Test selection, adaptation, and evaluation: a systematic approach to assess nutritional influences on child development in developing countries

Elizabeth L. Prado[1,2]*, Sri Hartini[1], Atik Rahmawati[1], Elfa Ismayani[1], Astri Hidayati[1], Nurul Hikmah[1], Husni Muadz[1], Mandri S. Apriatni[1], Michael T. Ullman[3], Anuraj, H. Shankar[1,4] and Katherine J. Alcock[2]

[1]SUMMIT Institute of Development, University of Mataram, Indonesia
[2]Lancaster University, UK
[3]Georgetown University, USA
[4]Johns Hopkins University, USA

*Requests for reprints should be addressed to Elizabeth Prado, SUMMIT Institute of Development, Gedung Pusat Penelitian Bahasa dan Kebudayaan (P2BK), University of Mataram, Jalan Pendidikan No 37, Mataram, NTB, Indonesia (e-mail: e.prado@lancaster.ac.uk).

Acknowledgments: This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Support for this project was also provided by the Allen Foundation. We would like to thank the SUMMIT Study Group, whose rigorous work in carrying out the SUMMIT program laid the foundation for this research an made it possible. We also thank the families and communities in Lombok who participated in the study and the Indonesian Institute of Science and National Institute of Health Research and Development for their support of the research.

Test Selection, Adaptation, and Evaluation: A Systematic Approach to Assess Nutritional Influences on Child Development in Developing Countries

Abstract

Background. Evaluating the impact of nutrition interventions on developmental outcomes in developing countries can be challenging since most assessment tests have been produced in and for developed country settings. Such tests may not be valid measures of children's abilities when used in a new context.

Aims. We present several principles for the selection, adaptation, and evaluation of tests assessing the developmental outcomes of nutrition interventions in developing countries where standard assessment tests do not exist. We then report the application of these principles for a nutrition trial on the Indonesian island of Lombok.

Sample. 300 children age 22-55 months in Lombok participated in a series of pilot tests for the purpose of test adaptation and evaluation. 487 42-month-old children in Lombok were tested on the finalized test battery.

Methods. The developmental assessment tests were adapted to the local context and evaluated for a number of psychometric properties, including convergent and discriminant validity, which were measured based on multiple regression models with maternal education, depression, and age predicting each test score.

Results. The adapted tests demonstrated satisfactory psychometric properties and the expected pattern of relationships with the three maternal variables. Maternal education significantly predicted all scores but one, maternal depression predicted socio-emotional competence, socio-emotional problems, and vocabulary, while maternal age predicted socio-emotional competence only.

Conclusion. Following the methodological principles we present resulted in tests that were appropriate for children in Lombok and informative for evaluating the developmental outcomes of nutritional supplementation in the research context. Following this approach in future studies will help to determine which interventions most effectively improve child development in developing countries.

The extent to which nutritional supplementation programs can improve developmental outcomes in developing countries is an important topic of ongoing research (Grantham-McGregor et al., 2007). However, few developmental assessment tests have been produced in and for developing country settings. Tests produced and standardized in one language, culture, and setting cannot be assumed to be valid in a setting that is different from that of the original target population (Greenfield, 1997; Rogler, 1999; van de Vijver & Tanzer, 1997). For example, children's test performance can depend on their familiarity with the test format (Greenfield & Childs, 1977) and materials (Serpell, 1979). Such familiarity levels may be quite different for children who grow up in different cultures and contexts.

Although cross-cultural research has demonstrated the importance of taking contextual differences into account, many nutrition studies in developing countries have used standard tests with minimal adaptation or evaluation on the local population (Cao et al., 1994; Caulfield, 2004; Hamadani et al., 2002; Husaini et al., 1991; O'Donnell et al., 2002; Soewondo, Husaini, & Pollitt, 1989; Whaley et al., 2003). One challenge to conducting this type of adaptation is the lack of a published systematic approach to adapting and evaluating tests for use in different contexts. This paper aims to address this gap by presenting methodological principles for test adaptation and evaluation in areas where standard tests do not exist.

In addition to test adaptation and evaluation, test selection is another important methodological step that influences the conclusions that can be drawn from nutrition research. This paper aims to promote improved assessment of the developmental outcomes of nutrition interventions in developing countries by presenting methodological principles for test selection, adaptation, and evaluation. We then report the application of these principles to assess the developmental outcomes of the

Supplementation with Multiple Micronutrients Intervention Trial (SUMMIT) (SUMMIT Study Group, 2008) on the Indonesian island of Lombok. SUMMIT was designed to compare the effects of a maternal supplement containing iron and folic acid to a multiple micronutrient supplement containing fifteen vitamins and minerals.[1] The children of a subset of SUMMIT participants were assessed on a battery of developmental tests at age 42 months in order to measure the relative effects of the two maternal supplements on subsequent child development. The methods for developing the assessment tests are reported here.

First, we argue that test selection, adaptation, and evaluation are important methodological steps to ensure that test scores are informative, appropriate, and valid measures of children's abilities in developing countries. The principles for test

[1]The World Health Organization (WHO) recommends distribution of iron and folic acid supplements to pregnant women. Change in practice towards distribution of multiple micronutrients has been hindered by a lack of empirical evidence for its benefits and risks. In recognition of the need for experimental evidence, UNICEF, WHO, and UN University specified the UN international multiple micronutrient preparation (UNIMMAP), a formulation of 15 micronutrients recommended for evaluation in trials of maternal supplementation (UNICEF/WHO/UNU, 1999). These micronutrients include iron, folic acid, iodine, zinc, vitamin A, vitamin C, vitamin D, vitamin E, vitamin B1, vitamin B2, niacin, vitamin B6, vitamin B12, copper, and selenium. SUMMIT was a double-blind cluster randomized trial carried out by the University of Mataram, the Government of Nusa Tenggara Barat Province, the Ministry of Health of Indonesia, and Helen Keller International (SUMMIT Study Group, 2008) which compared iron and folic acid to a multiple micronutrient supplement following the UNIMMAP formulation.

selection aim to ensure that the tests are maximally informative to clarify our understanding of nutritional influences on child development. Test adaptation is important so that the tests are appropriate for the local population. The principles for test evaluation aim to establish that the adapted test scores are reliable and valid measures of children's abilities in the local context.

Test Selection

To clarify the effects of undernutrition on child development, tests assessing specific developmental abilities and cognitive functions should be selected (Connolly & Kvalsvig, 1993; Horowitz, 1989; Hughes & Bryan, 2003). Many nutrition studies have used global tests of development or intelligence, such as the Bayley Scales of Infant Development (Bayley, 1993) or measures of a Developmental Quotient (DQ) or Intelligence Quotient (IQ) (Benton, 1992; Black et al., 2004; Castillo-Duran et al., 2001; Gardner et al., 2005; Hamadani et al., 2002; Hsueh & Meyer, 1981; Lind et al., 2004; Lozoff et al., 1987; Seshadri & Gopaldes, 1989; Southon et al., 1994; Zhou, Gibson, Crowther, Baghurst, & Makrides, 2006). While these global measures may be useful, especially for assessing development in early infancy when distinct cognitive abilities are difficult to measure separately, assessing specific abilities provides several advantages over global assessments.

First, performance on global tests often depends on a number of lower-level cognitive abilities. For example, performance on an IQ test probably depends in part on the ability to focus and sustain attention, working memory capacity, speed of information processing, reasoning ability, and executive function. The demonstration of any effects of a nutrition intervention on an IQ score would not indicate which lower-level ability or combination of abilities might have been specifically affected. Conversely, a lack of an effect on an IQ score does not necessarily mean that all

cognitive components are intact, particularly since children may be able to compensate for deficits in one area of ability while carrying out a more global task.

Second, nutrition is only one of a host of biological and environmental variables that influence children's development (Deater-Deckard & Cahill, 2006); thus, the effects of nutrition interventions might be quite subtle. Tests assessing specific abilities are more likely than global assessments to detect the potentially subtle cognitive changes that might follow nutritional supplementation (Hughes & Bryan, 2003).

