The Early Child Development Instrument (EDI): An item analysis using Classical Test Theory (CTT) on Alberta's data

Vijaya Krishnan, Ph.D.
E-mail: vkrishna@ualberta.ca

Early Child Development Mapping (ECMap) Project
Community-University Partnership (CUP)
Faculty of Extension, University of Alberta
2-410 Enterprise Square, 10230 Jasper Avenue
Edmonton, Alberta T5J 4P6

ECMap
Early Child Development
Mapping Project Alberta

Contents

Overview....3
Looking behind the EDI components....4
Why item analysis? 7
Theorizing item analysis....8
Some basic concepts in Classical Test Theory (CTT)....10
Some key definitions of terms in the CTT model....12
Observed score (X)....12
True score (T)....13
Measurement error (E)....13
Variability....13
What to look for in a CTT-based item analysis?....15
Item difficulty....15
Item discrimination....17
The item-test correlation....17
The item-rest correlation....17
The distractor-test correlation....18
Corrected Point-Biserial Correlation (CPBC)....19
Graphical item analysis....20
Reliability....21
Internal consistency reliability....22
Standard Error of Measurement (SEM)....24
Understanding the properties and reliability of EDI items....25
Basic statistics....26
Item difficulty (*p*-values)....31
Item discrimination....35
Distractor-test correlations....39
Graphical item analysis: DIF....44
Reliability....66
More on graphical analysis: Item difficulty and discrimination coefficients....71
Standard Error of Measurement (SEM)....75
Cross-area comparisons of perfect scorers....76
Discussion and Conclusions....79
References....83
Acknowledgements....86

Overview

The Early Development Instrument (EDI) is a tool to assess kindergarteners' development in the five areas of development: physical health & well-being, social competence, emotional maturity, language & thinking skills, and communication & general knowledge. The tool is designed to be universal enough to be relevant to most preschoolers around the world, allowing an assessment, an overview of the five key areas with no component of screening, yet constructed from the perspective of a Eurocentric epistemology. The multidimensional EDI is geared to provide a methodology and a framework for communities to effectively address developmental difficulties in children at a macro-level. Specifically, the EDI is a survey-based thematic tool primarily designed to assist and target communities at a local level, although data are collected at an individual level.[1]

Since its development in 1999 by the Offord Centre for Child Studies at McMaster University, researchers have been reporting the psychometric characteristics of the 103 items that make up the five areas, in terms of validity of the content and construction, and reliability (e.g., Forer & Zumbo, 2011; Hymel, LeMare & McKee, 2011; Janus, Brinkman & Duku, 2011). However, to understand the whys behind the EDI's ability to fit any context or to expand the scope of epistemological development, users must look behind the five components so that more explorations of the interaction between environment, cultural context, and epistemological development are possible if it is to be implemented to all population groups.

The EDI is being utilized in a growing number of countries and all provinces and two of the three territories within Canada. However, this perhaps is the only documentation of the reliability, and thereby the validity of specific components of the instrument utilizing the data for the province of Alberta. The pages that follow provides the reader with the analysis of the EDI survey questions, administered by kindergarten teachers across Alberta, through a collaborative effort led by Alberta Education and the Offord Centre. The Early Child Development Mapping Project (ECMap) (formulated in 2009) affiliated with the Community-University Partnership (CUP) at the University of Alberta is responsible for mining the data and develop an inter-community snapshot of developmental patterns of preschoolers. The data for this study cover four waves (2009, 2010, 2011, & 2012), are up to date, and represent 66,990 kindergarten children.

[1] The EDI developers make it all clear that the data can be aggregated so that it can be analyzed at the neighbourhood, sub-community or group level (based on age, sex, or ethnic characteristics).

Central to our effort in assessing the indicators of the five developmental areas is the theory upon which the analysis is based. In order to set the stage adequately for a non-technical reader, an approach I thought would work better is to first make the hammers and saws underlying the theory more handy before pursuing the theory itself. Consequently, the report provides an overview of the theoretical rationale for using the approach of Classical Test Theory (CTT) with the key concepts that envision it with not so much of a discussion of its development and the pros and cons. The subsequent sections of this report address the CTT with a description of the approach and the results of analysis of the EDI tool within the framework of CTT. The report concludes with an examination of the EDI's problematic items and the steps to be taken in order to better serve the diverse communities in the province of Alberta.

What is your takeaway?

Early Development Instrument (EDI) is designed to capture five areas, fundamental to young children's development. The survey questions are broad enough to be applicable to preschoolers in most contexts. Regardless, the survey questions need to be examined through a lens of equality and quality, since its ability to make comparisons across certain population groups may be questionable.

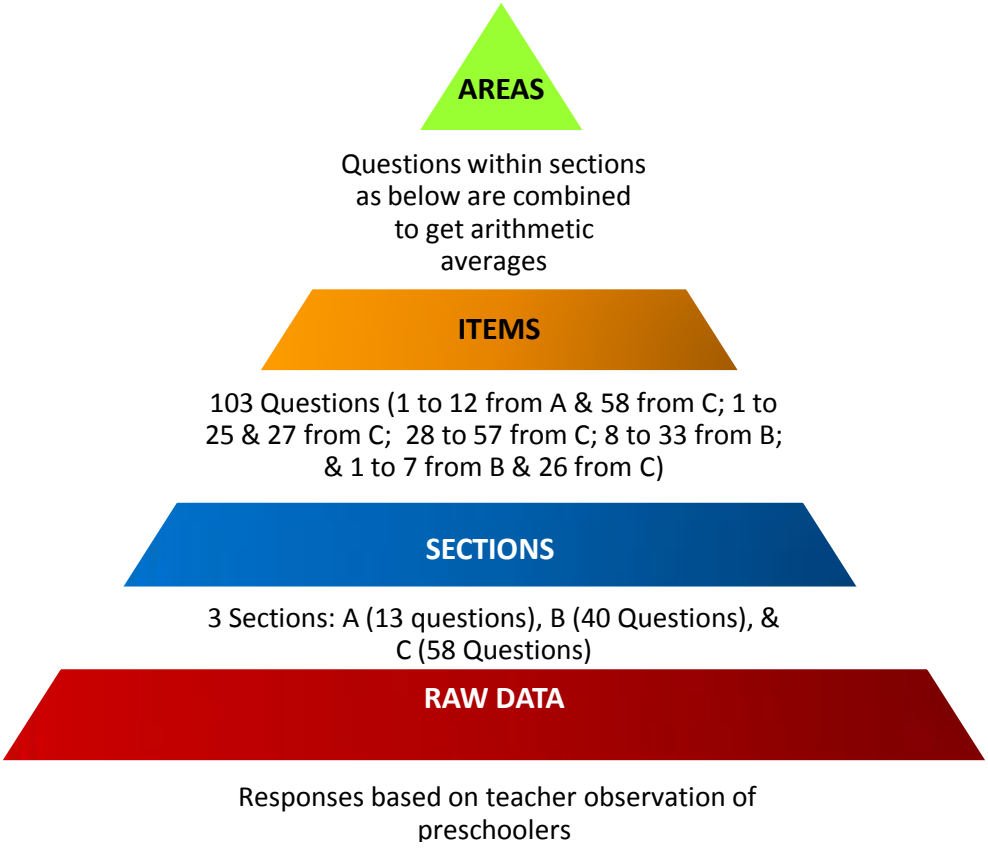
Looking behind the EDI components

The core areas of development along with their constituent parts must be adequately addressed, if assessments based on children's performance have practical implications. Consequently, before attempting to discuss the CTT procedures, a brief explanation of what is meant by item analysis is provided by outlining some of the basic concepts that appear in our discussions. However, before doing so, it is appropriate to inform readers how the teacher responses to survey questions are turned into numerical values or how the component scores are built. Figure 1 visually illustrates the steps to aggregation into component scores.

Once the data are collected through the surveys, they are checked following a rigorous quality control process. In order to arrive at one pool of data from multiple waves, they are merged keeping the original codes intact. The three sections of the questionnaire, A, B, & C provide the information to create component scores. Items from each of the three sections (in different combinations) are then themselves combined in order to yield the component values, assigning equal weights to individual items. In order to make information gathered from the 103

questions interpretable, simplification is necessary. This is where a composite construction becomes important.[2] It should be noted that, although the complexity of processing 103 items can be minimized if the focus is on components, rather than individual questions, moving up the pyramid to get to the composites introduces aggregation bias or the resolution is affected to some extent, adversely. The five component scores form the last step of aggregation, or they are not aggregated any further because the five EDI developmental areas are conceived as thematic.[3]

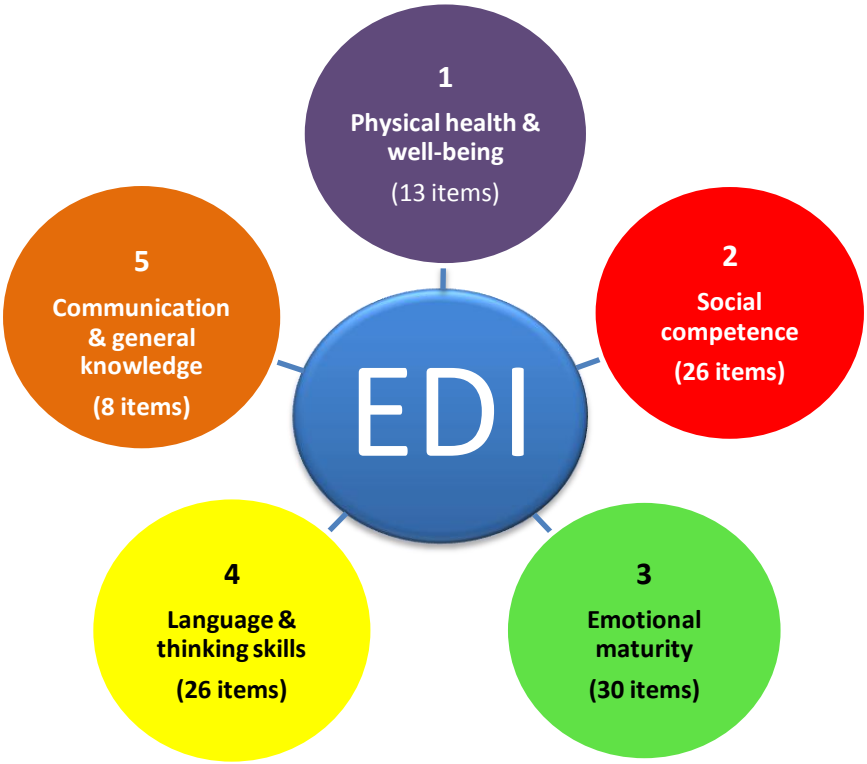
Figure 1: How EDI survey data are converted into five developmental area scores?



[2] A composite is an amalgamation of different questions that seeks to represent the individual questions. The most well-known of all composites is the Human Development Index (HDI).
It is important to recognize that there is a trade-off involved when we aggregate data, the more we aggregate, the more resolution is lost (Cohen, 2009). For example, the creation of the social competence composite from 26 questions by combining them all into one entity by using equal weights would essentially result in the individual influence of each question being irrelevant. That is, the average does wash away any differences in the values of individual question. However, this is unavoidable, especially when the patterns and trends need to be monitored.
[3] The five areas are themselves composites meant to represent five different constructs, based on the theoretical rationale upon which the tool was built. Therefore, a blend of multiple dimensions to yield a single numerical value was considered inappropriate (see, Cohen, 2009 for a discussion of thematic versus composites).

