paper and pencil for the graphic tasks (verbal tasks are completed orally). Judges who score the integrative tasks are trained on the criteria and benchmarks for scores on the 7-point Likert scale, and then scores are compared among the judges.

Scoring: For divergent-exploratory tasks scoring is based on the number of ideas generated or a count the number of verbal or graphic productions. To score the convergent-integrative tasks, a 7-point Likert scale (1=low, 7=high) is used by independent judges to rate each drawing or story. Concluding the tasks, four scores are computed: Divergent-Exploratory thinking in the Graphic domain (DG), Divergent-Exploratory thinking in the Verbal domain (DV), Convergent-Integrative thinking in the Graphic domain (IG), and Convergent-Integrative thinking in the Verbal domain (IV).

Technical properties: EPoC was developed and validated with a sample of French students. Test scores were reliable with inter-subset correlations ranging from .60 to .78, and external validity was reported to be satisfactory (Baptiste, 2011). In one study, 48 Chinese children from a primary school in Hong Kong were tested for creative potential using the EPoC (electronic and paper& pencil version). For the electronic version, the Cronbach's alpha for verbal divergent-exploratory, verbal convergent-integrative, graphic divergent-exploratory and graphic convergent-integrative dimensions were .92, .83, .51, and .41. For the paper & pencil version, Cronbach's alpha for the graphic divergent exploratory and convergent-integrative dimensions were .76 and .65. A second study consisted of four groups (Chinese children in Hong Kong (HK), Chinese children in Paris, French children in HK, and French children in Paris) of primary school students (total of 540 children in grades 1-6) used the electronic version. Inter-rater reliability of verbal convergent-integrative dimension for HK-Chinese group, HK-French, Paris-Chinese and Paris-French was reported as .99, .95, .95, and .92 respectively. Inter-rater reliability of the graphic convergent-integrative dimension for the HK-Chinese, HK-French, Paris- Chinese and Paris-French group was reported as .99, .71, .98, and .96 respectively.

Use: This assessment was developed in France and is now available in several languages including French, English, German, Turkish, and Arabic. This tool has been used as a monitoring tool to guide creativity development.

Strengths:

1. First tool among creativity assessments that combines an approach by domain of creative expression and by mode of thinking, instead of measuring a single component.

2. It offers a broader vision of creative potential in children.

3. Available in several languages.

Limitations:

1. A relatively new instrument.

2. Administration time is long for this instrument.

3. Need wider use to examine the reliability and validity of the instrument with a larger set of children.

4.10 Thinking Creatively in Action and Movement (TCAM) illustrates an assessment focused on imagination, creativity, and divergent thinking which are rarely measured in broad assessments.

Purpose: Designed to measure fluency, originality, and imagination in young children without having to use written/ verbal responses. It was developed based on 4 guidelines: 1) kinesthetic (not verbal) modality is the most appropriate for eliciting creativity, 2) preschool children require procedures for warm-up and motivation, 3) tasks for assessing creativity should be things pre-schoolers are familiar with, 4) the test should be easy to administer and score.

Age: Preschool- primary (ages 3-8)

Format and administration:

Consists of 4 activities:

- Activity 1: "How many ways?"- assesses fluency and originality in moving alternate ways across the floor
- Activity 2: "Can you move like?"- assesses imagination in moving like animals or a tree
- Activity 3: "What other ways?"- assesses fluency and originality in placing a paper cup in a waste basket
- Activity 4: "What might it be?"- assesses fluency and originality in generating alternate uses for a paper cup.

TCAM is administered individually. The examiner should record all responses (in movement, in words, or both) made by the child as completely and accurately as possible. Only one child should be in the activity room at a time and they should have enough space for movement. Before administering, warm-up activities should be done. Examiners are encouraged to participate with the child when instructions are given and during the introductory phase of each activity.

Scoring: Scoring guide is provided in the test manual. Activity 2 is scored for Imagination and the other 3 activities are scored for Fluency and Originality. Fluency scores are the number of relevant responses, and Originality scores range zero- three points for each response (they are based on comparing responses to the statistical frequencies of responses in the originality lists in the scoring guide). Imagination scores are based on a 5-point Likert scale ranging from "no movement" to "excellent; like the thing."

Developmental Domains Covered: The TCAM assesses motor, creativity, and cognition.

Time required: Administration takes 10-30 minutes (however no time limit should be imposed, the examiner should keep record of the time used).

Training and materials: Materials that are needed are: paper cups, wastebasket, pencils, red and yellow tapes.

Technical properties: Norms are based on 1,896 children ranging from ages 3-8 from 11 states and Guam. Inter-rater reliability is reported as coefficients between .90 and .99. Test-retest reliability is reported as .84 for a sample of 20 three-five year olds for a 2 week interval, and between .78 and .89 for a sample of

30, seven to eight year old boys with learning disabilities with a 1-14 day interval. Internal consistency was reported as .79. Significant positive correlations between TCAM and other creativity characteristics are reported. For example: correlations between the TCAM and production of various types of humour, between fluency scores and the Multidimensional Stimulus Fluency Measures, between TCAM and a modified Piaget measures of divergent thinking are reported. Scores on TCAM are showed only a low correlation to measures of intelligence. The TCAM results were not related to gender, socio-economic status, or race.

Use: This tool is used as a teaching tool. Teachers are more aware about the benefits of using creative movement in preschool and early elementary grades after using these tests.

Strengths:

1. This tool demonstrates acceptable reliability and validity.
2. This tool is easy to use for teachers.
3. The TCAM did not appear to be bias towards race, gender, community status, language/ culture.
4. Can examine abilities in young children and in children who are excluded from other testing instruments because of verbal restraints.

Limitations:

1. The TCAM has not been re-normed since 1981.
2. The originality lists associated with the TCAM have not been updated.
3. May not provide enough information about a child to make informed decisions or comparisons.
4. The assessment is designed as a teaching tool.

4.11 The Preschool Learning Behaviors Scale (PLBS; McDermott, Green, Francis, & Stott, 2000) **and Learning Behaviors Scale** (LBS; McDermott, Green, Francis, and Stott, 1999) illustrate relatively broad assessment of approaches to learning including motivation and executive functions.

Purpose: These scales were developed to examine the behaviours associated with learning.

Age: PLBS: Preschool age, 3-5; LBS: School age, 5-17

Format and Administration: The PBLS has 29 items each presenting a specific learning-related behaviour. The teacher indicates whether the behaviour *most often applies*, *sometimes applies*, or *doesn't apply*. The items are varied with positive and negative learning behaviours to reduce response sets. The item content between the two measures is very similar with the wording altered for PBLS to reflect less formal learning contexts. Teachers rate the student as accurately as possible and should rate all responses. Teachers should have seen the student in school for at least 6 school weeks or 30 days.

Developmental Domains Covered: This assessment has four subscales: Competence Motivation, Attitude Toward Learning, Attention/Persistence, Strategy/Flexibility. Content focuses on attentiveness, responses

to novelty and correction, observed problem solving strategy, flexibility, reflectivity, initiative, self-direction and cooperative learning.