Third, assessing specific abilities may help to clarify the effects of nutritional deficiency on brain development and the cognitive capacities that emerge as various areas of the brain mature (Hughes & Bryan, 2003). Throughout foetal development, infancy, and childhood specific areas of the brain develop at different rates during different periods of time. The long-term impact of nutritional deficiency during these periods may depend on which brain areas are developing at a rapid rate during the period of nutritional deprivation. Assessing specific rather than global abilities allows inferences concerning which periods of development and brain areas are specifically vulnerable to nutritional influences. It may also shed light on the biological mechanisms through which nutrition affects child development.

Test Adaptation

When adapting tests that originate in developed countries for use in developing countries settings, it is important to consider what aspects of the tests must be adapted and what modifications might be necessary to increase test appropriateness. Based on our review of the literature, we identified four aspects of tests that must be adapted: the test items, materials, instructions, and procedures.

First, the items in the test might not be relevant to the target culture, thus they may not measure the underlying construct they were designed to measure (Rogler, 1999). Traditional child development scales, such as the Bayley Scales of Infant Development, illustrate this problem. Such scales are based on the attainment of behavioural items in a normative sample of children in the country where the test originates. For example, the typical American child learns to squat after learning to crawl and stand. However, the order of attainment of these milestones may differ in other cultures. In Bali, crawling is explicitly discouraged because it is considered animal-like. Balinese children learn to squat as they progress from flexible movement on all-fours to sitting then squatting and standing (Super, 1981). While the failure of an American child of a certain age on a crawling item would indicate delayed motor development, the failure of a Balinese child at the same age might not give the same indication.

Direct translation of test items into a new language can also be problematic. For example, a vocabulary test in English and Spanish with items matched for meaning (directly translated) yielded different means and standard deviations for two groups of students matched on grade, age, sex, and academic achievement. However, when the items were matched on frequency of use rather than meaning, the two versions yielded similar means and standard deviations in both languages (Tamayo, 1987). Adapting the items based on language-specific criteria, that is frequency of use in each respective language, resulted in measures of vocabulary knowledge that were appropriate for each linguistic group. This type of adaptation can be particularly challenging in developing countries where published frequency norms (and other linguistic information) are not typically available. Thus, adaptation of language tests can require extensive surveys and piloting.

Second, test materials must be adapted to the local setting, since children's performance on a task can depend on their familiarity with the materials they are asked to use. The performance of British and Zambian children on a figure copying task illustrates this point. Using paper and pencil, British children, who were familiar with drawing and colouring, scored higher than Zambian children. On the other hand, Zambian children, who were familiar with making wire models, performed better than British children in copying wire figures. The children performed equally well reproducing models in plasticine, play material familiar to both groups (Serpell, 1979).

Third, test instructions must also be considered when using a test in a new context. For instance, modified instructions were required to use a Piagetian conservation task in Senegal (Greenfield, 1997). In the original task, the tester transferred water from a shorter, fatter glass to a longer, thinner one then asked the child if the quantity of water was more, less, or the same. When the tester asked the follow up question "Why do you think it is the same (or more or less)?" unschooled Senegalese children did not respond. When she instead asked "Why is the water the same (or more or less)?" the children responded with articulate reasons for their judgements Greenfield (1997) argued that these children did not make a distinction between their own thoughts about something and the thing itself. Thus, their poor response performance using unadapted instructions did not reflect their true ability to provide reasons for their judgements

Fourth, unadapted procedures can also lead to inaccurate assessment of children's abilities. Multiple choice is a common testing format in many formal education systems. When this format was used with Zincantecan Mayan children in a pattern continuation task, they were visibly confused and performed poorly. However,

when the patterns were presented as coloured sticks in a wooden frame and children were asked to continue the pattern by filling the frame with additional sticks, children were able to perform the task (Greenfield & Childs, 1977). Their confusion in the multiple choice test was not due to their inability to perform the task but rather to their unfamiliarity with the testing procedure.

As these examples illustrate, the valid assessment of children's abilities in developing countries requires the careful adaptation of test items, materials, instructions, and procedures to the target culture, language, and setting. The goal in this type of test adaptation is to assess the same underlying ability as the original test in a way that is appropriate in the local context. This is similar to the etic/emic approach in cross-cultural psychology in which an etic (universal) construct is identified then emic (culture-specific) ways of measuring this construct are developed and validated (Davidson, Jaccard, Triandis, Morales, & Diaz-Guerrero, 1976).

Test Evaluation

Once standard tests have been modified, they must be evaluated for the target population, since the test properties associated with the original version of the test cannot be assumed to apply to the modified version. Tests can be evaluated for several types of statistical properties that demonstrate their reliability, validity, and usefulness (Abubakar, Alcock, & Holding, 2008). First, the distribution of scores can be examined in order to determine the discriminatory power of the tests. If a test exhibits ceiling or floor effects then it is unable to distinguish between children with differing abilities. Second, test-retest reliability can be measured to determine whether or not a test score is a stable measure of a child's ability. This involves testing a group of children at two time points and examining the correlation between the scores. Third, high inter-rater agreement, which is the agreement between two testers

simultaneously scoring the same testing session, demonstrates that there is little chance for error in test scores due to scoring error. Fourth, internal consistency can be measured, usually with Cronbach's Alpha, providing evidence that all of the items in a test measure the same construct. Fifth, evaluation of the developmental sensitivity of the tests should demonstrate that test scores increase with age. Finally, test scores should show expected patterns with variables theorized to be related to them (convergent validity) and not related to them (discriminant validity).

These principles for test selection, adaptation, and evaluation were applied in Lombok to assess the developmental outcomes of SUMMIT. Test adaptation and aspects of test evaluation were accomplished through a series of pilot tests. The pilot testing phase was followed by the data collection phase, during which the finalized battery was administered to the main cohort of SUMMIT children. We expected that the careful selection and adaptation of the tests would result in tests that demonstrate a high level of discriminatory power, developmental sensitivity, test-retest reliability, inter-rater agreement, and internal consistency, as well as convergent and discriminant validity.

Methods

Research Site

Lombok is an island in the Nusa Tenggara Barat province of Indonesia. It is comprised of three administrative districts (East, Central, and West Lombok), all of which were included in the present study, as well as the capital city of Mataram, which was not included. Most of the 2.7 million inhabitants of Lombok are ethnically Sasak. Diverse socio-economic and living conditions exist throughout the island. Wealthier families live in brick houses with televisions and satellite dishes while poorer families live in bamboo huts without electricity or running water. Houses are

usually situated very close to each other, with shared walkways and yards where adults and children spend much of their time congregating with other members of the community.

Lombok is also quite diverse linguistically. Sasak, the principal language, is traditionally classified into five dialects (Syahdan, 1996; Teeuw, 1951), however, substantially more dialect variation seems to exist. Sasak speakers tend to identify their dialect with their particular village or neighbourhood and phonological, lexical and grammatical idiosyncrasies can be found from one village to another (Jacq, 1998). Most adults and school children also speak Indonesian, which is the official national language of Indonesia and is the medium of academic instruction from primary school to university, as well as the language of government, public meetings, and the media.

The literacy rate in Lombok among adults (age 15 and over) was 72.8% in 2004 (Kerjasama BAPPEDA Provinsi Nusa Tenggara Barat dengan Badan Pusat Statistik Provinsi Nusa Tenggara Barat, 2004). Among the pregnant women who participated in SUMMIT (n > 30,000), 12% had never been to school, 49% had 1-6 years of education, 21% had 7-9 years of education, and 15% had ten or more years of formal education (SUMMIT Study Group, 2008).

Thus, the living conditions, linguistic situation, and education levels in Lombok are quite different from most developed countries, where developmental assessments originate. All of these factors were taken into account when selecting and adapting the developmental tests for use in Lombok.

Participants

During the piloting phase, 167 children (90 girls) age 30-55 months in 15 villages across East, Central, and West Lombok participated in a series of pilot tests for the purpose of test adaptation. An additional 83 children (45 girls) age 39-45

months were then tested for the establishment of test-retest reliability. An additional 50 children (21 girls) age 22-50 months were then tested on a subset of the tests that were modified during reliability testing to assess the developmental sensitivity of these tests.