In summary, the EDI consists of five developmental areas with 103 questions to be answered on all five with 13 on physical health & well-being, 26 on social competence, 30 on emotional maturity, 26 on language & thinking skills, and eight on communication & general knowledge. The 103 questions associated with the five areas are referred to as items, in all our discussions. Thus, the composite of physical health & well-being is associated with 13 items, social competence with 26, and so on. The five components are presented in Figure 2 below. The order is not intended as a ranking of the components, but they are usually presented in this sequence.

Figure 2: The EDI components



What is your takeaway?

Since development is multifaceted, EDI's five areas are theorized as five different constructs. They are built into five single values representing different combinations of 103 questions on the survey.

Why item analysis?

"Never assume the obvious is true" William Safire

Item analysis broadly refers to the specific methods used to evaluate items on a test[4], both qualitatively and quantitatively, for the purpose of evaluating the quality of individual items. The goal is to help its developers to improve the instrument by revising or discarding items that do not meet a minimally acceptable standard. The qualitative review is essential during item development and involves experts who have a mastery of relevant material. Test review boards and content experts cannot always be equipped with the knowledge they require to identify "bad" or "defective" items because of such factors as the multidisciplinary nature of the test content and the demographic characteristics of test takers. The statistical analysis could help to identify problematic items that may have slipped the experts' attention, one way or the other. Thus, the quantitative analysis is conducted after the test/tool has been administered to the test takers. The objectives of both the qualitative and quantitative assessments remain the same – to assess the quality of items.

It is of critical importance to realize that there are numerous reasons why an item may fail to meet the minimum standard of quality, whatever the set standard is. Generally, they could come from: (1) the flaws in the question and (2) the flaws in the instruction of the content. More specifically, items can be problematic due to one or more of the following reasons:

- Items may be poorly worded causing test administers to be confused.
- Items may represent a different content area than that is measured by the rest of the items within the same area.
- The presence of bias in an item for or against a sub-group of the population (e.g., ethnic bias).
- The overall ability of the test taker in understanding the true meaning of an item, increasing the odds of guessing the correct answer.

[4] The term test here refers to a set of items that produces a total score for a specific developmental area (e. g., physical health & well-being area).

What is your takeaway?

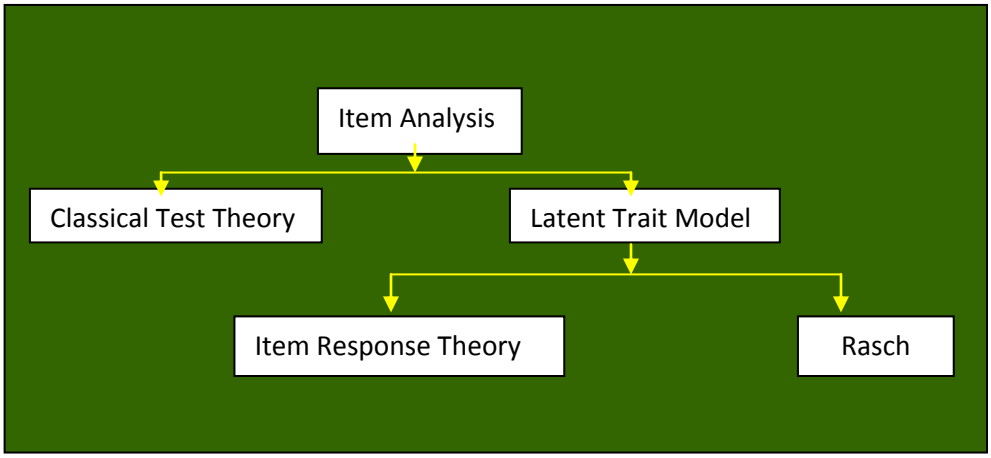
The EDI was developed to function as a reliable measure of population differences in developmental patterns of kindergartners and is the most widely used tool to-date in Canada. Accordingly, there has been emerging interest, not only to research about it, but also to assess it psychometrically. The assessment is important for its wider application so that the homogeneity of the population cannot become a prerequisite for its reliability.

Theorizing item analysis

The presence of bias in test items is a cause for concern to test developers and education researchers. Although diversity (e.g., ethnic or cultural) in populations is an inevitable phenomenon, if some items in a test function differently for a specific subgroup than the majority of the population being tested, direct comparisons of their performance on the items make no sense. This led to many item bias detection procedures or Differential Item Functioning (DIF) to flag for possible item bias.[5] There are two popular DIF detection methods, namely classical and latent trait models.

Classical Measurement Theory (CMT) or Classical Test Theory (CTT), Item Response Theory (IRT), and Rasch (identical to the most basic IRT model (IRT1)) are some of the tools available to identify the quality of items in a test (Box 1). It is beyond the scope of this report to provide a detailed discussion of the merits and demerits of these approaches, and readers may refer to the work of Croker & Algina (1986) for a better understanding of the theoretical and practical bases of CTT and IRT. However, the key differences between CTT and IRT approaches are outlined below.

[5] The term, item bias has been replaced by Differential Item Functioning (DIF) by recent researchers to reflect different empirical methods that relate to biases that are more statistical in nature.



Box 1: Theoretical frameworks for item analysis

IRT is relatively new, but CTT has been in use since the early 20th century. Whereas the key in CTT is the true score on a particular test, in IRT, the concept to be measured (e.g., emotional maturity) is the key (Council of Europe, 2004). IRT considers a score to be the direct result of the true score plus error. IRT methods are considered more powerful than methods based on CTT, but the use of IRT requires a lot of technical know-how. CTT is easier to explain and requires smaller sample sizes (100 or less) than IRT (Pope, 2009). A major shortcoming of CTT is that it is sample-dependent, meaning the statistics generated are not generalizable to similar populations taking a similar survey; they are applicable to only those test takers taking that and that survey only. An important advantage of IRT is that the estimates can be used to tailor an exam in terms of a person's ability. The tests get modified based on the test takers' ability. That is, first a person will be tested on a question of average difficulty, if he/she gets it correct then a harder question is given, and if the response is incorrect this time, then an easier question is given, and the process continues till it becomes satisfactory to the developer. However, with the exception of such test situations as Computerized Adaptive Testing (CAT), it is not highly practical to use estimates to tailor a test, especially when a large-scale population survey is involved.

Typically, both CTT and IRT have been used as standard methods of item analysis. Despite the claim that IRT is theoretically "superior" to CTT, CTT-based item discrimination indices have been found "enough", especially in the medical field, in flagging weak items. Based on findings from various studies that compare the two approaches, the bottom line is: despite its attractive feature, namely the capability of analyzing the unobservable (latent) variable, the results are very similar and the frameworks are quite comparable (Fan, 1998; Lord, 1980; Thorndike, 1982; Stage, 1998; 2003). There are also instances where CTT was found better than IRT in detecting item quality. In particular, a Swedish researcher, basing her work on data from the Swedish

Scholastic Aptitude Test (SweSAT), found CTT to work better than IRT (Stage, 2003). According to Lord (1980), IRT supplements rather than contradicts CTT. CTT is not a modern method, but it necessarily can meet the objective with little sophistication. An examination of whether or not the methods are radically different or they would yield similar results is beyond the scope of this study. The approaches discussed in this report have stemmed from CTT. Psychometric evaluation of the EDI was also given attention by applying two machine learning techniques, namely Artificial Neural Networks (ANN) and Natural Language Processing (NLP) elsewhere by Hollis & Krishnan (2012). The two procedures were chosen with this major purpose: to enhance community feedback on the EDI.

What is your takeaway?

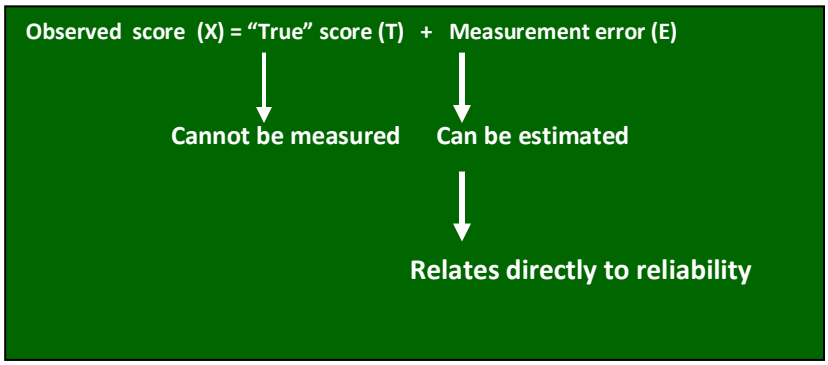
Classical Test Theory is not a modern approach to item analysis, but goes to great lengths to make sure we reach the finish line with easy, but readily available statistics that are sample-dependant; they can apply **only** to that group of respondents and **only** on those items/questions.

Some basic concepts in Classical Test Theory (CTT)

Since our focus here is on item analysis based on CTT, it is important to explore the basic ideas involved in order to fully understand the approach. CTT as a body of theory and research could predict or explain the difficulty of questions, provides insights into the reliability of test scores, and helps us toward coming up with an assessment of how to improve the test by maintaining and developing a pool of "good" items from which future assessments can be drawn. Thus, particular attention is given to individual items, item characteristics, the probability of answering items correctly, the overall ability of the test taker, and the extent to which an item conforms with the rest of the items in a test.

Mathematically, CTT is based on the premise that the observed score from a psychological testing is composed of an un-measurable "true score" and error (Box 2). It is the error that is the most essential component of the equation. Error is inherent in almost all measurement devices we can think of. To give an example, a weighing scale may be accurate to within 0.25 pounds 80 out of 100 times, and test score may be accurate to within 0.10 units 8 out of 10 administrations. The standard deviation of the distribution of random errors (called the standard error of measurement) around what we intend to measure (the true score), if low, it

will cluster around the true score. Therefore, an indicator of the degree of goodness of a single item is its standard error of measurement.



Box 2: CTT-based relationship between observed scores and true scores

It is worth giving a little more thought around the theory of true scores and error scores, using some examples from the EDI. A child's observed score is made up of his/her true score plus measurement error. If I were to administer a multiple choice question from the EDI with three possible response categories (0/5/10) myself to a kindergartner every day for 50 days (i.e., administering the question to the same child multiple times to produce a sample size of 50), my score for the child may sometimes be 0, sometimes 5, and sometimes 10, depending upon some external factors (e.g., my own mood of the day and the child's own behavior that day) but not of the question itself. After my 50-day observations, I get a distribution of scores and errors. The mean of the score based on my scores for the 50 days would be the best estimate of my true score. The fact that my score has its high's and low's, point to the idea that the true score may be normally distributed and the random errors around my true score may be normally distributed as well. It is important to further realize that the error I can make each time I administer the question has no relationship whatsoever between one another because the conditions under which the question being administered have no fixed pattern. Statistically speaking, the errors themselves are uncorrelated. The same can be said about the correlations between the errors and the true scores; we may not find a pattern. The assumptions regarding the true score and their errors form the foundations of CTT. The assumptions hold if I use 50 children all at once (i.e., administering the question once to 50 children to produce a sample size of 50) or if I use one child to get 50 observations. What follows from this discussion is the essential premise of CTT: If we assume that people are randomly selected, then the true score is also a random variable and the error(s):

- is normally distributed;
- are uncorrelated among themselves
- is uncorrelated with the true score; and
- has the expected value (the mean of the distribution of errors over series of trials) zero.