Time Required: This scale takes teachers approximately 5 to 10 minutes per child to complete.

Training and Materials: There is no specific training involved. Materials needed include the rating scale.

Scoring: The evaluator calculates raw scores and converts them to percentiles. Students who obtain scores at or above the 40th percentile are displaying learning behaviours at or above the average range.

Technical properties: A factor analyses yielded distinct and reliable dimensions of competence motivation, attention/persistence, and attitude toward learning from several studies for both the PLBS and LBS across countries. However, in the U.S. it appears that the LBS presents a four factor structure. A normative sample (N=100) was configured based on the U.S. Census for PBLS. The normative sample for LBS was conducted with 1500 US students from 5 to 17 years old and was based on the 1992 U.S. Census. The assessment showed acceptable test-retest reliability and inter-rater reliability. In addition, the PLBS demonstrated expected correlations with the Social Skills Rating Scale (Gresham & Elliott, 1990) for concurrent validity evidence.

Uses: These scales are used in the US to examine children's specific learning-related behaviours. The PLBS has been translated to Spanish and tested in Peru. There is cross-cultural construct validity of the LBS as a measure of differential learning behaviours observed in school-aged children in Trinidad and Tobago. The tested dimensions of learning behaviours were found to be generalizable across age, gender and ethnicity.

Strengths:

1. Several studies supporting the validity of the instrument.
2. Assesses domains that are often neglected such as attitudes toward learning and persistence.
3. Used with several populations in various countries including US, Peru, Trinidad, and Tobago.

Limitations:

1. This is completed by a teacher who has spent significant amount of time with the child. Not all children attend school at an early age.
2. Standardization on US students only.
3. Narrow in focus by examining only learning behaviours.

4.12 Teacher Rating Scales of Early Academic Competence (TRS-EAC; Reid, Diperna, Missall, & Volpe, 2014) provides an example of a rating scale that combines measures of broad measures skills well beyond the academic sphere with measures of approaches to learning.

Purpose: A strengths-based measure to screen a wide array of skills, behaviours, and attitudes that are indicative of school success for preschool-aged children.

Age: Preschool aged-children, 3-5

Format and Administration: Includes two broad scales named the Early Academic Skills (39 items) and Early Academic Enablers (49 items). Teachers rate each child's current skill level compared with children of the same age.

Developmental Domains Covered: Early Academic Skills: Early literacy, early language, early mathematics, and early thinking. Early Academic Enablers: engagement, motivation, self-regulation, motor, interpersonal, and emotional competence.

Time Required: Time to complete the rating scales is not reported. With a total of 88 items and an estimated 15-20 seconds per question one can estimate that the total time on this assessment is less than 30 minutes per child.

Training and Materials: Teachers completing the rating scales would do best with a firm understanding of child development and appropriate age-level skills and behaviours.

Scoring: Teachers rate children on a Likert scale ranging from 1 (significantly below age expectations) to 5 (significantly above age expectations).

Technical properties**:** This assessment was evaluated with 440 preschool children from 38-70 months and completed by their teachers (N= 60). Most children, 62 percent, were Caucasian, 25 percent were Hispanic, 6 percent were African American, 1 percent Asian, and 6 percent classified as "other." All children were from lower socioeconomic backgrounds. Factor analysis supported a five-factor solution for Early Academic Skills Scale (Creative Thinking, Critical Thinking Skills, Numeracy, Early Literacy, and Comprehension) and a five-factor solution for the Early Academic Enablers Scale (Approaches to Learning, Social and Emotional Competence, Fine Motor Skills, Gross Motor Skills, and Communication). Experienced preschool teachers evaluated the rating scale for appropriateness and importance as an examination of content validity. Content validity ratios demonstrated acceptable levels of validity for the items.

To examine concurrent validity within two weeks of the teachers' rating of the children research staff individually administered achievement measures. Measures used for this correlation were the Woodcock-Johnson Tests of Achievement, 3rd edition (WJ-III; Woodcock, McGrew, & Mather, 2001), The Test of Early Reading Ability, 3rd edition (TERA-3; Reid, Hresko, & Hammill, and the Test of Early Math Ability, 3rd edition (TEMA-3; Ginsburg & Baroody, 2003). The TRS-EAC scales were associated with these direct measures. Factor scores from the rating scales were correlated with WJ-III Literacy and Math raw composite scores, the TERA-3 raw composite scores, and the TEMA-3 raw composite scores. Additionally, TRS-EAC scales were moderately predictive of subsequent performance for mathematics when the fall teacher ratings were correlated with the spring TEMA-3 raw composite scores for a small subsample of participants. Reliability of the scales was examined through the internal consistency of factors. Cronbach's Alpha ranged from .67-.98.

Uses: The TRS-EAC is used to assess the early academic competence for at-risk preschool populations.

Strengths:

1. This is a comprehensive assessment covering several domains including those not often measured in comprehensive assessments such as approaches to learning: engagement and motivation.

2. This is an easy to administer teacher rating scale.

3. The TRS-EAC is a strength-based measure rather than deficit-based.

Limitations:

1. The person completing the scale must be knowledgeable about appropriate age expectations in all of the domains.

2. The results are reported by teachers of preschool children which can limit the studied population to those that attend a preschool program.

3. This instrument does not appear to be available in multiple languages at this time.

4. This is a new assessment not widely used yet and more research with a wider population is needed.

5. It has a fairly narrow assessment age range. Available for only preschool-aged children.

4.13 Assessment of Peer Relations

Purpose: Designed to improve the peer-related social competence of young children. This assessment can be specifically of value to all children experiencing problems in establishing and maintaining successful and productive relationships with peers. Although for this assessment, peer relations are assessed in their school setting, both family and community factors are included in evaluation and intervention.

Age: 3-5 years old.

Format and administration: The assessment consists of three components. In the first section, one learns the general nature of the child's observed peer interactions in conjunction with an assessment of processes that allow for effective peer interactions to occur. Summary statements provide a bridge between assessment and intervention and special considerations such as possible developmental issues. This section consists of a series of scales to be observed while watching the child play, as well as written summaries and notes to determine developmental levels of the child. Scales are weighted rarely, sometimes, often, and almost always. At the end of the first section, assessors are asked to design interventions for the child to improve peer interactions. The second section involves observations of three social tasks important to young children (peer group entry, conflict resolution, and maintaining play). These three social tasks are evaluated using a checklist of behaviours to be observed and rated and surveys (with same scales as stated above). The purpose of this step is to evaluate how children think about a particular problem during interactions with peers. Next, an assessment is made of the child's ability to recognize specific social tasks and consistently and effectively perform those tasks over time. This is evaluated using charts that list concerns with emotional regulation, social cognitive processes and higher-order processes during play, and the child's different responses to these conflicts. The observer is to note how the child gains entry into a group of peers, how they resolve conflicts, and how they maintain play. Finally, a special considerations summary report related to the social tasks is provided.