During the data collection phase, 487 children of SUMMIT participants (231 girls) were tested on the finalized test battery. These children were tested within 3 weeks of the date on which they were 42 months old. To distinguish these children from the pilot test participants, they will be referred to as the main cohort of SUMMIT children.

All testing was conducted at the homes of the participants. At each visit, the purpose of the research was explained to a parent or caregiver, who indicated their consent for their child to participate by signing an informed consent form. Children indicated their assent by their willingness to participate in the activities. Ethical approval for the informed consent and research procedures was obtained from the Lancaster University Psychology Department Ethical Committee and the Mataram University Ethical Research Committee.

Procedure

Test Selection

Our goal in test selection was to choose a battery of tests assessing specific abilities that develop during early childhood and are likely to be sensitive to nutritional influences. Research in maternal and child undernutrition in humans and animals suggests possible effects on motor development (Gorman, 1995), language development (O'Donnell et al., 2002; Pollitt, 1993), and non-verbal cognitive development, including visuospatial ability, attention, and executive function (Hughes & Bryan, 2003), as well as socio-emotional development (Black, 2003; Strupp &

Levitsky, 1995). Tests that assess each of those domains were chosen, focusing on tests that (1) are appropriate for three-year-old children, (2) are well-established and widely-used, (3) do not require special equipment such as computers or recording equipment, (4) do not require verbal responses, since children in Lombok were likely to be shy and (5) are brief to administer and objectively scored, that is do not require subjective judgements or ratings from the testers. The tests that were selected and piloted are presented in Table 1 and described in detail below.

TABLE 1 ABOUT HERE

Testers

Four university graduates were recruited and trained to administer the developmental tests. All testers were native speakers of Sasak and also fluent in Indonesian.

Manuals and Forms

Test instructions and procedures were translated from English to Indonesian. Since Indonesian is the medium of academic instruction, the testers were more comfortable reading and writing Indonesian than Sasak; therefore, the testing manuals were written in Indonesian. However, since Sasak is the predominant spoken language in Lombok, all testing sessions were conducted in Sasak, including test instructions and presentation of stimuli. To ensure uniformity in test administration, test instructions and items were printed on the testing forms in Sasak. To account for dialect variation across the island, instructions were translated into the main dialects of five areas of the island (north, west, central, south, and east Lombok) and five types of forms were printed. As the instructions and procedures were modified based on pilot test results, the testing manuals and the instructions printed on the forms were revised accordingly.

Test Adaptation

Our goal in test adaptation was to develop tests that assessed the same underlying ability as the original tests in ways that were appropriate for children in Lombok. To test the same ability, the instructions and procedures were adapted so that they elicited the target behaviour. To establish cultural appropriateness, the items and materials were adapted to the language, culture, and testing conditions in Lombok.

This adaptation was accomplished through a series of ten pilot tests. After each pilot test, each test was either eliminated from the test set, modified and included in the subsequent pilot test, or confirmed for inclusion in the test set without further modification. This decision was based in part on discussion of the practical aspects of the test session between the testers and their supervisor. This decision was also based on some aspects of test evaluation which were conducted in parallel with the test adaptation process and informed some of the adaptations that were made to the tests. For example, after each pilot test, the resulting distribution of scores was examined to evaluate discriminatory power. Any test with a negatively skewed distribution was modified to make the test more difficult and, conversely, any test with a positively skewed distribution was modified to make it easier. For some tests, such as the Sentence Complexity Scale, the developmental sensitivity of each item was used to determine which items to retain or eliminate. The adaptations that were made to each test based on pilot data are discussed in more detail below.

Test Evaluation

After the completion of test adaptation, two rounds of reliability testing were conducted. In each round, each tester administered the test battery to a group of children twice, with one week separating the two test sessions. The Pearson's

correlation between the first and second testing was calculated. Tests for which $r < .7$ in the first round of reliability testing were revised to further standardize test administration and subsequently evaluated in the second round of reliability testing. The age of participants was restricted to 39-45 months (within three months either side of the target age for the main cohort). At the end of the data collection period, each tester was also assigned to revisit a group of participants (from the main cohort) to ensure that the tests were administered consistently throughout the data collection period.

The tests that were modified after the first round of reliability testing were administered to an additional group of children age 22-50 months to evaluate developmental sensitivity. For the tests that were not modified, developmental sensitivity was evaluated based on the pilot data from the final version of the test.

To evaluate inter-rater agreement, two testers visited a group of children (from the main cohort), with one tester administering and scoring the tests, while the second tester independently scored the child's performance and recorded the parent's responses. Agreement was considered sufficient if the percent of items scored identically by the two testers was greater than 90%.

Discriminatory power, internal consistency, and convergent and discriminant validity were evaluated on the full set of data from the main cohort.

Description of Tests and Context-Specific Modifications

Here, we describe the tests that were chosen for pilot testing, presented in Table 1, and the modifications that were made to adapt the tests to the local context according to the principles and procedure described above.

Motor Development

The motor assessment was developed by selecting age-appropriate fine and gross motor items from the Bayley Scale of Motor Development (Bayley, 1993) and the Ages and Stages Questionnaire (Schaefer & DiGeronimo, 2000). Items were adapted using materials familiar to children in Lombok. For example, since few houses in Lombok contain stairs or store-bought toys, we replaced items assessing a child's ability to climb stairs with the item "Climbs onto a terrace 50 cm high" and we used a rope made of rubber bands strung together rather than a store-bought jump rope. The original version of the test resulted in a negatively skewed distribution of scores, therefore several higher difficulty items were added. The fine and gross motor scores were the total number of fine and gross motor items, respectively, that a child performed successfully.

Language Development

Two language tests were selected: a Picture Vocabulary Test assessing receptive vocabulary knowledge and a Sentence Complexity Scale measuring expressive sentence complexity, based on parent report.

Picture Vocabulary Test. This test was based on the British Picture Vocabulary Scale (Dunn, Dunn, Whetton, & Burley, 1997), which has been successfully adapted for other developing country settings (Holding et al., 2004). The child was shown four pictures and asked to point to the picture that matched the word spoken by the tester. To develop the target items, 65 Sasak words were chosen, comprising 40 nouns, 15 verbs, and 10 adjectives. Most of these target words were confirmed to be used in 107 villages throughout Lombok, based on a previously conducted dialect survey (unpublished data). For each target item, a phonological distracter, a semantic distracter, and an unrelated word were chosen, all of the same grammatical class as the target. A local illustrator produced colour drawings of these

target and distracter words. Based on several pilot tests, some items were eliminated due to ceiling effects or variation across dialects for the target item. Several items were also revised (i.e., pictures redrawn) to increase their clarity. The resulting test included 52 items, 2 practice items and 50 test items, which were arranged in order of difficulty based on the pilot results. All items were administered to every child. The score was the number of correctly identified items.

Sentence Complexity Scale. This test was based on the McArthur Communicative Development Inventory – Level III (Dale, Reznick, & Thal, 1998). For half of the items (items 1-12), the parent was given two example sentences, one comparatively more complex than the other, and asked to choose which one more closely resembled his or her child's speech. To develop these items, two hours of spontaneous speech was recorded and transcribed from 6 children age 39-50 months. We examined the type of construct in the English items and looked for a comparable example of child speech in the Sasak transcripts. For example the English item: "Don't read book/Don't want you read that book" was changed to "*Tas tə kadu/Tas tə kadu jauq robot*" which means "I'm using my bag/I'm using my bag to carry a robot."[2] Items 13-24, which consisted of questions concerning the child's word use or conceptual ability, were simply translated into Sasak.