What is your takeaway?

The true-score formula, $X=T+E$ is the heart of CTT, or there is this assumption that no score is error-free, errors come from many directions: uncontrolled testing conditions (e.g., distractions and differing context), random fluctuations in individual performance, etc. With the assumptions: (1) the errors are normally distributed (the scores have highs and lows); (2) the errors have no systematic pattern to why scores fluctuate; (3) the errors are unrelated to the true score (it can be positive or negative), and (4) having a normal distribution, the mean of the distribution of errors over an infinite number of trials is zero. CTT models the random errors of raw scores, but not as such, the systematic errors that may be attributed to changes in scores due to better training or experience.

Some key definitions of terms in the CTT model

The CTT model is based on the notion that the observed score that test takers obtain from a test is composed of a theoretical un-measurable true score plus some measurement error. The true score model can be expressed as*: $X=T+E$.* Let's examine, through the use of examples, what these letters stand for.

Observed score (X)

This is the raw score obtained by any one individual in a test. The summary of the results based on each of the 103 questions/items in EDI is represented by a number. The number associated with each item, is the observed EDI item score. Since we have 103 items and five composites, we have 103 observed item scores and five composite scores for each child. Thus, as earlier stated, the observed score for the physical health & well-being area is the sum of all 13 item scores, for the social competence area, the sum of all 26 item scores, and so on. In all analyses to be carried out in CTT, the observed item scores, not the composite scores are the quantities that enter in analyses.

True score (T)

The basic assumption of CTT is that a second administration of the same set of questions to the same child under similar circumstances as the first time will yield the same score as the first time. If a number of such administrations are carried out, a child can have a distribution of scores for his/her own. The mean of this distribution is the child's estimated true score. The true score is a hypothetical score that is unobservable. This is similar to comparing a statistic with a parameter; a sample statistic is known, but the population parameter is unknown, unless we have a way of surveying the entire population.

Measurement error (E)

The measurement error refers to the difference between the observed score (X) and the true score (T). If the observed score is greater than the true score, the measurement error is positive and if the observed score is smaller than the true score, the error is negative. Since the true score is unknown, the measurement error is unknown, and can only be estimated. Let's introduce a term, called variability to explain how the measurement error is related to consistency of scores or responses provided by a test taker.

Variability

Just as a blood pressure monitor cannot give the same blood pressure levels, due to differences in a person's energy, anxiety, or other conditions, variation in test scores is unavoidable. If there is no variability among scores, how informative can the test result be? If, for example, all 10 children taking a test score o where o being the incorrect response, all we can say is, the test is too difficult. Similarly, if all score 10 where 10 being the correct response the test is too easy for them. In simple terms, if the scores all hit either the floor or ceiling, the item restricts variance. In real life, it is not possible to get all wrong or all right for the kind of questions that comprise the EDI.

The most commonly used measure of variability is the standard deviation with variability being the square of it. Basing on the equation, X=T+E, the variability of the observed score is the variability of true score plus error variance. Theoretically, then reliability is the ratio of true score variance (unknown) to observed score variance (known). In terms of the error variance, it can be expressed as: (1-the ratio of error variance to observed variance) (Box 3).

$$\text{Reliability} = \frac{\text{True score variance (Unknown)}}{\text{Observed score variance (Known)}}$$

In terms of error variance, reliability can be estimated as:

$$\text{Reliability} = 1 - \frac{Error\ variance}{Observed\ score\ variance}$$

Box 3: CTT-based relationship between reliability and observed score and error variances

An important question at this point is: how the measurement error directly relates to reliability or the consistency of test scores? If I administer a series of questions an infinite number of times and I get the mean scores very close to one another, the better is the reliability in my measurement. It follows that, in terms of the known observed score, if I expect a high reliability for my observed score, its variance should be small or the error variance should be small (Kline, 2005). For practical purposes, internal consistency estimates measure the extent to which each item correlates with every other item. It is measured on a scale of 0 to 1; the value 0 means, all the variations in the observed score are due to measurement error and the value 1 means there is no measurement error. Thus, a value of 0.90 means that 90% of the observed score variance can be attributed to variation in the true scores and the remaining 10% can be attributed to measurement error. The reliability of test scores (not the reliability of a test as such) is a key concept in CTT, and it will be dealt with in our later discussions, as and when appropriate.

With some knowledge of the key terms in CTT, we can now turn to a discussion of the statistics that can be used to assess the performance of individual test items on the assumption that the overall quality of a test depends upon the quality of the items. In order to assess the statistical properties of the EDI items, both psychometrically appropriate and practically less complex, several analyses are available within CTT. In particular, the following aspects of CTT are explored here, as they are available through SPSS and/or Excel.

What is your takeaway?

Item analysis is concerned with examining responses to individual test items/questions to assess the item quality. The goal of achieving quality means minimizing the measurement error in scores. By using the internal criterion of test scores, item analysis presents such statistics as reliability coefficient to check for the internal consistency of items, which is also a first step in achieving the validity of test items.

What to look for in a CTT-based item analysis?

Drawing on the important key concepts at a theoretical level, we explore the essential things to look for in a typical item analysis based on CTT.

- Item difficulty or *p*- values
- Item discrimination
    - Item- test correlation
    - Item-rest correlation
    - distractor-test correlation
- graphical item analysis
- reliability
- Standard Error of Measurement (SEM)

Item difficulty

As earlier noted, items that are correctly answered by every person or that are wrongly answered by every person do not convey any message about individual differences in performance. Also, not all kindergarteners are likely to get all the items correct or wrong.[6] So the question is: how easy or difficult is the item? Let us explain difficulty using an example, I want to assess the concept extroversion by asking 50 teenagers using a series of questions that can be answered with "yes"/"no". I use an example of this sort, "I enjoy the company of people even when I don't know anyone." A useful statistic, based on all different responses is the proportion of teens who endorse the question, or its difficulty level. If 20 teens answered the question "yes", the difficulty level is calculated as: 20/50=0.40. The range to which difficulty can fall is 0%-100%, or written as a proportion of 0.0 to 1.00. The higher the value, the easier the item or lower the difficulty.

For a dichotomously coded item with two response categories, say 0 and 10 (as in Qa2 on the survey), the difficulty index, indicated by *p*, is calculated as the ratio of the number of persons

[6] Note: Yes/No responses are interchangeably used as Right/Wrong in our discussions in order to inform that the largest number always corresponds to the correct response or better outcome because some questions on the survey were reversely coded.

who answer the item correctly to the total number of test takers. When an item assumes more than two values (referred to as a partial credit item) as in Qa9 with response categories 0, 5, and 10, the *p* values are computed as the average relative item score (Box 4). An example each of a binary and a partial credit item is shown on Table 1.

For binary items (e.g., scored 0/10), the *p*-value of an item is calculated by the formula:

$$p_i = \frac{A_i}{N_i}$$

Where: p_i =Difficulty index of item *i*

A_i=Number of correct answers to item *i*
N_i=Number of correct answer plus number of incorrect answer to item *i*.

Box 4: Computation of item difficulty (*p*) for binary items

Table 1: The *p*-value calculations for binary and partial credit items, Merger #3, Alberta (N=52,035)

Item	Values*	0	5	10	*p* (%)
Qa2:dressed inappropriately	0=Yes; 10=No	4,694	0	47,312	90.97
Qa9:Proficient at holding pen	0=Poor/very poor; 5=Average; 10=Very good/good	4,506	19,002	28,498	73.07

The *p* for Qa2 is: (4694x0+47312x10)/(4694+47312)=0.9097

The *p* for Qa9 is: (4506x0+19002x5+28498x10)/(4506+19002+28498)=0.7307

The item, Qa2 has a p value of 0.91 – that is, 91% of the group got the item correct – provides the lowest level of differentiation between children for that item. This means, if there were 100 children responded to the item, then there will be 91x9 =819 differentiations made by that item, as each child who got the item correct is differentiated from each child who got the item wrong. In contrast, the item Qa9 with a p value of 0.73 will provide 73x27=1971 differentiations for that item. Items with p values closer to 0.50 are considered more useful in differentiating between individuals (Kline, 2005) and we will discuss the notion of optimal level of p below.

For maximizing variability and thereby reliability, the optimal item difficulty values can be calculated. The ideal value can be slightly higher than midway between chance (1.00 divided by the number of choices) and a perfect score (1.00) for an item.

For a dichotomous item with two response options as in Qa2,

The random guessing level = 1.00/2 =0.50

The optimal difficulty level= 0.50+ (1.00-0.50)/2=0.75.

For a three-alternative, multiple-choice item as Qa9,

The random guessing level= 1.00/3=0.33

The optimal difficulty level= 0.33+ (1.00-0.33)/2=0.67.

A general recommendation is to use items with p values within a range of 0.40 to 0.60 (with an average of 50% getting the item correct). If half of the group gets the item correct and half gets it wrong, whether or not we use a single item or a series of items, the end result can be the same in differentiating between the groups. In other words, it is better to create items of varying difficulty with an average p value around 0.50 (Ghiselli, Campbell, & Zedek, 1981). Items with p values above 0.90 and those below 0.20 warrant a careful evaluation, whatever the criterion one may use in determining a reasonable estimate for p.

Item discrimination

The term, item discrimination index, stands for the difference between the percentage of high performers and the percentage of low performers. First, using the p value, those who have the highest and lowest overall test scores are grouped into upper and lower groups. The upper group is made up of the best performers and the lower group is made up of the poorest performers. Most researchers use the upper 27% and the lower 27%, as they separate the tail from the mean of the standard normal distribution (Cureton, 1957). Second, determine the p values for each item for the two groups. Third, calculate the difference between the p values for the two groups. The higher the difference, the more the item discriminates. Items with p levels at 0.50 are often noted as having the highest discrimination values.

What if the item has three response categories? If the item separates the very best performers from the worst performers, it cannot separate those in the medium performance levels from the best or worst ones. This is where the correlation analysis becomes important. Within CTT, we may use:

The item-test correlation: To what degree do an item score and the total test score measure the same thing? The item-test correlation gives the strength of the relationship between an item score and the test score. If it is positive, the item discriminates between high and low scores, if 0, the item does not discriminate between high and low, and if negative, item scores and test scores disagree. If a child does well on an item, he/she is expected to do well on the test as a whole. In other words, item-total correlations are related to reliability because the

better each item correlates with the test as a whole, the greater the likelihood that all items correlate with each other.

The item-rest correlation: To what degree does an item score and total test score, without the item, measure the same thing? The item-rest correlation shows the strength of a relationship between an item score and the score on the test without that item. Like the item-test correlation, if it is positive, the item discriminates between high and low scores, if 0, the item does not discriminate between high and low, and if negative, item scores and test scores (without the item) disagree.