Developmental Domains Covered: Communication, Problem solving, Personal-Social, Relationships with other children

Time required: Not reported, but appears to be a lengthy assessment.

Training and materials: Guide includes templates for observation.

Scoring: This assessment does not result in a numerical score. Rather, it is meant to be used as a tool for creating a specific intervention program for each child that is evaluated. It is also for the purpose of determining possible developmental disabilities. A "special considerations summary report" related to each child's social tasks is generated from the assessment.

Technical properties: No information found on the standardization of this evaluation tool.

Use: Used by educators and also for clinical use. It is meant for both administrators to think about complex factors that influence young children's peer relations, and intervention methods on how children can be helped in why they may be expressing difficulties in peer relations.

Strengths:

1. This assessment is meant to bring a clinical understanding and educational understanding to developmental issues.

2. The Assessment of Peer Relations gives a strong qualitative perspective of specific developmental issues with individual children.

3. Different factors such as family and community are considered in this measure besides classroom behaviours with this assessment.

Limitations:

1. No information available on standardization for this instrument.

2. This test does not give overall scores that could be used in comparisons.

4.14 Child Behavior Scale (*Excluded by Peers* Subscale) offers an example of a teacher rating scale solely focused on children's peer relationships.

Purpose: To identify children who experience exclusion by peers.

Age: 5 to 13 years old (most commonly), but appears that it could work with younger children.

Format and administration: Teachers rate students as 0=doesn't apply, 1=applies sometimes, and 2=certainly applies on the following seven items:

1. Peers refuse to let this child play with them.

2. Not chosen as a playmate by peers.

3. Peers avoid this child.

4. Is excluded from peers' activities

5. Is ignored by peers.

6. Not much liked by other children

7. Ridiculed by peers

Developmental Domains Covered: This is an assessment of peer relationships only.

Time required: Administered in 1-2 minutes per child.

Training and materials: This is a teacher report and no training is required.

Scoring: Lower scores on this scale indicate more positive peer relations. To score the scale, sum the items and divide by the number of responses. Because most children are generally accepted by peers, receiving a rating of 1 or 2 on just one or two of these items may raise concern.

Technical properties: This scale has been found to be valid and reliable for children ages 5 to 13.

Use: Used to identify children who are experiencing exclusion by peers.

Strengths of Measure:

1. This scale is quick and simple to administer.
2. It is easy to score and an easy way to identify possible areas of concern.
3. Scale is focused on an important area of concern for young children.

Limitations:

1. The scale has very few items.
2. Narrow in scope.
3. It assesses only problems not positive aspects of peer relationships.

4.15 Parenting Stress Index provides an illustration of a parent rating scale that assesses children's social-emotional development and relationship with the parent.

Purpose: The Parenting Stress Index is designed to be a screening and diagnostic measure to identify stressful aspects of parent-child interactions.

Age: Used for children 3 months to 12 years.

Format and administration: The assessment consists of 101 items with optional 19-item life stress scale. The short form has 36 items within 3 subscales: parental distress, parent-child dysfunctional interaction, and difficult child.

Developmental Domains Covered: The full version includes 6 child subscales (adaptability, acceptability, distractibility/hyperactivity, demandingness, mood, reinforces parent) and 7 parent subscales (competence, social isolation, attachment, parent health, role restriction, depression, relationship with spouse). There are also optional total stress scores and life stress scores.

Time required: The completion time for this index is 30-minutes for original and 10-minutes for short form.

Training and materials: Parents are to complete the assessment and no training is required.

Scoring: Total scores are calculated for each subscale.

Technical properties: Normed on several different samples including 534 parents of children in paediatric practice in Virginia, 191 low-income mothers in paediatric primary care clinics, and 223 Spanish-speaking mothers in NYC. Reliability for parents ranges from .55 to .80 and for children from .62 to .70. Test-retest reliability after 1 year was .70 for parent (.71 after 3 weeks) and .55 for child (.82 after 3 weeks). Low scores on the parent section correlate with parents having little investment in parenting or dysfunction in parent-child system.

Use: Useful in prevention and intervention programs, assessment of child abuse risk, and forensic evaluation for child custody.

Strengths:

1. This is simple and relatively quick to complete with no training required.

2. It is available in multiple languages: English, Dutch, Korean, Chinese, Portuguese, French Canadian, Italian, French, Icelandic, Japanese, Polish, Serbian, Swedish, and Greek.

3. There is a short version available.

Limitations:

1. May be difficult to get accurate information from parents who are defensive or have dysfunctional relationships with children.

2. Assesses a specific area of concern.

4.16 The Student Teacher Relationship Scale (STRS) is an example of a teacher rating scale for the teacher-child relationship.

Purpose: This was designed to evaluate teachers' feelings and beliefs about individual student's actions toward them, based on teacher perceptions of the teacher-child relationship.

Age: Appropriate for preschool to grade 3.

Format and administration: Using a 5-point Likert-type scale that ranges from 1 = definitely does not apply to 5 = definitely applies, teachers rated how applicable each statement is to their current relationship with a particular child. Three subscales are included in the measure. The Conflict subscale taps the extent to which the teacher–child relationship is marked by antagonistic, disharmonious interactions (e.g., "This child and I always seem to be struggling with each other"). The Closeness subscale is an index of the amount of warmth and open communication present in the relationship (e.g., "I share an affectionate, warm relationship with this child"). The overall quality of the relationship is determined by the amount of closeness and conflict (reflected) in the relationship. The Dependency subscale measures the degree to which a teacher perceives a particular student as overly dependent on him/her. High dependency scores

suggest that the student reacts strongly to separation from the teacher, requests help when not needed, and consequently the teacher is concerned about the student's overreliance. Higher scores indicate more positive, higher quality teacher–child relationships. The items are based on attachment theory and the Attachment Q-Set (Waters & Deane, 1985).

Developmental Domains Covered: The full version includes 3 subscales – conflict (12 items), closeness (11 items), and dependency (5 items). The short form comprises 15 items that measure 2 dimensions of teacher-child relationships: Closeness and Conflict.

Time required: Time to complete this is 5 to 10-minutes for full version and 2-minutes for short form.

Training and materials: Teachers are to complete the assessment and no training is required.

Scoring: Each item is scored from 1 to 5. High total scores suggest higher teacher-child relationship quality, and specifically, a relative lack of conflict, lower dependency, and higher closeness.

Technical properties: The STRS was normed on a sample of more than 1500 students (and 275 teachers) that matched the 1990 US census data in terms of race/ethnicity and also reflected a wide range of socioeconomic status. It has also been shown to be psychometrically reliable and valid. Test-retest correlations over a 4-week period were .88 for closeness, .92 for conflict, and .76 for dependency. Validity studies indicate that the STRS correlates in predictable ways with concurrent measures of academic skills and performance on standardized tests (Hamre & Pianta, 2001).

Use: Primarily used as a tool for assessing student-teacher relationships in the context of efforts to prevent or to intervene early in the course of development of adjustment problems in school. The STRS can also be used in educational assessment batteries to determine the extent to which relationship problems or strengths should be addressed in program planning, and it can be used as a tool for researching classroom processes.