Based on pilot testing, three items were eliminated due to ceiling effects and six items were eliminated that did not correlate with the child's age (in months). We

[2] In this example, the number of morphemes in the Sasak item is similar to the English item ("Don't want you read that book"), even though the English translation of the Sasak item ("I'm using my bag to carry a robot") seems more complex. This is due to a lack of articles (*a*, *the*) or inflectional morphology (including tense marking) in the Sasak language.

also modified the instructions to increase their clarity by adding a context to the examples of child speech. For example, when interviewing the mother, we would say: "What if your child wanted a new shirt, what would she say to her friend? 'I want a new shirt' or 'I want my mother to buy me a new shirt.'" Sixteen items were retained in the final test; these items were summed for a total score.

Non-Verbal Cognitive Development

Visuospatial ability. Visuospatial ability was assessed by a Block Design Test, in which the child was asked to copy a shape using wooden blocks (Elliot, 1996). This test was developed based on two standard tests: the British Abilities Scale (BAS) (Elliot, 1996) and the WPSSI-III (Wechsler, 2002). The BAS items, which use plain uncoloured blocks, were more appropriate for children in Lombok than the red and white plastic blocks used in the WPSSI-III. However, the method of the administration from the WPSSI-III was more suitable. This test allows for a second attempt to build the design if the child does not succeed within 30 seconds on the first attempt. The chance for a second explanation and demonstration for each item improved children's performance and the time limit increased the efficiency of test administration. After various modifications based on the distribution of scores from several pilot tests, the final test consisted of the BAS items 1-7. The children were allowed a second attempt on items 1-4 only. Children scored two points for succeeding on the first attempt and one point for succeeding on the second attempt. Their time to successfully build each design was also recorded. Their scores on each item were summed for a total score and their average time per correct item was also calculated.

Attention. A Visual Search Test was developed based on the NEPSY Developmental Neuropsychological Assessment visual search subtest (Korkman,

Kirk, & Kemp, 1998). In this test, the child is shown an array of pictures and is asked to circle every instance of a target picture. This test measures the child's visual focus and ability to inhibit distracters. Each drawing on the NEPSY stimulus sheet was substituted with a locally-drawn picture that was more appropriate for children in Lombok. For example, a picture of a bunny was replaced with a picture of a chicken. Based on pilot testing, the procedures of the test were modified such that the child pointed to each instance of the target picture and the tester circled each picture indicated by the child. Two trials were administered; both accuracy and time were recorded for each trial. The score was calculated as the number of *hits* (number of targets correctly indicated) minus the number of *false alarms* (number of distracter pictures incorrectly indicated). Each child's average time per correct item was also calculated as the total number of hits divided by the time to complete the task.

Executive function. Five tests of executive function were piloted and two were retained in the battery used for data collection.

In the Knock and Tap Test, the child and the tester played a hand game requiring inhibition of automatic responses. Children were required first to imitate the tester, either knocking on the floor with their fist or tapping with their palm. Then children were instructed to knock when the tester tapped, and vice versa (Korkman et al., 1998). This test was piloted with various modifications, for example using a similar game with hand shapes that were more familiar to children in Lombok. However, the children were largely reluctant to play any sort of hand game, such that the test resulted in very low scores for some children or was impossible to score for others. Therefore, this test was eliminated from the test battery and was not administered to the main cohort.

In the Tower Test, the child was invited to help the tester build a tower with wooden blocks and was scored on his or her ability to take turns with the tester (Carlson, 2005; Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996). Many children were very shy and therefore reluctant to put blocks on the tower even when it was their turn. This test also resulted in low scores or missing data, thus was also eliminated.

In the Block Sorting Test, children were first asked to sort big blocks into a big bucket and small blocks into a small bucket. The tester then reversed the sorting rule (put big blocks into the small bucket and vice versa) (Carlson, 2005; Carlson, Mandell, & Williams, 2004). Scores on this test were negatively skewed. We tried several modifications to increase its difficulty, including introducing multiple switches between sorting rules and sorting according to colour rather than size. The version of the test in which children sorted according to colour then against colour (put the blue blocks in the red bucket and the red blocks in the blue bucket) resulted in the best distribution of scores. The score on this test was the number of blocks (out of twelve) placed in the correct bucket after the sorting rule was reversed. This test was retained for evaluation of test-retest reliability and developmental sensitivity then subsequently eliminated from the test battery (see below).

In the Snack Delay Test, the tester placed a snack under a clear cup and told the child to wait until she rang a bell before taking the snack (Carlson, 2005; Kochanska, Murray, & Harlan, 2000). After a brief demonstration, four trials were administered with delays of 5, 15, 30, and 45 seconds. If the child took the snack before the allotted time, the trial was scored as a fail and the amount of time the child waited before taking the snack was recorded. Based on pilot testing, this procedure was modified in two ways. First, since use of a bell was unfamiliar to children in

Lombok, they were instead instructed to wait until told before taking the snack. Most children scored at ceiling on this version of this test. The second modification, to make the test more difficult, was to place the snack in the child's hand and instruct him or her to wait until told before eating it. This procedure resulted in a better distribution of scores. The score on this test was the number of trials on which the child succeeded. The average amount of time the child waited before eating the snack across the four trials was also calculated.

In the Windows Test, the tester placed a treat inside one of two boxes, each of which had a clear window through which the child could see the treat. The child was first instructed to point to the box with the treat in order to obtain the treat. After several trials, the rule was reversed and the child was instructed to point to the box without the treat in order to obtain the treat (Russell, Mauthner, Sharpe, & Tidswell, 1991). Several versions of this test were piloted, with various numbers of pre-switch trials (point to the box with the sweet) and switch trials (point to the box without the sweet). In the version that resulted in the best distribution of scores, the child was given two pre-switch trials and six switch trials. The score was the number of correct switch trials out of six.

Socio-Emotional Development

A *Socio-Emotional Development Scale* was developed, based on the Brief Infant-Toddler Social and Emotional Assessment (BITSEA) (Briggs-Gowan & Carter, 2002). In this test, the child's parent or another caregiver rated 42 items probing specific aspects of the child's behaviour on a scale from 0 to 2 (0 = not true/rarely, 1 = somewhat true/sometimes, 2 = very true/often). The BITSEA items were translated into both Indonesian and Sasak, then examples were added for each item to further explain and clarify their meaning. Item surveys were then conducted

with mothers of young children to answer two questions: first, was item was clearly understood by the respondents, and second, was the behaviour being investigated appropriate to the local culture. Survey participants were asked whether they considered the behaviour in each item good, bad, or neither good nor bad. They were then asked to give reasons for their judgements Based on their reasons, we determined whether or not they had understood the intended meaning of the item and re-explained the meaning if necessary. Participants were allowed to change their answer (good, bad, or neutral) based on their new understanding of the item. The wording and examples that resulted in accurate understanding were noted and the forms and manuals were revised accordingly. Ten items for which more than seven women (28% of respondents) did not agree with the intended classification of the item (as good or bad) were eliminated. Three additional items were eliminated based on piloting due to apparent confusion in some respondents or lack of variability in responses. Following the BITSEA, a *competence score* was calculated as the sum of the competence item ratings and a *problem score* was calculated as the sum of the problem item ratings.