The distractor-test correlation: If an item has three answer options, 0=wrong, 5=sometimes, and 10=correct, how popular is 0 as an option (distractor wrong)? That is, what proportions pick the distractor 0 (wrong) when the correct answer is 10 (correct)? The distractor-test correlation is the correlation between the distractor and test score in multiple choice questions.

Of the three types of correlations, the first two should be positive and the third one should be negative. Therefore, distractor-test correlations are used to flag items if the correlation is positive. The notion of distractor analysis is given further consideration with the help of an example below and then we will introduce a concept, called point-biserial correlation coefficient in understanding the relationships between the score on an item or the score on a distractor against the score of the whole test (all items in the content area) .

The computation of correlations, in general, requires two series of data. For example, in item-test correlation, one is the array of item scores and the other is the array of test scores. With the help of the two arrays, the correlation coefficient can be computed using the usual formula for a product-moment correlation or Pearson correlation. It is common knowledge that the computation can only fail if there is no variance in the item score.

Distractors refer to all the available options in multiple choice questions: 10/0 – “yes”/ “no” for binary items, such as Qa2 and 0/5/10 – “poor (very poor)”/”average”/” very good (good)” for polytomous items, such as Qa9. To compute the correlation between a distractor “very good/good” and the test score containing Qa9, the response-categories need to be recoded first, as: 0= poor/very poor; 0= average; 10= very good/good, since 10=very good/good being the correct answer. To compute the correlation between a distractor, say “poor/very poor” and the test score, the item is recoded as: 10=poor/very poor; 0=average; 0=very good/good. Similar recodes apply for the other distractor, namely average; the category is recoded as 10 and the rest as 0.

The correlations between the distractors and test scores can either be positive or negative, depending upon whether or not the distractors are the correct or wrong answer ones. In other words, whereas those who selected the correct distractor are likely to score higher on the test,

those who selected the incorrect one are likely to score lower on the test. If the correlation between a wrong answer distractor and the test score is found positive, it may be an indication that it is an item that is confusing.

Corrected Point-Biserial Correlation (CPBC)

As noted earlier, to assess item discrimination, the correlation coefficient can also be used. Item discrimination indicates the relationship between how well children did on the item and their total test score. How do responses to an item relate to the total test score? The influence of an item on a test score with only eight items (as in communication & general knowledge) can be greater than on a test score with 30 items (as in emotional maturity). A correlation that is particularly important in such a situation is the Corrected Point-Biserial Correlation (CPBC) coefficient. CPBC is the correlation between the right/wrong scores that children receive on a given item within, say physical health & well-being and the total scores that the children receive when summing up their scores across the remaining items in physical health & well-being. To put it simply, CPBC is the correlation between an item and the rest of the test, **without** that item considered part of the test.

If an item is uncorrelated with the rest of the items, it does not contribute to the internal-consistency of the total score. This means, if an individual item is in good conformity with the rest of the items in the area, its CPBC should be a high positive number. In other words, if a child performed well on an item, he/she is expected to perform well on other items in the same area. As in all correlations, CPBC values range from -1.00 to +1.00. A low CPBC value indicates that a child who gets the item correct tends to perform poorly overall and vice versa. Low or negative CPBC coefficients may result from among other things, poor item wording, small sample, small number of items, or the multidimensional nature of the content. Point-biserial correlations can be computed using Excel or SPSS. Since the SPSS computation can be unfamiliar to many readers, the syntax is presented below (Box 5).

```
RELIABILITY
  /VARIABLES=Qa2 Qa3 Qa4 Qa5 Qa6 Qa7 Qa8 Qa9 Qa10 Qa11 Qa12 Qa13 Qc58
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /STATISTICS=DESCRIPTIVE SCALE
  /SUMMARY=TOTAL.
```

Box 5: The syntax for the calculation of CPBC in SPSS for physical health & well-being items

What is your takeaway?

Item statistics can be used to determine if an item is useful and how it performs in relation to other items or the whole test. They include, item difficulty (the proportion who answer correctly), item discrimination (e.g., item-total correlations and distractor-test correlations – how responses to each item or distractor correlate to the *corrected* (excludes the responses to the item) total score on the test – and graphs. An item will have low discrimination if it is difficult to guess it or most get it wrong or easy to guess it or most get it correct.

Graphical item analysis

Another way to judge the quality of items has to do with graphs. This is the pictorial depiction of the characteristics of a particular item, in relation to a homogenous group of test takers. First, the total number of individuals is split into a number of homogeneous groups based on their test scores. Second, the proportion of correct responses is computed for each group. Finally, the proportions are plotted against the group membership (e.g., 1=group with the lowest scores, 2=group with intermediate scores, and 3=group with the highest scores). In the graph, the group membership status (1, 2, & 3) is represented on the horizontal axis and the proportion of correct responses is represented on the vertical axis for each item.

This is probably the right place to introduce a well-known concept, called Differential Item Functioning (DIF), referred to as item bias. The ideal of fairness requires that an item does not favor any particular population. For example, in a fair test, the average scores for boys and girls should be the same. DIF analyses provide information on whether or not an item functions unfavorably across gender, linguistic background, or similar characteristics. Applying sex as an example, an item shows no DIF if boys and girls exhibit the same level of proficiency or when p values of the item are identical. By using p as a proxy for the level of proficiency, by computing the value of p for each item for each group and for each sexes and plotting them on a graph, some evidence for DIF can be determined in terms of sex.

What is your takeaway?

Graphs showing item difficulty among different test takers grouped by different characteristics (e.g., sex) according to their proficiency levels add additional information about the item than just tables of item difficulty and item discrimination.

Reliability

"Measure seven times before making a cut" Bulgarian proverb

From our discussion earlier, it becomes clear that the approach to calculating reliability exactly depends on estimation of true score and measurement error. Thus, we can only estimate reliability, and the measures thus obtained cannot be perfect. There are four general ways to estimate reliability, each of which has its own strengths and weaknesses. They are:

- Test-retest reliability, used to assess consistency of a measure by administering the same test to the same sample on two different occasions
- Parallel forms reliability, used to assess the consistency of the results of two tests constructed in the same manner from the same content
- Inter-rater reliability, used to assess the degree to which different raters give consistent results of the same content
- Internal consistency reliability, used to assess the consistency of results across items from a single administration of a test

The two methods that are commonly used in CTT are parallel testing and retesting. Two tests are said to be parallel if the true scores on both the tests are equal for all persons and the error variances are equal as well. Parallel testing can be a challenge because: (1) two administrations are required on the same sample of test takers; (2) the construction of the parallel forms themselves is difficult; and (3) the examination of them as to whether or not they are parallel is difficult. In the retesting situation, the same test needs to be applied at two different points in time. The amount of time allowed between measures is very critical here and the chances are that the longer the time gap, the greater the difference in estimates. How do we determine whether two raters/observers produce consistent observations and results? A number of factors impact the uniformity of judgments, the raters' familiarity, the amount of training they receive, the test length and format, to name a few. The sources of inconsistency between raters may be due to (Kaftandjieva, 2004): (1) different conception of mastery; (2) different interpretations of levels of proficiency; and (3) the raters' own different value systems. Since it is beyond the scope of this research to use either one of them, we won't be discussing the methods any further.

In discussing the choice of a procedure for estimating reliability, it should be mentioned that there is no best strategy that fits all circumstances. It is simply a matter of choice and depends entirely on the circumstances and the complexity of the procedure, including the complexity of the test. In principle, it is possible, although not perfect, to estimate the reliability from a single test administration. This is where we do a testing for internal consistency. There are several internal-consistency methods that require only one administration of the test and are not

rigorous. However, all that can be achieved is a lower bound to the reliability in terms of such coefficients as Cronbach's alpha for any mixture of item format (dichotomous or polytomous) and KR20 coefficient for only dichotomously coded items. In any case, the internal consistency reliability estimation should be explored further, since we only have a single administration of the test.

Internal consistency reliability

How do we estimate reliability if the test is administered only once? Internal consistency reliability estimation is based on a single test administered to a group of individuals in one occasion. It refers to the degree to which the items that make up the concept of interest are measuring the same underlying concept. There are different internal consistency measures that can be used. The most commonly used is the Cronbach's coefficient alpha that provides an indication of the average correlation among all items that make up the test. For example, if we have 13 items as in the physical health & well-being area, we will have 78 (=13! /11! x 2!) different correlations, and the mean of all these correlations is an estimate of the alpha.[7] Put it simply, reliability is estimated by computing the correlation between scores on every pair of test items.

As noted earlier, Cronbach's alpha, a lower-bound estimate of reliability can be used for any combination of dichotomous or partial credit items. Cronbach's alpha is sometimes called the index of internal consistency. The alpha values range from 0 to 1.00, with higher values indicating greater reliability. The Cronbach's alpha values are depended on the number of items in the test, the smaller the number (fewer than 10), the smaller the value of alpha (Pallant, 2007). According to Pallant (2007), in such situations, it is advisable to report the mean inter-item correlation for the items. Nunnally (1978) recommends a minimum level of 0.70 for alpha, and optimal mean inter-item correlation values are recommended as ranging from 0.20 to 0.40 (Briggs & Check, 1986; Pallant, 2007). A lower-bound estimate of alpha, say 0.70 means that the reliability can be at least 0.70.

Kuder-Richardson formula 20 (KR-20) is a frequently used method for determining internal consistency if the items are dichotomously coded.[8] Basically, the computation requires three pieces of information, namely the number of items, the mean, and the standard deviation.

In reality, most tests have one form. An alternative to administering two forms of a test to a group to estimate reliability is to artificially create two forms of the single test. In other words,

[7] In addition, if we compute a total score for the 13 items and use this as another entry in the computation of the correlation, we will get the 13 item-to-total correlation, with the mean of all these as the alpha value.

[8] The computational formula for KR-20 is: $\frac{k}{k-1}(1-\frac{\sum pq}{\sigma^2 x})$, where k=number of items; p=proportion of persons who got the item correct; q=proportion of persons who got the item wrong, and $\sigma^2 x$ is the total score variance.

what if we only have a single test and split this test into two halves which are meant to be parallel and test for internal consistency? This is what a split-half reliability means. A common procedure is to divide items into two groups. Usually, odd-numbered items are placed in one group and even numbered in the other. For instance, if the test contained 13 items as in the physical health & well-being area, one form would be created from the seven odd items and the second form from the six even items. The scores on one half of the test are correlated with the scores on the second half. The correlation between scores on the seven items and the scores on the six items is called the split-half reliability coefficient. The correlation between the two halves is not being the reliability of the test as a whole, but of the half test, we need to correct for split-half computations of reliability. This brings us to a concept called, the Spearman-Brown formula that can be used to compute the reliability for the entire test based on the split-half. The formula estimates the hypothetical correlation of a test assuming that each half of the test had been the length of the whole test (Box 6). In other words, the formula corrects for shortness of a test in a split-half reliability estimate with the assumption that reliability is a function of the test length. It gives answers to such questions as: what is the relation between reliabilities of a test length, say 20, and a test length, say 30? More specifically, the Spearman-Brown formula can be used to generate two types of information of the kind noted below.

- A test has 10 items and has a reliability of 0.60. What will the reliability be if 10 more items of the same quality are added?
- A test has 10 items and has a reliability of 0.60. If we want the reliability to be raised to 0.80, how many more items of the same quality we might need?