Strengths:

1. The STRS has been widely used in studies with preschool and elementary school children. It is associated with children's and teachers' classroom behaviours and correlates with observational measures of quality of the teacher–child relationship (e.g., Birch & Ladd, 1997; Howes & Hamilton, 1992; Howes & Ritchie, 1999).

2. STRS scores correlate with Attachment Q-Set ratings of teachers and students such that higher STRS scores are associated with more secure relationships (Howes & Ritchie, 1999).

3. This scale can be used with a preschool to grade 3 age range.

Limitations:

Only teacher perceptions are relied upon and children's perceptions are not considered.

4.17 Early Years Foundation Stage Profile (EYFSP)

Purpose: The EYFSP was developed to inform parents about their child's development against the early learning guidelines and the characteristics of their learning, to support a smooth transition to key stage 1

by informing the professional discussion between EYFSP and key stage 1 teachers, and to help year 1 teachers plan an effective, responsive and appropriate curriculum that will meet the needs of all children.

Age: This assessment offers a two-year-old "check" between the ages of two and three and the EYFS profile is completed by the end of the year in which the child reaches age five.

Format and administration: The EYFSP profile summarizes and describes children's attainment at the end of the EYFS. Practitioners' assessments are primarily based on observing a child's daily activities and events. The assessor notes the learning which a child demonstrates spontaneously, independently and consistently in a range of contexts. Accurate assessment takes into account the perspectives of the child, parents and other adults who have significant interactions with the child.

Developmental Domains Covered: The EYFSP assesses 17 early learning goals in six areas of learning. **Communication and language**: listening and attention, understanding, speaking; **Physical development**: moving and handling, health and self-care; **Personal, social and emotional development**: self-confidence and self-awareness, managing feelings and behaviour, making relationships; **Literacy**: reading, writing; **Mathematics**: numbers, shape, space and measures; **Understanding the world**: people and communities, the world, technology; **Expressive arts and design:** exploring and using media and materials, being imaginative.

The measure also examines the child's three learning characteristics: Playing and exploring- engagement; Active learning- motivation; Creating and thinking critically- thinking.

Time required: The profile is completed over time after observations of the child in an ongoing process.

Training and materials: The Local Authority is responsible for training and supporting the teachers/practitioners. They provide support and guidance for all teachers/practitioners in making accurate assessments of children's achievements and progress through a range of strategies grounded in observations over time. It is unclear how much training is required to administer the profile.

Scoring: First, the report includes the child's attainment in relation to the 17 ELG descriptors. These are scored on a nine point scale. The first three points of each scale describe a child who is still progressing towards the achievements described in the early learning goals. The next five points are from the early learning goals themselves. They are not necessarily in hierarchical order and a child may achieve a later point without achieving some of the earlier points. The final point in each scale describes a child who has achieved all points one through eight and has developed further and is consistently working beyond the level of the early learning goals. These scores are categorized into emerging (1-3), expected (4-7) and exceeding (8-9).

Second, a short narrative describing the child's three characteristics of effective learning is generated by the assessor.

Standardization and psychometrics: Teachers are moderated in their use of this instrument. There are moderators that visit schools to sample students. The moderator secures consistency and accuracy of judgments made by teachers and assures that the setting has achieved an acceptable level of accuracy and validity. The moderator does this by evaluating several profiles to establish if the practitioner has understood what constitutes an appropriate outcome and judgment.

Analysis of data from the EYFSP indicated that six scales provide reliable measures of underlying skills. The simplest factor to measure uniformly is the Literacy factor. The least clear factor is Physical Development. The different scales appear to tap quite similar things as demonstrated by high correlations among domains. However, this may reflect how teachers make generalizations about pupils across

domains and has been documented in other similar assessments. It was also reported that the EYFSP correlated with other language measures and was predictive of later achievement (Snowling, Hulme, Bailey, Stothard, and Lindsay, 2010).

Use: The EYFSP is used to inform parents of children's progress, to inform instruction in school, and to report children's progress nationally in England.

Strengths:

1. The EYFSP is being redeveloped to become more quantitative than qualitative (this new measure is not yet available).

2. The EYFSP is very comprehensive as it examines all key learning domains.

3. Although results are not reported on the moderation of the instrument, a strong program of moderating the use of the instrument is in place.

4. Used widely in England.

Limitations:

1. The EYFSP is being redeveloped to be more quantitative which could potentially compromise some domains that are currently evaluated (i.e., creativity).

2. The moderation of the instrument is likely an expensive endeavour.

3. The EYFSP is not currently used outside of England.

5. Conclusions and Recommendations

55. The assessments available offer many choices for measuring children's physical, social, emotional, linguistic, and cognitive development with respect to age, mode of assessment, the source or respondent, and burdens on respondents. There are fewer choices for assessments of executive functions and for some cognitive measures in the areas of math and science. Very few options are available for assessing development in the arts and culture and for approaches to learning; this is primarily done through performance assessments including clinical interviews (conversations and storytelling would be included here). Measures we reviewed that addressed aspects of approaches to learning including the specific topics of curiosity, creativity, critical thinking, and problem solving. None of the assessments we reviewed measured self-esteem, self-efficacy, values and respect, or subjective states of wellbeing such as happiness. We did not identify any comprehensive assessments for young children that addressed these domains.

56. For those domains that are measured rarely or not all by comprehensive assessments, specific assessments are sometimes available. Most often these are rating scales (except for executive functions) completed by adults. Specific measures of wellbeing identified include those that assess relationships with parents and peers, and engagement and participation in ECEC. We did not identify measures of general happiness and satisfaction specific to young children, but these could be constructed. In our opinion, general wellbeing and measures of children's rights to engagement and participation in decision making would be most readily assessed through clinical interviews or time diaries (with the latter requiring inferences from activities about quality of life as indicated by the engagement of children's capacities).

57. Clearly, some assessments have stronger evidence of technical adequacy than others. Concerns with technical adequacy are greatest for performance assessments and ratings, particularly in the domains that are not well-covered by tests. The technical adequacy of performance assessments can be improved by standardization of assessment procedures and training of assessors. This has costs, of course.

58. Given the available assessments, the most efficient strategy to selection of instruments for an international study would appear to be choosing one very broad assessment to be supplemented by a small number of highly specific assessments in domains that often are neglected. However, it would be possible to construct a broad assessment that is carefully tailored based on judgments regarding the best choice in each domain as was done with ZCAT. To fully cover all of the domains of interest to OECD representatives, some instrument construction may be necessary. Instrument adaptations will be required for language and culture. Given the extent of these adaptations, international pilot-testing to evaluate performance is recommended before use in a full scale international study.