Analysis and Results

Test-Retest Reliability

The test-retest reliabilities reported in Table 2 represent the combined data from the reliability testing at the beginning and at the end of the period of data collection from the main cohort.[3] Reliability coefficients were greater than 0.7 for all

[3] The reliability coefficients were also analysed separately and contrast coefficients were generated to test whether the reliability coefficient for each test was significantly different at the beginning versus at the end of the data collection period. The only score for which the coefficients differed significantly was the Block Design time per

of the test scores except the Visual Search, Block Sorting, and Snack Delay tests. The level of reliability that is considered sufficient ranges generally from $r > 0.7$ to $r > 0.9$ (Cohen & Swerdlik, 2005; Kline, 1993); however, this standard depends on the type of test and the purpose for which the test is being used. In the case of developmental tests, it is not entirely clear what level of reliability is sufficient. Unlike, for example, the measurement of the length of a room, test scores at different time points might vary due to reasons other than measurement error (Murphy & Davidshofer, 1988). Specifically, the attribute being measured may actually change in the intervening time period or the experience of having taken the test previously may change a child's true score. For example, a parent may notice which items a child failed to identify correctly on a vocabulary test and teach those words to the child before the second test session. Similarly, after observing a Snack Delay test, a parent may speak sternly to a child about eating the snack before being told, which could lead to improved performance at the second testing. Although the reliability coefficients for the Visual Search and Snack Delay tests did not quite reach 0.7, evidence from the other statistics associated with these tests (e.g., relationship with the child's age and the mother's education, reported below) suggests that these are valid and useful measures even though test-retest reliability was slightly lower than expected. In contrast, for the Block Sorting Test, test-retest reliability was relatively low ($r(44) = 0.514$, $p < .001$) and the correlation with age was non-significant ($r(24) = 0.236$, $p = 0.246$). These two results together suggested that this was not an optimal test; therefore, it was eliminated from the test battery and was not administered to the main cohort. In

---

correct item score (pre-data collection: $r(31) = 0.854$, $p < .001$, end of data collection: $r(31) = 0.383$, $p = .028$, difference: $p = .037$; for all other contrasts ps > .07).

general, the high test-retest reliability of the tests demonstrates that they were stable measures of children's abilities.

TABLE 2 ABOUT HERE

Inter-Rater Agreement

Inter-rater agreement was greater than 90% for all test scores except the average time per correct item on the Block Design test (Table 2). This score was timed independently by the two testers; thus, small discrepancies may be expected. Indeed, the differences between the times were small and the time per correct item scores from the two testers correlated strongly with each other ($r(7) = 0.990$, $p < .001$). The high inter-rater agreement demonstrated that there was little chance for error in test scores due to scoring error.

Internal Consistency

Cronbach's Alpha was calculated for the data from the main cohort of SUMMIT children. Alpha > 0.7 was considered to demonstrate sufficient inter-correlation between test items, suggesting that all items in the test measured the same construct (Nunnally, 1978). Since the Visual Search Test consisted of only two items, the internal consistency was calculated as the correlation between the two items, rather than Cronbach's Alpha.

Cronbach's Alpha was greater than 0.7 for all of the test scores except the Fine Motor score and the Socio-Emotional Scale scores (Table 2). The low internal consistency for the socio-emotional scores was consistent with the expectations of the authors of the original instrument. The BITSEA items were drawn from many different subscales of the Infant-Toddler Social and Emotional Assessment, thus, were not expected to correlate highly with each other (Briggs-Gowan & Carter, 2002). The criteria that Alpha should be greater than 0.7 was not relevant for the Visual Search

Test, since the internal consistency represents the correlation between the two trials, rather than Cronbach's Alpha. The correlations between the two trials for both the score and the time per correct item were significant ($p < .0001$). In general, the high internal consistency of the tests provided evidence that the items in each test measured the same construct.

Developmental Sensitivity

Developmental sensitivity was evaluated by calculating the Pearson's correlation between the test score and the child's age in months based on the pilot data for the final version of each test. This measure was evaluated using pilot data since a greater range of ages was represented than participants in the main cohort of SUMMIT children (all of whom were tested within three weeks of 42 months).

All test scores correlated significantly with age except the Block Sorting Test (which was eliminated from the battery; see above) and the socio-emotional competence and problem scores (Table 2). The lack of correlation for the socio-emotional scores was consistent with the results of the age analysis on the original test, which compared scores between four groups of children: age 12-17 months, age 18-23 months, age 24-29 months, and age 30-35 months. Consistent with our findings, the problem score did not differ significantly between the four age groups. Unlike our results, the age effect was significant for the competence score; however, pairwise comparisons revealed that only the youngest group of children scored significantly lower than the other three groups, which did not differ from each other (Briggs-Gowan & Carter, 2002). We probably did not observe an effect of age on the competence score since we did not test children as young as 12-17 months. In general, the significant correlations with age demonstrated that children's performance on the tests improved with age, hence the test scores are meaningful developmentally.

Discriminatory Power

Table 3 presents the descriptive statistics from each test for the data from the main cohort of SUMMIT children. We were able to obtain complete data from all 487 children for the Sentence Complexity Scale only. For the other tests, missing data represents cases where the child refused to perform or complete the test or the parent was not able to answer all test items. Also excluded was one negative Visual Search score (the calculation *hits* minus *false alarms* resulted in a negative score). In general, scores were well distributed and means were slightly above the centre of the possible range of scores. These results demonstrate that the tests were able to distinguish between children with differing abilities.

TABLE 3 ABOUT HERE

Convergent and Discriminant Validity

Multiple Regression Models

Typically when assessments of cognitive or language development are constructed in developed countries, they are validated against existing standardised tests or with groups of children who have been diagnosed with a learning disability. As no such tests or diagnoses exist in Lombok, we assessed our instrument's convergent validity against maternal factors known to affect cognitive, language, and socio-emotional development (education and depression).

Maternal education has been found to predict children's cognitive and socio-emotional function in both developed (Dollaghan et al., 1999; Duncan, Brooks-Gunn, & Klebunov, 1994) and developing countries (Khandke, Pollitt, & Gorman, 1997). Therefore, we expected to find a strong relationship between maternal education and all child development scores. Maternal depression is expected to be strongly related to socio-emotional outcomes (Caplan, Cogill, & Alexandra, 1989; Murray et al., 1999)

and may also predict motor, language, and cognitive scores (Cummings & Davies, 1994; NICHD Early Child Care Research Network, 1999; Petterson & Albers, 2001). Maternal education was quantified as the number of years of completed formal education and depression was measured by an adaptation of the Center for Epidemiological Studies Depression Scale (Radloff, 1977), which was administered to the mother at the time of child testing.

Maternal age, on the other hand, is unlikely to be related to children's developmental levels. Although children of teenage parents have been shown to score poorly on cognitive and socio-emotional tests (Roosa, Fitzgerald, & Carlson, 1982), evidence suggests that this is due to social factors (Geronimus, Korenman, & Hillemeier, 1994), which may not apply in developing countries. Age was the mother's reported age in years.

The results of the multiple regression models with maternal education, depression, and age predicting each developmental test score are reported in Table 4. These analyses excluded 175 participants for whom any of the three independent variables was not known. As expected, maternal education predicted every child development score, though the regression coefficient for the Visual Search average time per correct item was only approaching significance ($p = .069$).[4] Maternal depression significantly predicted the socio-emotional problem and competence scores, as well as the picture vocabulary score. Also as expected, maternal age was unrelated to most scores; socio-emotional competence was the only score significantly predicted by the mother's age. The demonstration of the expected pattern of

[4] For the Visual Search and Block Design time per correct item scores, the coefficients for maternal education were negative, since a smaller (faster) score indicates better performance.

relationships between these variables and the test scores provides evidence that the tests reflected the constructs they were intended to measure.

TABLE 4 ABOUT HERE

Validity of the Sentence Complexity Scale

We also performed an additional evaluation of the validity of the Sentence Complexity Scale. Although parent report questionnaires have been found to accurately reflect children's abilities (Fenson et al., 1994), even in developing country settings (Alcock, Rimba, Abubakar, & Holding, 2005), it was important to ensure that this was true for the expressive language measure we developed. Two hours of spontaneous speech from 14 of the children in the main cohort was recorded and transcribed using the CHILDES transcription system (MacWhinney, 2006). Ten percent of each transcript was re-transcribed by a second transcriber and compared to the original transcription for validation. Ten of the transcripts were found to be valid (agreement >80%). Each child's mean number of words per utterance was calculated (excluding singing and nonsense words). The correlation between a child's mean words per utterance and score on the Sentence Complexity Scale was computed. This correlation was significant ($r(8) = .759$, $p = .011$), validating this parent-report measure as an accurate reflection of a child's expressive language ability.

Discussion

In this paper, we have presented several principles for the selection, adaptation, and evaluation of child development assessment tests for nutrition interventions in developing countries. First, tests assessing specific rather than global abilities should be selected (Connolly & Kvalsvig, 1993; Horowitz, 1989; Hughes & Bryan, 2003). Second, tests should be adapted with the goal of assessing the same underlying ability as the original test in locally-appropriate ways. To accomplish this,

test instructions and procedures must be adapted to ensure that they elicit the target behaviour and test items and materials should be appropriate for the target population. Third, the adapted tests should be evaluated for psychometric properties that demonstrate their reliability, validity, and usefulness.