The Spearman-Brown formula:

$$\rho(k) = \frac{k\,\rho(l)}{1 + (k-1)\rho(l)}$$

Where: $\rho(k)$ =Reliability of test length k, a positive number

$\rho(l)_l$ =Reliability of test length l, a positive number

Note: k stands for the ratio of a new test length to some existing test length.

Box 6: The Spearman-Brown computation formula

Using the formula, a test with 8 items and a reliability of 0.60, if extended to 24 or three times more items, will yield a reliability of (3x0.60)/[1+(3-1)x0.60]=0.82. On the other hand, if we

want to know how many items should be added in order to get a reliability of 0.80 from a test with 8 items and 0.60 reliability, the computation is:

0.80= (kx0.60)/[1+(k-1)x0.60)]

k= (0.80) x (1-0.60) / (0.60 x (1-0.80) =2.667 or the test should be 2.667 times longer than the present length or 2.667x8=21.3 items.

What is your takeaway?

A reliable test is one we can trust or we can use to measure a person's performance approximately the same way each time. Whether we measure it twice and evaluate stability over time, measure agreement among different scorers, measure two versions of the test with same content, or correlate two halves of a test with a homogeneous content, we are testing how well the test reflect stability and consistency. Internal consistency estimates, indicating the extent to which each item correlated with every other item, are the easiest and practical of all types. Cronbach's coefficient alpha that evaluates the degree to which different items "pull together" the same content area and the Spearman-Brown formula that uses a statistical correction to estimate the correlation between two halves allow making inferences about reliability.

Standard Error of Measurement (SEM)

The Standard Error of Measurement (SEM) is a number that indicates the accuracy with which an individual's score approximates the true score for the same individual. It is the error expected in an individual's test score. SEM is directly related to the reliability of the test. If it were possible to administer an infinite number of parallel tests, a child's score would be expected to change from one administration to the next for a number of reasons. For each child, we thus obtain a distribution of scores, and the mean of this distribution is believed to be the child's true score. The standard deviation of this hypothetical distribution is called the SEM, and it reflects the amount of change one would expect from one administration to another (see, Nunnally, 1967 for a detailed discussion of the standard error of measurement).

In simple terms, SEM can be used to answer such questions: How much would the sample mean vary, on average, over hypothetically infinite number of samples? How would the standard deviation of the sample mean be over such infinite number of samples?

Mathematically, SEM can be computed using sample data (because in reality, we cannot administer infinite number of tests) as follows (Box 7):

$$SE_X = SD_X \times \sqrt{1 - Reliability\ (X)}$$

Where: SE_X is the SEM
SD_X is the Standard deviation of the observed test score
$Reliability\ (X)$ = Estimated reliability coefficient of test score

Box 7: Computational formula for SEM

From the computational formula above, it is safe to say that, the smaller the standard error of measurement, the more reliable the test is. For standardized tests, the computation involves plugging the values of coefficient alpha and standard deviation. It is important to mention that SEM is distinct from standard error of mean (samples or population) and the standard error of estimate (used in predictions). SEM can help better interpret scores and it can be used to calculate confidence intervals (Verhelst, 2004). Let us see what we can say about a child's observed score = 8.45; SEM = 0.40, and the reliability for the test score = 0.95. The observed score is not being perfect, contains some measurement error and the best we can do is to say that the child's true score will fall within an interval. If we assume the error distribution to be normal, we can define a 90% confidence interval as:

Prob (8.45-1.645x0.40<= true score<=8.45+1.645x0.40)
=8.45-0.658<=true score<=8.45+0.658
=7.792~9.108

This means that the child's true score with a 90% confidence interval is: 1.316 units, or we can expect 10 out of 100 times the true score can fall lower than 7.792 or higher than 9.108. By substituting the values 1.96 and 2.58 in the place of 1.645, we will get the confidence intervals at the 95% and 99% levels, respectively.[9]

Understanding the properties and reliability of EDI items

Table 2 presents the items with their descriptions, formats and descriptive statistics. All in all, there are 33 dichotomous and 70 polytomous items. Physical health & well-being area has both dichotomous and polytomous items. Dichotomously- and polytomously-coded items are

[9] A general rule of thumb to predict the amount of difference which can be expected in individual test scores is to multiply the SEM value by 1.5.

analyzed separately depending upon the item analysis statistics (e.g., *p*-values), since the computational methods differ based on the format. The number of questionnaires analyzed included the four waves of data (2009, 2010, 2011, & 2012) totaling to 66,990 children (9,641 in 2009; 21, 976 in 2010; 20, 881 in 2011; and 14, 492 in 2012). Of these, 14,955 questionnaires were removed for all or one of the following reasons:

- Children in class less than 1 month
- Children having special needs (diagnosed disabilities)
- Children whose scores are missing in more than one area
- Children with less than 4 and older than 7 years of age
- Children with no parental consent

This resulted in a sample of 52,035 for our analysis. The children's mean age was found to be 5.67 years, with little or no difference between boys' and girls' mean ages (males: 5.69 years; female: 5.65 years). Boys outnumbered girls only by a small margin (50.73% vs. 49.26%). Only 10 per cent of children had English or French as their second language. This variable has a large number of missing cases (22.2%) with 67.9% reporting that English or French is not their second language. And, Aboriginal children represented only 5% of the sample. However, the variable is based on a family's "self report", and is not based on any official records of ancestry. Due to the small number cases, we will use this variable only at a test level, not item-level. The percentage who repeated grade constituted only 3.26% of the sample. Although we computed the difficulty levels based on this variable, they need to be interpreted with some caution.

Basic statistics

Item-level descriptive statistics (Table 2) indicate that all items, with the exception of Qc30, Qb16, and Qb17, are negatively skewed. That is, scores for 100 items are clustered at the high end (to the right hand side of a graph at the high values). High mean scores and extreme high kurtosis values for more than 50% of the items indicate that the scores are all clustered to the right with no tail on the right and very thin tail on the left. Since many scales in the social sciences have scores that are skewed, either positively or negatively, this may not indicate a serious problem with the scale. Rather, it simply means that a large majority of children in our sample are performing reasonably well. However, the highest levels of skewness and kurtosis occurred on the language & thinking skills items, and 14 out of the 26 items in the area had means above 9.0.

Table 2: The five developmental areas with 103 items and the descriptive statistics of items, Merger #3, Alberta (N=52,035)

Item	Values	Mean	Sd	Se	Skewness	Kurtosis
Physical health & well-being (13)						
Qa2:dressed inappropriately	0=Yes; 10=No	9.10	2.87	0.01	-2.86	6.18
Qa3:Too tired	0=Yes; 10=No	8.91	3.11	0.01	-2.52	4.34
Qa4:Late	0=Yes; 10=No	7.91	4.06	0.02	-1.43	0.06
Qa5:Hungry	0=Yes; 10=No	9.55	2.08	0.01	-4.37	17.05
Qa6:Washroom	0=No; 10=Yes	9.85	1.22	0.01	-7.99	61.77
Qa7:Hand preference	0=No; 10=Yes	9.76	1.52	0.01	-6.29	37.53
Qa8:Well coordinated	0=No; 10=Yes	9.31	2.54	0.01	-3.40	9.56
Qa9:Proficient at holding pen	0=Poor/very poor; 5=Average; 10=Very good/good	7.31	3.25	0.01	-0.81	-0.43
Qa10:Manipulates objects	0=Poor/very poor; 5=Average; 10=Very good/good	7.85	2.82	0.01	-0.89	-0.22
Qa11:Climbs stairs	0=Poor/very poor; 5=Average; 10=Very good/good	8.11	2.60	0.01	-0.88	-0.44
Qa12:Level of energy	0=Poor/very poor; 5=Average; 10=Very good/good	7.71	2.89	0.01	-0.82	-0.32
Qa13:Overall physical	0=Poor/very poor; 5=Average; 10=Very good/good	7.88	2.69	0.01	-0.74	-0.61
Qc58:Sucks thumb	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.74	1.36	0.01	-5.65	33.14
Social competence (26)						
Qc1:overall soc/emotional	0=Poor/very poor; 5=Average; 10=Very good/good	7.23	3.13	0.01	-0.68	-0.52
Qc2:gets along with peers	0=Poor/very poor; 5=Average; 10=Very good/good	7.49	2.99	0.01	-0.75	-0.42
Qc3:cooperative	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.44	2.52	0.01	-1.27	0.54
Qc4:plays with various children	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.33	2.64	0.01	-1.27	0.61
Qc5:follows rules	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.46	2.49	0.01	-1.26	0.48
Qc6:respects property	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	9.07	2.09	0.01	-2.06	3.38
Qc7:self-control	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.31	2.67	0.01	-1.27	0.63
Qc8:self-confidence	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	7.83	2.88	0.01	-0.94	-0.13
Qc9:respect for adults	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	9.17	1.97	0.01	-2.17	3.79
Qc10:respect for children	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.74	2.31	0.01	-1.51	1.18
Qc11:accept responsibility	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.39	2.70	0.01	-1.44	1.13
Qc12:listens	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	7.68	2.92	0.01	-0.83	-0.30
Qc13:follows directions	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.16	2.69	0.01	-1.08	0.13
Qc14:completes work on time	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.02	2.96	0.01	-1.21	0.44
Qc15:independent	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.1	2.94	0.01	-1.28	0.61
Qc16:takes care of materials	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	9.07	2.10	0.01	-2.10	3.64
Qc17:works neatly	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	7.89	2.97	0.01	-1.08	0.15
Qc18:curious	0=Never/not true; 5=Sometimes/somewhat true;	9.06	2.10	0.01	-2.06	3.42

Item	Values	Mean	Sd	Se	Skewness	Kurtosis
	10=Often/very true					
Qc19:eager new toy	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	9.39	1.71	0.01	-2.70	6.56
Qc20:eager new game	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	9.29	1.85	0.01	-2.51	5.60
Qc21:eager new book	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.91	2.30	0.01	-2.00	3.23
Qc22:independent solve problems	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	7.65	3.04	0.01	-0.92	-0.17
Qc23:follow simple instructions	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	9.15	2.01	0.01	-2.23	4.22
Qc24:follow class routines	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.37	2.68	0.01	-1.39	0.97
Qc25:adjust to change	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.76	2.41	0.01	-1.77	2.26
Qc27:tolerance for mistake	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.5	2.59	0.01	-1.49	1.28
Emotional maturity (30)						
Qc28:help hurt	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	7.01	3.32	0.02	-0.67	-0.62
Qc29:clear up mess	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	6.2	3.59	0.02	-0.39	-1.01
Qc30:stop quarrel	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	4.69	3.72	0.02	0.10	-1.19
Qc31:offers help	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	5.62	3.68	0.02	-0.20	-1.14
Qc32:comforts upset	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	5.76	3.65	0.02	-0.24	-1.10
Qc33:spontaneously helps	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	5.95	3.56	0.02	-0.29	-1.01
Qc34:invite bystanders	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	5.07	3.59	0.02	-0.02	-1.07
Qc35:helps sick	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	5.26	3.73	0.02	-0.08	-1.20
Qc36:upset when left	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.9	2.51	0.01	-2.24	4.17
Qc37:gets into fights	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.46	1.73	0.01	-3.35	11.31
Qc38:bullies or mean	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.17	2.08	0.01	-2.48	5.62
Qc39:kicks etc.	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.55	1.61	0.01	-3.76	14.57
Qc40:takes things	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.52	1.64	0.01	-3.60	13.28
Qc41:laughs at others	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.42	1.71	0.01	-2.91	8.09
Qc42:restless	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.24	3.00	0.01	-1.50	1.13
Qc43:distractible	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.18	3.05	0.01	-1.46	1.00
Qc44:fidgets	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.09	3.06	0.01	-1.36	0.73
Qc45:disobedient	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.12	2.14	0.01	-2.39	5.14
Qc46:temper tantrums	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.55	1.65	0.01	-3.97	16.18
Qc47:impulsive	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.63	2.68	0.01	-1.84	2.46
Qc48:difficulty awaiting turns	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.64	2.65	0.01	-1.81	2.38