5.1 Age

59. As the validity and reliability of assessment, and children's abilities to actively contribute to the assessment increase with age, the quality of the information obtained will be improved by conducting the assessment at age 3 or later. Prior to age 3, the assessments are primarily reports by adults. One then must choose whether to assess children at a particular age or at an educational transition such as entry to preschool or primary school. As entry to preschool can be well before age 3, this suggests entry to primary school as possible assessment point. However, this seems to us somewhat artificial, as what constitutes primary school in one country constitutes preschool in another. For this reason, we would recommend a uniform age across countries, perhaps age 4. One could also consider age 5 if by this age universal or near universal participation in ECEC (or primary education) has been achieved in the relevant country or countries, so that they can be assessed outside the home (including ratings by teachers). A final consideration is whether policy makers and others want just a point in time measure or wish to know how children are developing over time during the pre-primary years; in the latter case it will be necessary to administer comparable assessments at more than one age.

5.2 Final thoughts on decision making for national and international assessments

60. As pointed out previously, what, how and when children are assessed depends on the purpose or purposes of the assessment, judgments about what is important, and budgets. It is also limited by what is currently available, and some aspects of LDWB will require investment in assessment development if these are to be assessed on a large scale at an affordable cost. Whatever approach is taken, it would be wise to invest in some development, adaptation, and piloting before large scale use. Given the limitations of existing instruments, the most feasible course in the near future may be to administer a broad measure that addresses the domains identified as most important. This could be a single existing measure or a composite of existing measures. If some aspects of the measure are more costly (in time as well as money) to employ then it might be possible to administer those only to samples of the population or subsamples of a larger sample (matrix sampling in which only some items are administered to each child and then these are aggregated is another possibility, but it limits usefulness for teachers and policy studies). Lastly, we remind the reader that this paper reviews types and exemplars, and not all the available assessments. When selecting specific assessments for specific purposes, policy makers can consult experts in the relevant country or countries as well as the existing compendia or early childhood assessments.

REFERENCES

Atkins-Burnett, S. (2007). *Measuring children's progress from preschool through third grade* (No. 5687). Plainsboro, NJ: Mathematica Policy Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barbot, B., Besançon, M. & Lubart, T. (2011). Assessing creativity in the classroom. The Open Education Journal 4, 58-66.

Barnett, W. S., & Boyce, G. C. (1995). Effects of children with Down syndrome on parents' activities. *American journal of mental retardation: AJMR*, *100*(2), 115-127.

Benítez, I., & Padilla, J. L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*(1), 52-68.

Berry, D.J., Bridges, L.J., & Zaslow, M.J. (n.d.). Early childhood measures profiles. Washington, DC: Child Trends.

Borg, W. R., Gall, M. D., & Gall, J. P. (1989). *Educational research: an introduction* (5th ed.). New York: Longman.

Creswell, J. W. (2008). Educational research: Planning, conducting, and evaluating quantitative and qualitative research. (3rd ed.). Saddle River, NJ: Pearson.

Dunn, L. & Dunn, L. (2006). Peabody Picture Vocabulary Test, Fourth Edition (PPVT -1V). Bloomington,MN: NCS Pearson.

Dunphy, E. (2008). Supporting early learning and development through formative assessment: a research paper. Dublin: National Council for Curriculum and Assessment.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(3-4), 199-215.

Ettling, D., J. T. Phiri, et al. (2006). Child Development Assessment in Zambia: A study of developmental norms of Zambian children aged 0-72 months Lusaka, Zambia, Ministry of Education, Republic of Zambia.

Fernald, L. C., Kariger, P., Engle, P., & Raikes, A. (2009). *Examining early child development in low-income countries.* Washington DC: The World Bank.

Fink. G., Matafwali, B., Moucheraud, C., & Zuilkowski, S. S. (2012). The Zambian Early Childhood Development Project 2010 Assessment Final Report. Cambridge: Harvard University.

Frongillo, E. A., Tofail, F., Hamadani, J. D., Warren, A. M., & Mehrin, S. F. (2014). Measures and indicators for assessing impact of interventions integrating nutrition, health, and early childhood development. *Annals of the New York Academy of Sciences*, *1308*(1), 68-88.

Glascoe, F.P. (2002). The brigance infant and toddler screen: Standardization and validation. *Journal of Developmental & Behavioral Pediatrics* 23, 145-150.

Godfrey, J. R., & Galloway, A. (2004). Assessing early literacy and numeracy skills among Indigenous children with the Performance Indicators in Primary Schools test. *Issues in Educational Research*, *14*(2), 144-155.

Hamilton, S. (2006). Screening for developmental delay: Reliable, easy-to-use tools. *Journal of family practice* 55, 415.

Hofferth, S. L., & Sandberg, J. F. (2001). How American children spend their time. *Journal of Marriage and Family*, *63*(2), 295-308. Kim, D. H., & Smith, J. D. (2010). Evaluation of two observational assessment systems for children's development and learning. NHSA Dialog, 13, 253-267.

Korkman, M., U. Kirk, et al. (1998). NEPSY: A developmental neuropsychological assessment. San Antonio, TX, The Psychological Corporation.

Lau, Sing, et al. (2013). Bicultural effects on the creative potential of Chinese and French children. *Creativity Research Journal* 25, 109-118.

Matafwali, B. (2010). The relationship between oral language and early literacy development: Case of Zambian languages and English. Ph.D. Dissertation in progress. Lusaka, University of Zambia.

Melton GB. Young children's rights. In: Tremblay RE, Boivin M, Peters RDeV, eds. *Encyclopedia on Early Childhood Development* [online]. Montreal, Quebec: Centre of Excellence for Early Childhood Development and Strategic Knowledge Cluster on Early Child Development; 2011:1-8. Available at: http://www.child-encyclopedia.com/documents/MeltonANGxp1.pdf.

Merrell, C., & Tymms, P. B. (2001). Inattention, hyperactivity and impulsiveness: their impact on academic achievement and progress. *British Journal of Educational Psychology*, *71*(1), 43-56.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Rossbach, H. G. (1988). Daily Routines of Young Children. Paper presented at the Annual Meeting of the American Educational Research Association.

Sheridan, S., & Pramling Samuelsson, I. (2001). Children's conceptions of participation and influence in pre-school: A perspective on pedagogical quality. *Contemporary Issues in Early Childhood*, *2*(2), 169-194.

Slentz, K.L., Early, D., & McKenna, M. (2008). A guide to assessment in early childhood: Infancy to age 8. Washington State Office of Superintendent of Public Instruction.

Snow, C. E., & Van Hemel, S. B. (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.

Snowling, M. J., Hulme, C., Bailey, A. M., Stothard, S. E. & Lindsay, G. (2011). Better Communication Research Project: Language and Literacy Attainment of Pupils during Early Years and through KS2: Does Teacher Assessment at Five provide a Valid Measure of Children's Current and Future Educational Attainments? (DFE-RR172A). London: Department for Education. Available http://dera.ioe.ac.uk/13689/1/DFE-RR172a.pdf

Standards and Testing Agency (2013). 2014 Early Years Foundation Stage Profile Handbook. London: Department of Education, Standards and Testing Agency.

Stevens, P. A., & Dworkin, A. G. (Eds.). (2014). *The Palgrave Handbook of Race and Ethnic Inequalities in Education*. Palgrave Macmillan.