Following these principles in the assessment of the developmental outcomes of SUMMIT resulted in a battery of tests that were appropriate for children in Lombok, demonstrated good psychometric properties, and showed the expected pattern of relationships with maternal education, depression, and age. The confirmation of the expected pattern of relationships between these variables, together with the establishment of the discriminatory power, developmental sensitivity, test-retest reliability, inter-rater agreement, and internal consistency of the tests, validated the adapted tests as informative measures to assess the effects of maternal multiple micronutrient supplementation on children's development in Lombok.

One important aspect of this study was the evaluation of multiple measures that provide converging evidence for the validity and reliability of the tests. Evaluation of a single measure alone usually cannot determine whether or not a test will provide useful information. However, if multiple measures are evaluated, these measures can provide a body of evidence supporting the validity and reliability of a given test. Even if one or two measure are slightly lower than expected, the other measures may indicate that the test is likely to result in meaningful information concerning children's developmental levels.

We have also demonstrated the validity of two parent-report measures to assess children's development, specifically their expressive sentence complexity and socio-emotional development. Given children's relatively short attention span, test sessions usually must be limited to an hour or less, depending on the child's age. In a

test battery assessing specific rather than global abilities, time constraints might prevent assessing all of the abilities one wishes to evaluate in one test session. Under these circumstances, parent-report measures, if carefully adapted and systematically administered, can provide useful additional information concerning the child's developmental level in certain domains.

The systematic administration of the tests, not only for parent-report measures but also for child assessments, is another important aspect of valid and accurate data collection. Uniformity in test administration was accomplished in the present study in several ways. Standardized instructions were printed on the testing forms and the testing manuals stipulating detailed directions for test administration and scoring. However, detailed documentation alone does not guarantee uniform data collection; the testers must also be trained to follow the proscribed procedures. In the present study, this training was accomplished through field evaluations, which consisted of a detailed list of the instructions and procedures required for each test. Evaluation and feedback were repeated until the testers administered all of the tests without error. Supervisory visits were made throughout the data collection period to ensure that the testers continued to administer the tests correctly. Along with rigorous methodology, this rigorous training was another important aspect of the present study that produced accurate and valid data.

The results presented here suggest several directions for future research. First, the specific changes that were necessary to adapt the tests for use in Lombok reveal cross-cultural differences that may prove to be theoretically interesting. For example, the high rates of success on the original version of the Snack Delay Test (in which the snack was placed under a clear plastic cup) demonstrated a high level of inhibition ability among children in Lombok. Only when the snack was placed in the child's

hand was the task sufficiently difficult that some children failed. This high level of performance may be due to the emphasis that Indonesian parents place on compliance, even in young children. A similar pattern of high performance on executive function tasks has been found in pre-school children in China (Sabbagh, Xu, Carlson, Moses, & Lee, 2006) and Korea (Oh & Lewis, 2008), with enhanced performance in the latter study particularly for tasks involving self-control. Further research comparing British and Indonesian children on the Snack Delay Test and other executive function tests may elucidate the influence of culture and/or parenting styles on the development of inhibition ability, as well as possibly other executive functions.

Second, the principles presented here for test selection, adaptation, and evaluation may be useful to evaluate the developmental outcomes of future nutrition interventions in developing countries where standard developmental tests do not exist. Although the test battery described here focused on pre-school development, these principles can also be applied to studies assessing motor, cognitive, and socio-emotional development throughout the school years and into adulthood. These principles may also be applied to other types of interventions aimed at improving child development, such as reducing the prevalence of malaria infection, exposure to heavy metals, exposure to violence, and maternal depression (Walker et al., 2007). These principles can also be applied to education interventions, for example, interventions aimed at improving early childhood development through parental education or improving the quality of academic instruction children receive in school. The study reported here demonstrates the usefulness of the principles presented to develop a maximally informative set of test scores that accurately reflect children's abilities in different contexts. Using this approach in future studies will help advance

towards the goal of determining which interventions most effectively improve developmental and educational outcomes for children in developing countries.

Table 1

The Developmental Domains and Tests Selected for Pilot Testing

Developmental Domain	Developmental Test
Motor Development	
	Fine Motor Development Scale
	Gross Motor Development Scale
Language Development	
	Picture Vocabulary Test
	Sentence Complexity Scale
Non-Verbal Cognitive Development	
Visuospatial Ability	
	Block Design Test
Attention	
	Visual Search Test
Executive Function	
	Knock and Tap Test[a]
	Tower Test[a]
	Block Sorting Test[a]
	Snack Delay Test
	Windows Test
Socio-Emotional Development	
	Socio-Emotional Development Scale

[a]The Knock and Tap Test and Tower Test were eliminated from the test set after pilot testing and the Block Sorting Test was eliminated after reliability testing; these tests were not administered to the main cohort of SUMMIT children.

Table 2

Inter-Rater Agreement, Internal Consistency, Test-Retest Reliability, and Correlation with Age for the Test Battery Assessing Child Development in Lombok

	Test-Retest Reliability		Inter-Rater Agreement		Internal Consistency		Correlation with Age		
Test Score	*n*	*r*	*n*	*%*	*n*	*Cronbach's Alpha*	*n*	*r*	*p*
Motor Development Scale									
Fine Motor Test score	77	0.715	10	90.0%	437	0.514	77	0.417	<.001
Gross Motor Test score	74	0.740	10	97.8%	341	0.722	78	0.261	.021
Picture Vocabulary Test score	81	0.723	11	99.5%	395	0.741	49	0.744	<.001
Sentence Complexity Scale score	81	0.861	12	99.5%	376	0.736	50	0.876	<.001
Block Design Test									
Score	69	0.797	10	100%	466	0.743	78	0.547	<.001
Time per Correct Item	66	0.705	10	78.6%	--	--	77	-0.514	<.001
Visual Search Test	80	0.585			475		78	0.255	.025
Score	73	0.691	10	100%	467	0.480[a]	76	-0.392	<.001

Time per correct item			10	90.0%		0.802[a]			
Block Sorting Test	46	0.514	--	--	--	--	26	0.236	.246
Snack Delay Test									
Score	80	0.599	10	100%	448	0.722	40	0.453	.003
Average delay across trials	77	0.634	10	97.5%	--	--	39	0.542	<.001
Windows Test Score	70	0.797	10	100%	470	0.879	67	0.528	<.001
Socio-Emotional Development Scale									
Competence score	79	0.781	12	96.6%	478	0.418	50	0.056	0.699
Problem score	79	0.803	12	96.6%	474	0.624	47	-0.197	0.185

Note. Internal consistency was not calculated for the Block Design average time per correct item score or for the Snack Delay average delay across trials, since these scores were not calculated for every item individually. Inter-rater agreement and internal consistency were not calculated for the Block Sorting Test, since this test was not administered to the main cohort of SUMMIT children. For the test-retest reliability coefficients, all *p*s < .0001.

[a]For the Visual Search Test, the internal consistency represents the correlation between the two trials rather than Cronbach's Alpha. For both correlations $p < .0001$.

Table 3

Descriptive Statistics of the Finalized Test Battery for the Data from the Main Cohort of SUMMIT Children

Test Score	*N*	Maximum possible	Range	*Mean*	*SD*
Motor Development Scale					
Fine Motor Test score	468	10	0 – 10	5.92	2.00
Gross Motor Test score	444	10	1 – 10	7.16	2.02
Picture Vocabulary Test score	477	50	16 – 47	32.35	5.82
Sentence Complexity Scale score	487	16	0 – 16	11.43	2.86
Block Design Test					
Score	468	14	0 – 14	9.06	3.42
Time per correct item	458	--	6 – 55	18.81	7.83
Visual Search Test					
Score	478	40	3 – 40	32.13	5.90
Time per correct item	470	--	4 – 43	8.54	2.52
Snack Delay Test					
Score	468	4	0 – 4	2.73	1.38
Average delay across trials	445	24	1 – 24	18.88	6.91
Windows Test score	472	6	0 – 6	3.57	2.32
Socio-Emotional Development Scale					
Competence score	478	16	8 – 16	13.00	1.87
Problem score	474	42	0 – 19	6.86	3.55

Note. For the Block Design and Visual Search time per correct item scores, there is no maximum possible score; smaller (i.e., faster) scores indicate better performance.