Item	Values	Mean	Sd	Se	Skewness	Kurtosis
Qc49:can't settle	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.05	2.32	0.01	-2.46	5.38
Qc50:inattentive	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	7.92	2.98	0.01	-1.13	0.25
Qc51:seems unhappy	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.11	2.14	0.01	-2.34	4.89
Qc52:fearful	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.18	2.07	0.01	-2.51	5.80
Qc53:worried	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	8.96	2.24	0.01	-2.01	3.25
Qc54:cries a lot	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.41	1.83	0.01	-3.24	10.47
Qc55:nervous	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.33	1.95	0.01	-2.99	8.72
Qc56:indecisive	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	9.02	2.21	0.01	-2.15	3.95
Qc57:shy	0=Often/very true; 5=Sometimes/somewhat true; 10=Never/not true	7.8	3.18	0.01	-1.15	0.18
Language & thinking skills (26)						
Qb8:handles a book	0=No; 10=Yes	9.95	0.70	0.00	-14.24	200.63
Qb9:interested in books	0=No; 10=Yes	9.67	1.79	0.01	-5.20	25.08
Qb10:interested in reading	0=No; 10=Yes	9.15	2.79	0.01	-2.97	6.84
Qb11:identifies letters	0=No; 10=Yes	9.11	2.85	0.01	-2.88	6.28
Qb12:sounds to letters	0=No; 10=Yes	8.41	3.66	0.02	-1.86	1.47
Qb13:rhyming awareness	0=No; 10=Yes	8.01	3.99	0.02	-1.51	0.28
Qb14:group reading	0=No; 10=Yes	9.19	2.74	0.01	-3.06	7.37
Qb15:reads simple words	0=No; 10=Yes	7.52	4.32	0.02	-1.17	-0.63
Qb16:reads complex words	0=No; 10=Yes	2.31	4.21	0.02	1.28	-0.37
Qb17:reads sentences	0=No; 10=Yes	4.88	5.00	0.02	0.05	-2.00
Qb18:experiments writing	0=No; 10=Yes	9.34	2.48	0.01	-3.50	10.26
Qb19:writing directions	0=No; 10=Yes	9.37	2.44	0.01	-3.59	10.86
Qb20:writing voluntarily	0=No; 10=Yes	6.79	4.67	0.02	-0.77	-1.41
Qb21:write own name	0=No; 10=Yes	9.77	1.51	0.01	-6.33	38.02
Qb22:write simple words	0=No; 10=Yes	8.64	3.42	0.02	-2.13	2.54
Qb23:write simple sentences	0=No; 10=Yes	6	4.90	0.02	-0.41	-1.83
Qb24:remembers things	0=No; 10=Yes	8.4	3.67	0.02	-1.86	1.45
Qb25:interested in maths	0=No; 10=Yes	9.25	2.63	0.01	-3.24	8.47
Qb26:interested in number games	0=No; 10=Yes	9.29	2.57	0.01	-3.34	9.12
Qb27:sorts and classifies	0=No; 10=Yes	9.57	2.03	0.01	-4.51	18.34
Qb28:1 to 1 correspondence	0=No; 10=Yes	9.45	2.28	0.01	-3.90	13.18
Qb29:counts to 20	0=No; 10=Yes	8.24	3.81	0.02	-1.70	0.90
Qb30:recognizes 1-10	0=No; 10=Yes	8.7	3.37	0.02	-2.20	2.82
Qb31:compares numbers	0=No; 10=Yes	8.97	3.04	0.01	-2.61	4.80
Qb32:recognizes shapes	0=No; 10=Yes	9.54	2.10	0.01	-4.33	16.74
Qb33:time concepts	0=No; 10=Yes	9.42	2.33	0.01	-3.79	12.37
Communication & general knowledge (8)						

Item	Values	Mean	Sd	Se	Skewness	Kurtosis
Qb1:effective use-English	0=Poor/very poor; 5=Average; 10=Very good/good	7.55	3.19	0.01	-0.95	-0.19
Qb2:listens - English	0=Poor/very poor; 5=Average; 10=Very good/good	7.99	2.84	0.01	-1.07	0.14
Qb3:tells a story	0=Poor/very poor; 5=Average; 10=Very good/good	7.12	3.36	0.02	-0.75	-0.57
Qb4:imaginative play	0=Poor/very poor; 5=Average; 10=Very good/good	7.72	2.87	0.01	-0.82	-0.33
Qb5:communicates needs	0=Poor/very poor; 5=Average; 10=Very good/good	7.62	3.13	0.01	-0.96	-0.14
Qb6:understands	0=Poor/very poor; 5=Average; 10=Very good/good	7.72	3.08	0.01	-1.00	-0.04
Qb7:articulates clearly	0=Poor/very poor; 5=Average; 10=Very good/good	7.27	3.39	0.02	-0.85	-0.46
Qc26:knowledge about world	0=Never/not true; 5=Sometimes/somewhat true; 10=Often/very true	8.83	2.44	0.01	-1.99	3.18

Note: Both the physical and emotional areas have a mix of negatively and positively worded questions.

The ten most highly skewed items, extracted from Table 2 are presented in Table 3 for easy reference. All ten are skewed to the left (negatively skewed) or scores tend to be lower and have high peakedness or some scored very high and some scored very low on these items.

Table 3: Ten most highly skewed items, Merger #3, Alberta (N=52,035)

	Mean	Skewness	Kurtosis
Physical			
Qa5:Hungry	9.55	-4.37	17.05
Qa6:Washroom	9.85	-7.99	61.77
Qa7:Hand preference	9.76	-6.29	37.53
Qa8:Well coordinated	9.31	-3.4	9.56
Qc58:Sucks thumb	9.74	-5.65	33.14
Emotional			
Qc45:disobedient	9.55	-3.97	16.18
Language			
Qb8:handles a book	9.95	-14.24	200.63
Qb9:interested in books	9.67	-5.2	25.08
Qb21:write own name	9.77	-6.33	38.02
Qb27:sorts and classifies	9.57	-4.51	18.34

What is your takeaway?

The items are divided into 33 dichotomously-scored items (0/10 or 10 points for each one) and 70 polytomously-scored items (0/5/10 or 15 points for each). Both item formats used the same estimates, item totals and entire area totals. Almost all items are negatively skewed. The highest levels of skewness and kurtosis occurred on the physical health and well-being and language and thinking skills areas.

Item difficulty (*p*-values)

As earlier noted, the proportion of children who get the correct answer to a dichotomous item is termed its *p* value. It is also called the item's difficulty level in CTT. The *p*- values (the average and average relative item scores) for all the 103 items by age are presented on Table 5. The *p*- values for the items in physical health & well-being range from 73% for Qa9 to 99% for Qa6. This means, Qa9 was correctly answered by 99% of the children or *p* for that item is 0.99.

In order to understand the *p*- values on the table better, let us recall the question we posed earlier: how difficult should a good item be? Several things must be taken into consideration in order to determine the difficulty level. A major question to be asked is: what is the probability of answering an item correctly by chance alone? If the question is in a true/false format, the chance for answering it correctly is 50% because there are only two options. For such a question, if the *p*- value is 0.50, the correct answer could be obtained by guessing alone. Similarly, a trichotomous item with three response categories could be answered correctly by 33% of the time. In this case, the item difficulty would be greater than 0.33 in order to discriminate between individuals' ability to guess correctly. Desirable difficulty levels can be estimated as midway between 100% and the percentage of expected guessing. Thus, for multiple choice items with two response categories, *p* should be around 0.75 (average of 100 and 50) and for those with three response categories, it should be around 0.67 (average of 100 and 33). Although the *p*-values are not the only statistics one should use to judge item quality, in general, tests are more reliable when the *p*- values range from 0 to 1 with a large concentration of items with medium difficulty or *p*=0.75 or 0.67, respectively for a dichotomously-or polytomously-coded item. The items are classified according to their difficulty levels, following Ebel (1965) as presented in Table 4.

Table 4: Classification rules for difficulty levels and the number of EDI items in each category, Merger #3, Alberta (N=52,035)

An item is:	If it has a difficulty index (*p*) of:	No. of EDI items
Very easy	0.91 or above	39
Easy	0.76 to 0.90	46
Optimum difficult	0.26 to 0.75	17
Difficult	0.11 to 0.25	1
Very difficult	0.10 and below	0

Based on the rule, almost 38% of the items were very easy, 45% were easy, and 17% were of optimum difficulty. [10] With this rule, there is not a single item that was very difficult. In conformity with the rule, 64 items are either "good" (with optimum difficulty level) or "fair" (easy, optimum, or difficult). The mean *p* values across areas ranged from 0.71 (communication & general knowledge) to 0.84 (physical health & well-being) with almost 83% of the items easy or very easy. Since variance is depended upon the mean, items with optimum difficulty have a better chance to show the most variance, and consequently more discrimination.

Out of all the 103 items, only one item from the language & thinking skills – Qb16: reads complex words – was found difficult. As the *p* values by age indicate, the older the children, the higher their proficiency levels. However, regardless of their age, all children found Qb16 (reads complex words) as difficult.

Table 5: The five developmental areas with 103 Items and the *p*-values by age, Merger #3, Alberta (N=52, 035)

Item	All ages	Age (yrs.) <=5.40	5.41 -5.65	5.66-5.90	>=5.91
Physical health & wellbeing (13)					
Qa2:dressed inappropriately	**90.97**	89.52	91.29	91.36	91.71
Qa3:Too tired	89.14	86.24	88.81	90.43	90.96
Qa4:Late	79.14	77.65	78.84	79.78	80.18
Qa5:Hungry	95.46	94.78	95.71	95.94	95.41
Qa6:Washroom	98.50	97.81	98.24	98.86	99.06
Qa7:Hand preference	97.65	96.44	97.40	98.16	98.55
Qa8:Well coordinated	93.10	90.47	92.81	94.32	94.55
Qa9:Proficient at holding pen	73.07	65.68	71.20	76.32	78.71
Qa10:Manipulates objects	78.53	72.32	77.15	80.93	83.43
Qa11:Climbs stairs	81.10	76.29	80.22	82.81	84.83
Qa12:Level of energy	77.06	71.92	76.31	78.94	80.88
Qa13:Overall physical	78.79	73.64	77.67	80.86	82.72
Qc58:Sucks thumb	97.38	96.62	97.18	97.62	98.01
Social competence (26)					
Qc1:overall soc/emotional	72.29	66.86	71.41	74.98	75.62
Qc2:gets along with peers	74.92	71.01	74.23	76.75	77.38
Qc3:cooperative	84.42	81.36	83.92	85.94	86.23
Qc4:plays with various children	83.33	80.44	82.93	84.63	85.13
Qc5:follows rules	84.62	81.26	84.04	86.13	86.83
Qc6:respects property	90.67	88.67	90.62	91.31	92.00
Qc7:self-control	83.09	80.52	82.83	84.33	84.54
Qc8:self-confidence	78.32	73.39	77.28	80.31	81.95
Qc9:respect for adults	91.67	90.46	91.48	92.18	92.51
Qc10:respect for children	87.40	86.00	87.36	87.86	88.26

[10] A better distribution of the values of p is: 5% easy; 20% medium-low difficulty; 50% medium difficulty; 20% medium-hard; and 5% difficult, although there is no consensus on such break downs. However, with 39 items exceeding 0.9 difficulty levels, the variance of these items can be very low (=0.90x0.10=0.09 and sqrt (0.09) =0.30).