Teaching Strategies (2013). Teaching Strategies GOLD Assessment System: Technical Summary. Washington, DC: Author.

Tinajero, A.R., & Loizillon, A. (2012). Early childhood development and wellbeing. The review of care, education, and child development indicators in early childhood. Paris: OECD.

William, F., & Monge, P. (2001). *Reasoning with statistics: How to read quantitative research.* (5th ed.). Belmont, CA: Thomson Higher Education.

Woodhead, M., & Brooker, L. (2008). A sense of belonging. Early Childhood Matters No. 111, 3-17. The Hague, The Netherlands: Bernard van Leer Foundation

ANNEX – MATRICES OF ASSESSMENTS

Matrix A – Assessment Descriptives

	Assessment - Author and Publication Date	Ages	Type of Assessment	Language	Countries Used in	Who administers
Assessor-report	Zambian Child Assessment Test (ZCAT) - Fink, Matafwali, Moucheraud, & Zuilkowski, 2010	Preschool	1-1 test	Nyanja, Bemba, Tonga, Lozi, English	Zambia; various subtests used in multiple countires	Trained assessors
	Battelle Developmental Inventory (BDI) - J. Newborg, J.R. Stock, J. Wnek, J. Guidubaldi, and J.S. Svinicki, 1988	Birth to 7 years 11 months	1-1 test; also parent and teacher interview items			Parents or Examiners
	NIH Measures - NIH Blue print of Neuroscience Research. Principal Investigator: Dr. Richard Gershon, 2004	3-85 years	Proctored; self-administered; computer-administered	English and Spanish	US , Colombia	Trained assessors
	Brigance Early Childhood Screens (BECS) - Albert H. Brigance, 1999	Birth to 68 months	Child observation and performance; also parent interview items		US	Professionals with child development knowledge
	Denver II - 1990	Birth to 6 years	Direct observation; also parent observation	English and Spanish		Professionals or para-professionals
	Griffiths Mental Development Scales Extended revised (GMDSER) - 2006	2 - 8 years	1-1 test	English	UK	Pediatricians and health professionals

	Assessment - Author and Publication Date	Ages	Type of Assessment	Language	Countries Used in	Who administers
	Mullen Scales of Early Learning (MSEL)- 1995	Birth to 68 months	1-1 test		US	"highly trained" professionals
	Schedule of Growing Skills (SGS) - 1996	Birth to 5 years	1-1 test		UK, England	Trained assessors
	Hong Kong Early Childhood Development Scale (HKECDL) - Nirmala Rao, Sun Jin, Sharon Ng, Kitty Ma, YvonneBecher, Diana Lee, Carrie Lau, Dr. CB Chow, & Patrick Opper (1992, 1996)	3-6 years	1-1 test	English, Cantonese, Chinese	China (Hong Kong)	Trained assessors
	Early Years Foundation Stage Profile (EYFSP) – Snowling, Hulme, Bailey, Stothard, & Lindsay, 2011	2-5 years	Observation based assessment	English	England	Trained Teachers
	Early Learning System (ELS) - Riley-Ayers, Stevenson-Garcia, Frede, and Brenneman 2012; Riley-Ayers, Stevenson-Garcia, Brenneman, Thompson, & Thompson, 2014; Developed at NIEER	3-6 years	Authentic observation-based assessment	English	US, China	Teachers
	International Performance Indicators in Primary Schools (iPIPS) - Peter Tymms and Colleagues: http://www.ipips.org/the-team Note: could be listed as direct assessment for cognitive domains.	4-7 years (first year of school)	1-1 test with teacher rating and supplemental parent report	Dutch, German, Russian, Spanish, French, Slovenian, Chinese, Afrikaans, Sepedi	Australia, Netherlands, Scotland, New Zealand, Abu Dhabi, Germany, South Africa	Trained teachers
	Work Sampling System (WSS) - Meisels, Jablon, Dichtelmiller, Dorfman, & Marsden, 1998	3yrs to sixth grade	Checklists used 1-1 or in group setting	English	US	Teachers

	Assessment - Author and Publication Date	Ages	Type of Assessment	Language	Countries Used in	Who administers
	Teaching Strategies GOLD - Teaching Strategies, 2010	Birth through Kindergarten	Authentic observation-based assessment	English, Spanish	US; Other countries, but do not have details about which or how many---have reached out to a contact at GOLD to inquire.	Teachers
	High Scope Child Observation Record (CORE)- High Scope, 2013	Birth through Kindergarten	Authentic observation-based assessment	English, Spanish	US, Canada, Chile, Indonesia, Ireland, Korea, Mexico, The Netherlands, Portugal, South Africa, UK	Teachers
	Early Development Instrument (EDI) - 1998	4-7 years	Questionaire done by teachers or parents.	English, French	Canada , US, Australia, Chile, England, Holland, Egypt, Mexico, Jamaica,	Teachers, Early childhood educators
	Kindergarten Entrance Inventory for Connecticut (KEI – Connecticut)	K	Observation-based assessment	English	US	Teachers
Parent-report	Child Development Inventory (CDI) - Harold Ireton, 1992	15 months to 6 years	Parent report with professional assistance	English and French	US, France, Canada	Parents

	Assessment - Author and Publication Date	Ages	Type of Assessment	Language	Countries Used in	Who administers
	Ages & Stages Questionnaire (ASQ) - Bricker, D., and Squire, J., 1999	1 month to 66 mos.	Parent report	English, Spanish, French, Korean	France, Norway, Finland, Spain, Netherlands, Turkey, North America, South America, Asia, Australia	Parents
	Parents' Evaluation of Developmental Status (PEDS) - 1997	0 to 8 years old	Parent report	English, Spanish, Vietnamese, Hmong, Somali, Chinese, Malaysian	US, Australia, Great Britain, England	Parents

Matrix B – Assessment Details

	Assessment	Administration time	Cost per administration	Assessment format (electronic/paper & pencil)	Strengths	Weaknesses
Assessor-report	ZCAT	1 hour	?	Paper and pencil	Comprehensive, culturally specific	Evidence of validity limited to original assessments; long administration
	BDI	45 - 90 minutes	$312.50	1. Direct assessment with toys, games, tasks 2. Observation 3. Parent report	It covers all the developmental domains, and adaptions are available for children with disabilities.	Training is required to administer, measure takes 1.5 hours, not standardized for use in the UK, caution is required for using with kids not from US, does not involved parents, evidence is lacking about acceptability by parents, relatively costly.
	NIH	9 tests + 2 supplemental tests. Time ranges from 2-7 mins for each test.	No cost for assessment. Fees apply for user & tech support: >100 subjects = $1500 or <100 subjects = $5000 covered until 12/31/2014 if non NIH support. If Study is NIH supported then the fees are paid as long as it is funded by NIH.	computerized	Easily incorporates multiple areas of neurological function; Inexpensive, no royalties, low per subject costs.	
	BECS	10-15 mins	$149	paper/pencil; computer scoring available	Covers the developmental domains of interest, quick and easy to use, has acceptable psychometric properties, flexibility with administration.	More focused on academic performance, used in educational settings and not in health settings, has not been standardized in the UK; primarily used as screener
	Denver II	20-30 mins	$90 for English packet, $120 for Spanish packet		Covers the developmental domains, is well known and widely used, is reported to have good sensitivity.	Training is required, meticulous administration, has poor specificity, the measurement was standardized in 1980 so the norms are outdated and specific to US; used primarily as developmental screener.
	GMDSER	50-60 mins		Kit: building blocks, drawing book, record forms	Covers the developmental domains but personal-social rather than social-emotional, widely used, acceptable sensitivity and specificity, standardized recently on a UK population.	Intensive training required, no evidence for use as a population measure, no evidence about acceptability by parents, lengthy to administer, little published evidence on validity.