Table 4

Results of the Regression Models with Maternal Education, Age, and Depression Predicting Each Child Development Score

Test Score	*df*	Maternal Education			Maternal Depression			Maternal Age		
		B	*t*	*p*	*B*	*t*	*p*	*B*	*t*	*p*
Motor Development Scale										
Fine Motor Test score	292	0.090	2.58	.010	-0.017	0.82	.415	-0.015	0.73	.467
Gross Motor Test score	278	0.094	2.75	.006	0.008	0.42	.678	0.021	1.05	.294
Picture Vocabulary Test score	299	0.303	3.19	.002	-0.135	2.37	.018	0.005	0.09	.926
Sentence Complexity Scale score	308	0.159	3.32	.001	-0.029	1.02	.306	-0.009	0.31	.759
Block Design Test										
Score	293	0.232	4.04	<.001	-0.027	0.79	.429	0.013	0.39	.698
Time per correct item	288	-0.283	2.04	.042	0.002	0.02	.985	0.049	0.59	.554
Visual Search Test										
Score	300	0.259	2.45	.015	-0.080	1.27	.207	0.057	0.92	.360
Time per correct item	293	-0.092	1.82	.069	0.019	0.64	.523	0.003	0.10	.924
Snack Delay Test	296	0.049	2.08	.038	-0.027	1.91	.057	0.003	0.25	.804
Score	280	0.281	2.35	.019	-0.116	1.64	.103	0.058	0.84	.401

Average delay across trials										
Windows Test score	297	0.124	3.08	.002	-0.034	1.42	.156	0.018	0.76	.448
Socio-Emotional Development Scale										
Competence score	303	0.075	2.31	.022	-0.044	2.25	.025	0.054	2.81	.005
Problem score	298	0.128	2.26	.025	0.185	5.45	<.001	-0.056	1.64	.102

References

Abubakar, A., Alcock, K., & Holding, P. A. (2008). Adapting Western developmental measures for use in resource poor settings: Methodological issues. Archives of Disease in Childhood.

Alcock, K. J., Rimba, K., Abubakar, A., & Holding, P. A. (2005). First words in two east African languages. Paper presented at The International Congress for the Study of Child Language, Berlin.

Bayley, N. (1993). Bayley Scales of Infant Development (2nd ed.). San Antonio, TX: Psychological Corporation.

Benton, D. (1992). Vitamin-mineral supplements and intelligence. Proceedings of the Nutrition Society, 51, 295-302.

Black, M. M. (2003). The evidence linking zinc deficiency with children's cognitive and motor functioning. Journal of Nutrition, 133, 1473S-1476S.

Black, M. M., Baqui, A. H., Zaman, K., Persson, L. A., Arifeen, S. E., Le, K., et al. (2004). Iron and zinc supplementation promote motor development and exploratory behavior among Bangladesh infants. American Journal of Clinical Nutrition, 80, 903-910.

Briggs-Gowan, M. J., & Carter, A. S. (2002). Brief infant-toddler social and emotional assessment (BITSEA) manual, version 2.0. New Haven, CT: Yale University.

Cao, X.-Y., Jiang, X.-M., Dou, Z.-H., Rakeman, M. A., Zhang, M.-L., O'Donnell, K. J., et al. (1994). Timing and vulnerability of the brain to iodine deficiency in endemic cretinism. New Engl J Med.

Caplan, H. L., Cogill, S. R., & Alexandra, H. (1989). Maternal depression and the emotional development of the child. British Journal of Psychiatry, 154, 818-822.

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology, 28(2), 595-616.

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2 to 3. Developmental Psychology, 40(6), 1105-1122.

Castillo-Duran, C., Perales, C. G., Hertampf, E. D., Marin, V. B., Rivera, F. A., & Icaza, G. (2001). Effect of zinc supplementation on development and growth of Chilean infants. J Pediatr, 138, 229-235.

Caulfield, L. (2004). Maternal Zinc Deficiency and Maternal and Child Health in Peru: The 2000 Avanelle Kirksey Lecture, Purdue University. Nutr Today, 39(2), p78-87.

Cohen, R. J., & Swerdlik, M. E. (2005). Psychological Testing and Assessment: An Introduction to Tests and Measurement (6th ed.). Boston: McGraw Hill.

Connolly, K. J., & Kvalsvig, J. D. (1993). Infection, nutrition and cognitive performance in children. Parasitology, 107, S187-S200.

Cummings, E. M., & Davies, P. T. (1994). Maternal depression and child development. Journal of Child Psychology and Psychiatry, 35, 73-112.

Dale, P. S., Reznick, J. S., & Thal, D. J. (1998). A parent report measure of language development for three-year-olds. Paper presented at the International Conference on Infant Studies, Atlanta, Georgia.

Davidson, A. R., Jaccard, J. J., Triandis, H. C., Morales, M. L., & Diaz-Guerrero, R. (1976). Cross-cultural model testing: Toward a solution of the etic-emic dilemma. International Journal of Psychology, 11, 1-13.

Deater-Deckard, K., & Cahill, K. (2006). Nature and nurture in early childhood. In K. McCartney & D. Phillips (Eds.), Blackwell handbook of early childhood development (pp. 3-21). Malden, MA: Blackwell Publishing.

Dollaghan, C. A., Campbell, T. F., Paradise, J. L., Feldman, H. M., Janosky, J. E., Pitcairn, D. N., et al. (1999). Maternal education and measures of early speech and language. Journal of Speech, Language, and Hearing Research, 42, 1432-1443.

Duncan, G., Brooks-Gunn, J., & Klebunov, P. (1994). Economic deprivation and early childhood development. Child Development, 65, 296-318.

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). British Picture Vocabulary Scale: Second Edition. London: NFER-Nelson Publishing Co., Ltd.

Elliot, C. D. (1996). British Ability Scales: Second Edition. London: NFER-Nelson Publishing Co., Ltd.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, 59(5), 1-189.

Gardner, J. M., Powell, C. A., Baker-Henningham, H., Walker, S. P., Cole, T. J., & Grantham-McGregor, S. (2005). Zinc supplementation and psychosocial stimulation: effects on the development of undernourished Jamaican children. American Journal of Clinical Nutrition, 82, 399-405.

Geronimus, A. T., Korenman, S., & Hillemeier, M. M. (1994). Does young maternal age adversely affect child development? Evidence from cousin comparisons in the United States. Population and Development Review, 20(3), 585-609.

Gorman, K. S. (1995). Malnutrition and cognitive development: evidence from experimental/quasi-experimental studies among the mild-to-moderately malnourished. Journal of Nutrition, 125(8 Suppl), 2239S-2244S.

Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., et al. (2007). Developmental potential in the first 5 years for children in developing countries. Lancet, 369, 60-70.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. American Psychologist, 52, 1115-1124.

Greenfield, P. M., & Childs, C. P. (1977). Weaving, color terms, and pattern representation: Cultural influences and cognitive development among the Zincantecos of Southern Mexico. Interamerican Journal of Psychology, 11, 23-48.

Hamadani, J. D., Fuchs, G. J., Osendarp, S. J., Khatun, F., Huda, S. N., & Grantham-McGregor, S. (2002). Zinc supplementation during pregnancy and effects on mental development and behavior of infants: a follow up study. Lancet, 360, 290-294.

Holding, P. A., Taylor, H. G., Kazungu, S. D., Mkala, T., Gona, J., Mwamuye, B., et al. (2004). Assessing cognitive outcomes in a rural African population: Development of a neuropsychological battery in Kilifi District, Kenya. Journal of the International Neuropsychological Society, 10(2), 246-260.