Table 5: The five developmental areas with 103 Items and the *p*-values by age, Merger #3, Alberta (N=52, 035)

Item	All ages	Age (yrs.) <=5.40	5.41 -5.65	5.66-5.90	>=5.91
Qc11:accept responsibility	83.90	81.15	83.46	85.38	85.46
Qc12:listens	76.80	71.50	75.49	79.20	80.66
Qc13:follows directions	81.58	76.41	80.42	83.83	85.31
Qc14:completes work on time	80.18	72.83	78.81	83.04	85.61
Qc15:independent	80.96	73.13	79.79	83.98	86.46
Qc16:takes care of materials	90.71	87.92	90.22	91.86	92.66
Qc17:works neatly	78.90	73.35	77.65	81.32	82.92
Qc18:curious	90.57	87.13	90.36	91.97	92.60
Qc19:eager new toy	93.94	92.50	93.68	94.60	94.87
Qc20:eager new game	92.91	91.17	92.66	93.66	94.01
Qc21:eager new book	89.15	86.25	88.71	90.55	90.87
Qc22:independent solve problems	76.54	69.80	75.62	78.98	81.27
Qc23:follow simple instructions	91.52	87.82	91.06	93.01	93.95
Qc24:follow class routines	83.74	78.92	82.99	85.81	86.96
Qc25:adjust to change	87.64	83.82	87.05	89.32	90.08
Qc27:tolerance for mistake	85.04	82.77	85.35	85.77	86.13
Emotional maturity (30)					
Qc28:help hurt	70.13	66.75	70.09	71.27	72.23
Qc29:clear up mess	62.02	58.34	61.64	63.53	64.34
Qc30:stop quarrel	46.88	41.64	45.81	48.63	51.15
Qc31:offers help	56.24	49.77	55.54	58.44	60.82
Qc32:comforts upset	57.55	54.23	57.48	58.67	59.63
Qc33:spontaneously helps	59.47	55.63	58.92	61.09	61.93
Qc34:invite bystanders	50.71	46.49	50.47	52.35	53.39
Qc35:helps sick	52.58	48.99	52.12	53.88	55.13
Qc36:upset when left	88.98	86.78	88.62	89.81	90.50
Qc37:gets into fights	94.62	94.49	94.66	94.75	94.51
Qc38:bullies or mean	91.68	91.89	91.95	91.56	91.29
Qc39:kicks etc.	95.48	95.05	95.44	95.69	95.65
Qc40:takes things	95.23	94.60	95.13	95.47	95.65
Qc41:laughs at others	94.17	94.36	94.30	94.13	93.84
Qc42:restless	82.37	79.13	81.69	83.97	84.53
Qc43:distractible	81.84	77.14	80.80	84.02	85.11
Qc44:fidgets	80.87	77.38	80.11	82.61	83.20
Qc45:disobedient	91.18	90.14	90.97	91.78	91.77
Qc46:temper tantrums	95.54	94.83	95.36	96.00	95.88
Qc47:impulsive	86.34	84.94	85.91	87.41	87.03
Qc48:difficulty awaiting turns	86.37	84.48	86.26	87.33	87.32
Qc49:can't settle	90.54	87.61	90.06	91.72	92.59
Qc50:inattentive	79.25	74.67	78.15	81.37	82.60
Qc51:seems unhappy	91.07	90.31	90.84	91.51	91.60
Qc52:fearful	91.82	90.84	91.49	92.16	92.70
Qc53:worried	89.59	89.13	89.25	89.82	90.12
Qc54:cries a lot	94.12	92.51	93.85	94.90	95.09
Qc55:nervous	93.28	93.23	92.96	93.43	93.52
Qc56:indecisive	90.17	87.52	89.49	91.36	92.14
Qc57:shy	78.04	74.34	77.00	79.36	81.18
Language & thinking skills (30)					
Qb8:handles a book	99.51	99.14	99.49	99.62	99.77
Qb9:interested in books	96.67	95.35	96.59	97.37	97.29

Table 5: The five developmental areas with 103 Items and the *p*-values by age, Merger #3, Alberta (N=52, 035)

Item	All ages	Age (yrs.) <=5.40	5.41 -5.65	5.66-5.90	>=5.91
Qb10:interested in reading	91.48	88.04	91.29	92.99	93.39
Qb11:identifies letters	91.05	86.99	90.62	92.38	94.02
Qb12:sounds to letters	84.07	77.31	83.30	86.37	89.02
Qb13:rhyming awareness	80.11	71.18	78.83	83.61	86.16
Qb14:group reading	91.85	87.81	91.70	93.58	94.12
Qb15:reads simple words	75.24	66.35	73.72	78.64	81.81
Qb16:reads complex words	**23.07**	**16.51**	**21.55**	**25.16**	**28.84**
Qb17:reads sentences	48.80	39.83	46.76	52.11	56.10
Qb18:experiments writing	93.42	91.10	92.99	94.42	95.08
Qb19:writing directions	93.67	89.98	93.42	95.05	95.97
Qb20:writing voluntarily	67.87	60.32	66.36	70.84	73.65
Qb21:write own name	97.67	95.81	97.58	98.40	98.79
Qb22:write simple words	86.44	80.19	85.81	88.74	90.74
Qb23:write simple sentences	60.00	51.61	58.56	63.18	66.29
Qb24:remembers things	84.06	78.74	82.73	86.38	87.81
Qb25:interested in maths	92.53	89.31	92.29	93.81	94.52
Qb26:interested in number games	92.88	89.98	92.70	93.98	94.66
Qb27:sorts and classifies	95.71	93.27	95.27	96.74	97.37
Qb28:1 to 1 correspondence	94.48	91.07	94.16	95.68	96.78
Qb29:counts to 20	82.42	74.77	81.22	85.26	87.95
Qb30:recognizes 1-10	86.96	81.03	85.90	89.47	91.14
Qb31:compares numbers	89.67	84.47	89.20	91.36	93.27
Qb32:recognizes shapes	95.39	93.16	95.23	96.22	96.76
Qb33:time concepts	94.22	90.89	94.04	95.45	96.25
Communication & GK (8)					
Qb1:effective use-English	75.54	69.15	87.73	90.52	91.76
Qb2:listens - English	79.92	74.31	93.21	95.41	96.15
Qb3:tells a story	71.19	64.02	82.89	87.07	88.28
Qb4:imaginative play	77.23	73.39	93.52	94.91	94.67
Qb5:communicates needs	76.16	69.92	88.69	91.67	92.40
Qb6:understands	77.16	70.23	90.02	92.62	93.74
Qb7:articulates clearly	72.68	66.83	83.32	86.92	87.36
Qc26:knowledge about world	88.33	83.24	96.28	97.65	97.72

Note: Items that are very easy (>0.91) are highlighted in green, easy (0.76<p<0.90) in blue, optimum difficulty (0.26<p< 0.75) in pink, and difficult (0.11<p<0.25) in orange shades.

What is your takeaway?

A large majority of items are (very) easy (85 out of the 103 items or 82.52%), and consequently will have low ability to discriminate, although both easy and difficult items are needed to adequately select test content and objectives. Although *p* values are not the best and only statistics of difficulty levels and they can only be used as estimates, there are indications that in the case of many questions, observers did guess or they got the items correct.

Item discrimination

How related is an item to its scale or construct? Or simply, what is the tendency of children getting the correct answer to a given item also get a high overall score for the test? There are several indices of discrimination one can use o answer this, but we are interested in the correlation coefficient measuring the strength of a relationship between performance on an item and performance on a test (area). The coefficient should be positive, indicating that those answering correctly tend to have higher overall scores and those answering incorrectly tend to have lower overall scores. The higher the correlation, the better is the item discrimination.

The SPSS output with item-total statistics is presented in Table 6. The column labeled Corrected item-total correlation provides the corrected biserial correlation (CPBC). The suffix, corrected here refers to taking out the item score from the calculations so that the item being examined is not contributing to itself in terms of the statistics. There are seven items with lower than the recommended level of point-biserials (<0.30)[11]: Qa4 (late), Qa6 (washroom), Qa7 (hand preference), Qc58 (sucks thumb), Qc36 (upset when left), Qc57 (shy), and Qb8 (handles a book). With four out of the 13 items showing low biserial correlation in the physical health & well-being area, the indication is that the four items are not really measuring the same thing the rest of the items are trying to measure. The two items with very low point-biserial correlations (Qa6 and Qc58) in the area showed very high *p*-values (0.99 and 0.98), indicating that they are problematic items. There are three items, Qa6 (washroom), Qc58 (sucks thumb), and Qc57 (shy) with corrected item-total correlations as low as 0.13.

Five out of the eight items in the communication & general knowledge area have extremely high item-total correlations. Undoubtedly, they exhibit excellent discriminatory power. However, one might wonder, how far this can be true? In fact, we expect to find a nice break line between low scorers and high scorers. Is there a possibility that the items in the area are inflated, especially because it is comprised of only eight items? Items in the area appear to have conflicting *p*-values and point-biserial correlations. Most have optimal *p*-values, but high point-biserial correlations. The high *p*-values should not be taken as indicative of superior quality. However, because point-biserial correlations are strongly

[11] It is advisable to use a minimum threshold value for the point-biserial correlation. Generally, a value of at least 0.15 is recommended, although there are indications that 'good" items have point-biserial above 0.30 (Pallant, 2007).

influenced by the p-values, the low correlations may have something to do with the underlying construct; it may very well be the result of multidimensionality.[12]

In the column headed *Cronbach's alpha if item deleted*, the values represent the impact of removing each item from the area. These values can be compared to the overall alpha for the area, assuming it is unidimensional, and if any of the values in the column are higher than the overall alpha, it is recommended to remove this item from the test.[13] The items, Qa4 (late), Qc58 (sucks thumb), Qc36 (upset when left), Qc57 (shy), and Qc26 (knowledge about world) all have values higher than the final alpha. Although only by a small amount, the reliability would increase to 0.786 from 0.782 for the physical health & well-being area if either of Qa6 and Qc58 were deleted. The reliability for the emotional maturity area would increase to 0.917 and 0.919 from 0,915, if Qc36 and Qc57 were removed.[14] For scales with a small number of items (i.e. <10), as in the communication & general knowledge area, Cronbach's alpha values can be misleading and it is better to consider the mean inter-item correlation value instead. In the case of communication & general knowledge, the mean inter-item correlation is 0.63, suggesting a strong relationship among the items, although less than the recommended level of 0.70.