	Assessment	Administration time	Cost per administration	Assessment format (electronic/paper & pencil)	Strengths	Weaknesses
	MSEL	+/- 30 mins		Electronic and paper/ pencil	Covers the developmental domains and is relatively easy to score.	Professionals need to be highly trained, not a population measure, standardized 30 years ago, excludes children with disabilities, no UK norms, not used in general child developmental assessment, unknown how parents feel about it, parents are not involved, costly.
	SGS				Covers the developmental domains, widely used, the original specificity is good, completion time is relatively short.	The original estimates of sensitivity range from poor to good depending on the domain, the validity was carried out over 30 years ago so is now outdated, although SGS is being used in the Flying Start program, there is no published evidence of this, does not involve parents, no information available on acceptability by parents or professionals, little use in reviewed journals
	HKECDS	2 sessions of 30-45 mins	No info	Direct assessment: verbal and physical actions.	Considers both holistic development and current expectation of early child development in Hong Kong.	No representative sample of Prek school children in HK; Recruited different school children from different PreKs and family background of 2 pilots' sample size were small.
Teacher-report	EYFSP	Ongoing over the course of the school year		Observational; paper/pencil	Currently being redeveloped to become more quantitative; comprehensive in examination of all key learning domains; used widely in England	Moderation of the instrument is expensive; not currently used outside of England.
	ELS	Ongoing; Generally with 3 score periods per school year.	$250.00 for print kit for 25 children; Online price dependent on the number of children assessed: range as high as 21.95 plus one time set up fee ($450-$900)	Paper and Pencil and Online	Measure that examines child development across several domains through a manageable assessment system for teachers. Provides valuable information for teachers to inform instuction. Provides a developmental trajectory of children's development.	ELS does not examine development of children in the arts; It is not used wide-scale at this time.
	iPIPS (only part is teacher report)	20-minutes for direct assessment (cognitive) portion	In England PIPS is £80 per school plus £3.10 per pupils	computerized for the objective part and booklet with app	Teachers and students reported to enjoy the computer delivery; program is easy to use; has been used wide-scale in several countries and in several languages for 20 years.	Technology and IT support may be necessary for some users; psychometric properties extrapolated from PIPS to iPIPS; vocabulary and phonological awareness scales have proved most challenging in terms of generating equivalent versions for the different languages and cultures.
	WSS	Ongoing	$75 for basic Teacher Reference Pack and goes up from there	Paper and pencil	Comprehensive; charts children's progress overtime; gives insight for individualizing instruction	WSS may vary depending on teacher experience and training; sample used for reliability and validity was not representative of US population

	Assessment	Administration time	Cost per administration	Assessment format (electronic/paper & pencil)	Strengths	Weaknesses
	GOLD	Ongoing; Generally with 3 score periods per school year.	$199.00 for print kit for 25 children; Online price dependent on the number of children assessed: range as high as 21.95 plus one time set up fee ($450-$900)	Paper and Pencil and Online	Comprehensive; developmental trajectory of learning and development; gives insight for individualizing instruction	May vary depending of teacher experience and training (teacher bias evident in one study); Large number of items to collect data on and evaluate; Minimal concurrent/ construct validity with direct assessments of language, cognition, and self-regulation.
	CORE	Ongoing; Generally three score periods per school year	$225 for kit for 24 children paper and pencil administration	Paper and pencil and Online	Comprehensive; developmental trajectory of learning and development; gives insight for individualizing instruction	May vary depending on teacher experience and training; no normative data
	EDI	15 mins per child. Administered by teacher in 2nd half of the year.	No info	Electronic (e-EDI) or paper/pencil	Emphasizes community aspects of readiness; can be adapted for different locations; considers both teacher and parent observation; predicts reading and writing outcomes in later years	Limited use in the U.S.; narrow age applications; cultural modifications need to be arranged with a professional assessor; limited agreement with other instruments with similar purpose; results reported at population level only, not for individual children
	KEI - CONNECTICUT	Ongoing; generally observations at beginning and end of the kindergarten year.		Paper and pencil	Evaluates children comprehensively across several domains; minimal training and materials required.	Only used small-scale in the US, only appropriate for children entering kindergarten.
Parent-report	ASQ	10-15 mins	$190	Paper and pencil/ electronic	Covers developmental domains and produces scores for each, can be a population measure, flexibility to administer, gets parents involved, provides a good basis for discussion, can be used with children at high risk of developmental problems, quick and easy, one-off purchase	Lack of standardized norms for the UK population, focuses on personal-social instead of social-emotional, lack of info about UK parents accepting the measure, need further analysis to see if it can be used with parents with language barriers and literacy problems, age specific questionnaires, and some of the language is "Americanised"; used primarily as screener
	PEDS	5 mins and 2 mins to score	$60 for 100 survey forms	Interview or paper and pencil or electronic	Covers the developmental domains, used in population surveys, encourages parent involvement, flexible format, quick and easy, requires minimal training, low cost, and can be used among children at risk of developmental problems	May be hard to track small changes overtime, hard to discriminate in between children, less useful as a population monitoring measure, may be subjective, it assumes parents' knowledge about development; used primarily as screener.
	CDI	30 - 50 mins to complete and 10 mins to score	Manual $30; Booklet $10-$15; Answer sheets $10; Profiles $10	Paper and pencil	Covers the developmental domains of interest, low costs involved, parents find it easy to complete, has been used with high-risk kids.	It is not a assessment to find developmental delay, but for kids who may have one, the standardization was over 30 years ago and with a largely white US sample, there are no UK norms for this measure, authors state it is not appropriate for parents with less than a high school degree; used primarily as screener.