Horowitz, F. D. (1989). Using developmental theory to guide the search for the effects of biological risk factors on the development of children. American Journal of Clinical Nutrition, 50, 589-597.

Hsueh, A. M., & Meyer, B. (1981). Maternal dietary supplementation and 5 year old Stanford Binet IQ test on the offspring in Taiwan. Federation Proceedings, 40, 897.
Hughes, D., & Bryan, J. (2003). The assessment of cognitive performance in children: considerations for detecting nutritional influences. Nutrition Reviews, 61(12), p413-422.
Husaini, M. A., Karyadi, L., Husaini, Y. K., Sandjaja, Karyadi, D., & Pollitt, E. (1991). Developmental effects of short-term supplementary feeding in nutritionally at-risk Indonesian infants. American Journal of Clinical Nutrition, 54(5), 799-804.
Jacq, P. (1998). How many dialects are there? In P. K. Austin (Ed.), Working Papers in Sasak (Vol. 1, pp. 67-90). Melbourne: University of Melbourne.
Kerjasama BAPPEDA Provinsi Nusa Tenggara Barat dengan Badan Pusat Statistik Provinsi Nusa Tenggara Barat. (2004). Penghitungan indeks pembangunan manusia provinsi Nusa Tenggara Barat tahun 2004 (Accounting of human development index in Nusa Tenggara Barat province 2004).
Khandke, V., Pollitt, E., & Gorman, K. S. (1997, April 3-6, 1997). Maternal education and its influences on child growth and cognitive development in rural Guatemala. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Washington, DC.
Kline, P. (1993). The Handbook of Psychological Testing. London: Routledge.
Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. Child Development, 67, 490-507.
Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. Developmental Psychology, 36, 220-232.
Korkman, M., Kirk, U., & Kemp, S. (1998). NEPSY: A developmental neurospychological assessment. Orlando: The Psychological Corporation.
Lind, T., Lonnerdal, B., Stenlund, H., Gamayanti, I. L., Ismail, D., Seswandhana, R., et al. (2004). A community-based randomized controlled trial of iron and zinc supplementation in Indonesian infants: effects on growth and development. American Journal of Clinical Nutrition, 80, 729-736.
Lozoff, B., Brittenham, G. M., Wolf, A. W., McClish, D. K., Kuhnert, P. M., Jimenez, E., et al. (1987). Iron deficiency anemia and iron therapy effects on infant developmental test performance. Pediatrics, 79(6), 981-995.
MacWhinney, B. (2006). The CHILDES Project: Tools for analyzing talk. Electronic edition. Volume 1: Transcription Format and Programs. Part 1: The CHAT Transcription Format.
Murphy, K. R., & Davidshofer, C. O. (1988). Psychological Testing: Principles and Applications: Prentice-Hall International, Inc.
Murray, L., Sinclair, D., Cooper, P., Ducournau, P., Turner, P., & Stein, A. (1999). The socioemotional development of 5-year-old children of postnatally depressed mothers. Journal of Child Psychology and Psychiatry, 40, 1259-1271.
NICHD Early Child Care Research Network. (1999). Chronicity of maternal depressive symptoms, maternal sensitivity, and child functioning at 36 months. Dev Psychol, 35, 1297-1310.
Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

O'Donnell, K. J., Rakeman, M. A., Zhi-Hong, D., Xue-Yi, C., Mei, Z. Y., Delong, N., et al. (2002). Effects of iodine supplementation during pregnancy on child growth and development at school age. Dev Med Child Neurol, 44(2), 76-81.
Oh, S., & Lewis, C. (2008). Korean preschoolers' advanced inhibitory control and its relation to other executive skills and mental state understanding. Child Development, 79, 80-99.
Petterson, S. M., & Albers, A. B. (2001). Effects of poverty and maternal depression on early child development. Child Development, 72(6), 1794-1813.
Pollitt, E. (1993). Early supplementary feeding and cognition: Effects over two decades. Monographs of the Society for Research in Child Development, 58(7), 1-118.
Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1, 385-401.
Rogler, L. H. (1999). Methodological sources of cultural insensitivity in mental health research. Amerian Psychologist, 54, 424-433.
Roosa, M. W., Fitzgerald, H. E., & Carlson, N. A. (1982). Teenage parenting and child development: A literature review. Infant Mental Health Journal, 3, 4-18.
Russell, J., Mauthner, N., Sharpe, S., & Tidswell, T. (1991). The 'windows task' as a measure of strategic deception in preschoolers and autistic subjects. British Journal of Developmental Psychology, 9, 101-119.
Sabbagh, M., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development lf executive functioning and theory of mind: A comparison of Chinese and US preschoolers. Psychological Science, 17, 74-81.
Schaefer, C. E., & DiGeronimo, T. F. (2000). Ages and stages : a parent's guide to normal childhood development. New York: Chichester, John Wiley.
Serpell, R. (1979). How specific are perceptual skills? A cross-cultural study of pattern reproduction. British Journal of Psychology, 70, 365-380.
Seshadri, S., & Gopaldes, T. (1989). Impact of iron supplementation on cognitive functions in pre-school and school-aged children: the Indian experience. American Journal of Clinical Nutrition, 50, 675-686.
Soewondo, S., Husaini, M., & Pollitt, E. (1989). Effects of iron deficiency on attention and learning processes in preschool children: Bandung, Indonesia. American Journal of Clinical Nutrition, 50, 667S-673S.
Southon, S., Wright, A. J. A., Finglas, P. M., Bailey, A. L., Loughride, J. M., & Walker, A. D. (1994). Dietary intake and micronutrient stauts of adolescents: effect of vitamin and trace element supplementation on indices of status and performance in tests of verbal and non-verbal intelligence. British Journal of Nutrition, 71, 897-918.
Strupp, B. J., & Levitsky, D. A. (1995). Enduring cognitive effects of early malnutrition: a theoretical reappraisal. Journal of Nutrition, 125(8 Suppl), 2221S-2232S.
SUMMIT Study Group. (2008). Effect of maternal multiple micronutrient supplementation on fetal loss and infant death in Indonesia: a double-blind cluster-randomised trial. Lancet, 371(9608), 215-227.
Super, C. M. (1981). Behavioral development in infancy. In R. H. Munroe, R. L. Munroe & B. B. Whiting (Eds.), Handbook of cross-cultural human development. New York: Garland Press.
Syahdan. (1996). Sasak-Indonesian code-switching. Unpublished PhD Thesis, University of Arizona.

Tamayo, J. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. Educational and Psychological Measurement, 47, 893-902.
Teeuw, A. (1951). Atlas dialect pulau Lombok [Dialect atlas of the island of Lombok]. Jakarta: Biro Reproduksi Djawatan Togografi.
UNICEF/WHO/UNU. (1999). Composition of a multiple micronutrient supplement to be used in pilot programmes among pregnant women in developing countries. Report of a workshop held at UNICEF Headquarters, New York.
van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 47, 263-279.
Walker, S. P., Wachs, T. D., Meeks-Gardner, J. M., Lozoff, B., Wasserman, G. A., Pollitt, E., et al. (2007). Child development: risk factors for adverse outcomes in developing countries. Lancet, 369, 145-157.
Wechsler, D. (2002). Wechsler Preschool and Primary Scale of Intelligence - Third Edition (WPPSI - III). San Antonio, TX: Harcourt Assessment, Inc.
Whaley, S. E., Sigman, M., Neumann, C., Bwibo, N., Guthrie, D., Weiss, R. E., et al. (2003). The impact of dietary intervention on the cognitive development of Kenyan school children. Journal of Nutrition, 133, 3965S-3971S.
Zhou, S. J., Gibson, R. A., Crowther, C. A., Baghurst, P., & Makrides, M. (2006). Effect of iron supplementation during pregnancy on the intelligence quotient and behavior of children at 4 y of age: longterm follow-up of a randomized controlled trial. American Journal of Clinical Nutrition, 83(5), 1112-1117.