Matrix C – Assessment Psychometrics

	Assessment	Purpose of Assessment (summative/formative)	Purpose of assessment (clinical/screening)	Normative data	Reliability	Validity	Use of Scores
Assessor-report	ZCAT	Summative	Research tool	Administered 1686 Zambian preschool children in 2010	Cronbach's alpha	Validity for individual assessments but not for full adapted battery	Census
	BDI	Summative	Depicts child progress in intervention programs; identifies children with special needs; provides comprehensive analysis of functional capabilities.	Normative data were gathered from 2,500 children (closely resembling the 2000 US census) btwn the ages of birth to 7yrs 11mos	Test-retest r = .71-1.0	Moderate correlations with other established tests; concurrent validity .566 with PPVT-R; .66 with Preschool Language Scale	Comprehensive test used by professionals to assess the development of children. Has been used among children with autism, developmental delays, motor delays, speech and language delays and prematurity.
	NIH	Summative	Research/clinical/longitudinal/epidemiological	Administered to 4,859 participants 3-85 y.o. Normative data available for ages 3-17, 19-29,30-39,40-49,50-59,60-69, 70-79, 80-85.	Strong	16,000 subjects for response theory approach. 450-500 subjects compared with Gold standard measures where available.	Used for variety of settings by researchers and clinicians with emphasis on measuring outcomes for longitudinal, epidemiologic studies + prevention and intervention trials.
	BECS	Summative	Screening to identify weaknesses, assess readiness for school; identify interventions needed	Sensitivity 82%, specificity 84%. Also identifies 86% of children under the age of 2 with potential giftedness.		Comparison with a battery of age-appropriate developmental assessment tools such as Bayley's.	Identifies potential learning delays or academic giftedness

	Assessment	Purpose of Assessment (summative/formative)	Purpose of assessment (clinical/screening)	Normative data	Reliability	Validity	Use of Scores
	Denver II	Summative		The authors present norms based on the 1980 US census, which is criticized.		Authors made no attempt to measure the validity of the tool.	To detect developmental delay. The Denver II is widely used in clinical settings and as the gold standard which other measures are compared.
	GMDSER	Summative	Clinical and research purposes	The measure was normed on a national representative sample of children in UK.			To measure the rate of development of young children, for clinical and research purposes.
	MSEL	Summative		Normative sample as based off of a US sample of 1849 children over 8 years.		Little evidence found for validity.	Now commonly used as a measure of cognitive language skills. Used in research, clinical evaluations, and longitudinal evaluation of children with autism.
	SGS	Summative				Validity of the original tool was assessed by comparison with the Griffiths test with the NCES tool.	To establish children's developmental level: widely used to assess children as they enter the program Welsh Flying Start and when they leave at 3 years of age.
	HKECDS	Summative	Developmental readiness.	Sample in Hong Kong is largely comparable with the Canadian normative references in the physical, social and emotional domains, but higher in the language/cognitive and communication/general knowledge domains.	The test-retest reliability of the CEDI after a four-week interval was analyzed in 30 participants using the kappa statistic (k). The kappa coefficient was 0.89, thus demonstrating the instrument's stability over time.		
Teacher-report	EYFSP	Formative	To inform parents about their child's development against the early learning guidelines and the characteristics of their learning.	Teachers are moderated in their use of the instrument. Moderators visit schools to sample students and evaluate several profiles to establish whether practitioners understand what constitutes an appropriate outcome and judgment.	Six scales have proven to be reliable measures of underlying skills. The simplest factor to measure is literacy and the least clear to measure is physical development.	The EYFSP has been correlated with other language measures and is predictive of later achievement.	Scores are used to inform parents of child progress, inform school instruction, and report child progress nationally in England.
	ELS	Formative	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning	This assessment is not normed.	High inter-rater reliability.	Moderate correlations with other established tests. High internal consistency.	Inform instruction and track children's progress over time

	Assessment	Purpose of Assessment (summative/formative)	Purpose of assessment (clinical/screening)	Normative data	Reliability	Validity	Use of Scores
	iPIPS (only part is teacher report)	Formative and Summative	Assess children's abilities and development at entry to school and assess progress during the first year of school.	The PIP baseline assessment has been standardized to have a mean of 50 and a standard deviation of 10. The assessment provides conversion charts that offers age corrected standardized test scores.	Test-retest reliability for 29 students was .98 for the full instrument; subtests ranged from .34 to .99.	Predictive validity has been demonstrated through correlations ranging from .48 to .66 on assessments given up to 6 years later.	Used to assess children at the start of school and after one year of schooling within and across countries.
	WSS	Formative	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning	Scores provided for diverse, but not standardized sample. Norms based off of 5 public schools in Pittsburgh over 3 years	Internal consistencies range from .87 to .94; high inter-rater reliability	High correlation between WSS and WJ-R subscales	Inform instruction and track children's progress over time
	GOLD	Formative	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning	Nationally representative norm sample of 18,000 children from 50 states, PR, and DC; across age cohorts; Provides norm tables across all six areas of development. Each norm table includes expected scores for children across 24 different 3-month age bands from 0-71 months. Includes norms for fall, winter, and spring.	Strong internal consistency; Inter-rater reliability for kindergarten teachers seems weak;	Varying validity based on study report and age of children; seems to be most valid for math and literacy; low concurrent and construct validity with direct measures of language, self regulation, and cognition.	Inform instruction and track children's progress over time; Statewide progress assessed at kindergarten entry in US
	CORE	Formative	Track child's efforts, achievements, and progress; designed to enhance instruction and improve learning	None?	85.7% agreement for inter-rater reliability for 70 teachers; acceptable internal consistency	Demonstrated significant differences in scores at age category and moderate to high correlations with other standardized measures.	Inform instruction and track children's progress over time
	EDI	Summative	Assesses children's level of development during the 1st year of school and readiness to learn.	116,860 4-5 year old children from Canada + several countries. Data available @Offord Centre website.	Internal: .84-.96; Inter-rater: High between teachers and early childhood professionals, but moderate between these and parents.	International comparison of normative data is valid.	To determine the % of children at various levels of readiness to benefit from school; also plan interventions and resource investments

	Assessment	Purpose of Assessment (summative/formative)	Purpose of assessment (clinical/screening)	Normative data	Reliability	Validity	Use of Scores
	KEI - CONNECTICUT	Formative	Gives a statewide snapshot of the skills and behaviors students demonstrate			Validity evaluated by comparing the content to state preschool framework and curriculum was reviewed by teachers in preschool and kindergarten. Demonstrates relationship to later grade 3 reading proficiency.	Gives comprehensive evaluation of children entering kindergarten.
Parent-report	ASQ	Summative	Comprehensive screening program, used to identify children who need additional evaluations	Standardized on 15,138 children from families of different educational and economic backgrounds	Inter-observer agreement r=.92; Test-retest r=.95	The ASQ was validated against the Bayley Scales of Infant Development. Concurrent =.84; Sensitivity =.72; Specificity = .86	Screening for developmental delay
	PEDS	Summative	A surveillance tool and screening test to elicit parents' concerns about their child's development and health	It was standardized on 2823 families in the US from various backgrounds		PEDS has been compared with 14 other developmental assessment including Bayley's.	To elicit parents' concerns about their child's development and health.
	CDI	Summative	Screening requiring in-depth developmental information from parents.	The CDI came from 30 years of research with the Minnesota Child Development Inventory		Concurrent validity with IQ and achievement.	To screen and assess children where there are concerns about development, or follow high-risk children