[12] We acknowledge that it is somewhat premature to make such a claim, and we may require additional analyses in order to identify the conceptual structure of the tool. A crude measure of dimensionality is the reliability coefficient, Cronbach's alpha. A more appropriate method for assessing the unidimensionality of a test is factor analysis (Hambleton & Rovinelli, 1986).
The results of a factor analysis, using Maximum Likelihood procedures showed the following results: (1) three factors for the physical health & well-being area with 10 items having initial eigen values less than 1 ; (2) four factors for the social competence area with 22 items having initial eigen values less than 1.0; (3) four factors for the emotional maturity area with 25 items having initial eigen values less than 1.0; (4) five factors for the language and thinking skills area with 21 items having initial eigen values less than 1.0; and (5) one factor for the communication & general knowledge area with all eight items having initial eigen values less than 1.0. Total variance explained were respectively,40.93%, 61.71%, 56.20%, 42.63%, and 64.16%..

[13] Removal of the items, however, makes comparison of results with other studies difficult. In addition, in well-validated tools, it becomes a cause for concern only when the alpha falls below 0.70 (Pallant, 2007).

[14] The deletion of items may be performed one item at a time and repeating the analysis. However, it is important to ensure that the overall alpha is not lowered in the process. If, on the other hand, the alpha value is higher than the alpha with the item included, one should consider deleting the item not only to improve the overall reliability, but also to reduce the time and energy in administering a lengthy questionnaire.

Table 6: The five developmental areas with 103 items and Item-total statistics, Merger #3, Alberta (N=52,035)

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Physical health & wellbeing (13)				
Qa2:dressed inappropriately	104.11	289.637	**.288**	.781
Qa3:Too tired	104.29	277.405	**.377**	.772
Qa4:Late	105.28	280.284	**.214**	.802
Qa5:Hungry	103.67	299.942	**.298**	.777
Qa6:Washroom	103.37	318.154	**.138**	.786
Qa7:Hand preference	103.46	311.444	**.224**	.782
Qa8:Well coordinated	103.90	285.227	**.403**	.769
Qa9:Proficient at holding pen	105.89	255.109	**.581**	.749
Qa10:Manipulates objects	105.34	257.305	**.671**	.741
Qa11:Climbs stairs	105.10	265.166	**.634**	.746
Qa12:Level of energy	105.49	256.358	**.663**	.741
Qa13:Overall physical	105.32	258.323	**.698**	.739
Qc58:Sucks thumb	103.49	317.321	**.132**	.786
Cronbach's Alpha=0.782				
Social competence (26)				
Qc1:overall soc/emotional	212.57	1838.988	**.670**	.952
Qc2:gets along with peers	212.31	1844.318	**.682**	.952
Qc3:cooperative	211.36	1861.963	**.737**	.951
Qc4:plays with various children	211.47	1868.690	**.670**	.952
Qc5:follows rules	211.35	1859.320	**.757**	.951
Qc6:respects property	210.75	1898.615	**.689**	.952
Qc7:self-control	211.50	1860.114	**.699**	.951
Qc8:self-confidence	211.97	1883.097	**.548**	.953
Qc9:respect for adults	210.65	1911.968	**.654**	.952
Qc10:respect for children	211.07	1886.586	**.678**	.952
Qc11:accept responsibility	211.42	1853.920	**.718**	.951
Qc12:listens	212.13	1841.524	**.712**	.951
Qc13:follows directions	211.65	1845.053	**.762**	.951
Qc14:completes work on time	211.79	1854.856	**.647**	.952
Qc15:independent	211.71	1843.785	**.700**	.951
Qc16:takes care of materials	210.74	1897.944	**.689**	.952
Qc17:works neatly	211.92	1860.756	**.620**	.952
Qc18:curious	210.75	1923.426	**.548**	.953
Qc19:eager new toy	210.42	1961.898	**.420**	.954
Qc20:eager new game	210.51	1950.876	**.455**	.954
Qc21:eager new book	210.89	1916.621	**.531**	.953
Qc22:independent solve problems	212.14	1840.741	**.685**	.952
Qc23:follow simple instructions	210.65	1908.099	**.665**	.952
Qc24:follow class routines	211.43	1855.201	**.720**	.951
Qc25:adjust to change	211.04	1878.657	**.691**	.952
Qc27:tolerance for mistake	211.31	1887.193	**.597**	.952
Cronbach's Alpha=0.954				
Emotional maturity (30)				
Qc28:help hurt	235.18	1781.772	**.625**	.911

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Qc29:clear up mess	235.94	1760.515	**.649**	.910
Qc30:stop quarrel	237.45	1763.037	**.614**	.911
Qc31:offers help	236.51	1746.195	**.677**	.909
Qc32:comforts upset	236.40	1758.682	**.643**	.910
Qc33:spontaneously helps	236.20	1761.429	**.649**	.910
Qc34:invite bystanders	237.09	1769.826	**.614**	.911
Qc35:helps sick	236.92	1756.141	**.639**	.910
Qc36:upset when left	233.32	1924.661	**.176**	.917
Qc37:gets into fights	232.75	1902.192	**.425**	.914
Qc38:bullies or mean	233.04	1884.704	**.446**	.914
Qc39:kicks etc.	232.66	1905.937	**.434**	.914
Qc40:takes things	232.69	1908.409	**.407**	.914
Qc41:laughs at others	232.80	1905.803	**.405**	.914
Qc42:restless	233.96	1815.375	**.570**	.912
Qc43:distractible	234.01	1807.919	**.591**	.911
Qc44:fidgets	234.11	1812.869	**.568**	.912
Qc45:disobedient	233.09	1861.571	**.564**	.912
Qc46:temper tantrums	232.65	1905.729	**.425**	.914
Qc47:impulsive	233.57	1831.830	**.570**	.912
Qc48:difficulty awaiting turns	233.57	1837.298	**.553**	.912
Qc49:can't settle	233.15	1851.858	**.564**	.912
Qc50:inattentive	234.27	1809.166	**.598**	.911
Qc51:seems unhappy	233.10	1886.691	**.422**	.914
Qc52:fearful	233.03	1905.816	**.330**	.915
Qc53:worried	233.25	1903.350	**.315**	.915
Qc54:cries a lot	232.80	1913.611	**.331**	.915
Qc55:nervous	232.88	1906.500	**.350**	.915
Qc56:indecisive	233.20	1880.998	**.436**	.914
Qc57:shy	234.38	1922.296	**.135**	.919
Cronbach's Alpha=0.915				
Language & thinking skills (26)				
Qb8:handles a book	209.61	1990.231	**.213**	.901
Qb9:interested in books	209.87	1946.277	**.356**	.899
Qb10:interested in reading	210.37	1877.266	**.503**	.897
Qb11:identifies letters	210.43	1849.124	**.605**	.895
Qb12:sounds to letters	211.14	1791.729	**.645**	.893
Qb13:rhyming awareness	211.52	1787.453	**.598**	.894
Qb14:group reading	210.35	1872.148	**.532**	.896
Qb15:reads simple words	211.97	1750.715	**.657**	.893
Qb16:reads complex words	217.16	1848.994	**.373**	.901
Qb17:reads sentences	214.57	1759.082	**.525**	.897
Qb18:experiments writing	210.21	1918.163	**.370**	.899
Qb19:writing directions	210.18	1898.851	**.471**	.897
Qb20:writing voluntarily	212.72	1791.063	**.487**	.898
Qb21:write own name	209.78	1954.188	**.364**	.900
Qb22:write simple words	210.89	1839.676	**.525**	.896
Qb23:write simple sentences	213.49	1790.298	**.460**	.899
Qb24:remembers things	211.14	1813.225	**.572**	.895
Qb25:interested in maths	210.29	1881.631	**.512**	.897

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Qb26:interested in number games	210.26	1891.581	**.477**	.897
Qb27:sorts and classifies	209.98	1920.757	**.450**	.898
Qb28:1 to 1 correspondence	210.11	1893.046	**.533**	.897
Qb29:counts to 20	211.29	1803.027	**.581**	.895
Qb30:recognizes 1-10	210.84	1824.037	**.591**	.895
Qb31:compares numbers	210.58	1844.625	**.577**	.895
Qb32:recognizes shapes	210.01	1919.851	**.438**	.898
Qb33:time concepts	210.14	1912.898	**.419**	.898
Cronbach's Alpha=0.901				
Communication & GK (8)				
Qb1:effective use-English	54.35	298.563	**.850**	.917
Qb2:listens - English	53.91	314.386	**.792**	.922
Qb3:tells a story	54.78	294.442	**.838**	.918
Qb4:imaginative play	54.19	322.195	**.695**	.929
Qb5:communicates needs	54.29	301.028	**.844**	.918
Qb6:understands	54.19	304.610	**.821**	.919
Qb7:articulates clearly	54.64	304.994	**.724**	.927
Qc26:knowledge about world	53.07	345.464	**.562**	.937
Cronbach's Alpha=0.933; mean inter-item correlation=0.630				

Notes:
1. In terms of discrimination coefficients, items that are "good" (>=0.30) are highlighted in green and those that are "bad" (<0.30) are highlighted in blue shades. The corrected item-total correlations highlighted in red font represent those values lower than 0.15 ("very bad" items).
2. In the column headed Cronbach's alpha if item deleted, values that are higher than the final alpha for the area are highlighted in orange shades.

What is your takeaway?

Problematic items, regardless of high *p* values show low point-biserial correlations. Seven items have <0.30 thresh-hold level biserials. An indicator of the overall test reliability (Cronbach's alpha if item deleted) also supports the removal of these items.

Distractor-test correlations

Neither the item difficulty nor the item discrimination index considers the performance of the incorrect response options, or distractors. A distractor analysis addresses the performance of incorrect response options. Just as those who are skilled or knowledgeable in the area or whose responses endorse a particular statement, distractors should be a

reasonable selection among those who do not possess the necessary skill or knowledge in the or whose responses fail to support a particular statement. If a distractor appears so unlikely that almost no individual will select it, it is simply not contributing to the quality of the item. In fact, the presence of very few distractors in a multiple choice item can make the item or the test in general, too easy. In other words, if distractors were either not selected at all or selected by a minority, it is likely that the content area behind the distractors were well understood by the individuals so that distractors were not behaving like "distractors" after all.

Table 7 presents the results of a distractor analysis, namely "wrong" response in all items. To compute the correlation between the total score and the distractor ("wrong"/0), a new binary variable was created, giving a score of 10 to every child who got 0 and a score of 0 to others. The correlations between this new variable and total score are given in the column, corrected item-total correlation. As noted earlier, these correlations should be negative. In other words, children who got the incorrect answer for an item should tend to score lower in the area containing the item.

Probably, the most puzzling result in Table 7 is that the distractor, "wrong" is a positive distractor because it has all positive point-biserial correlations between individuals' scores on this particular distractor and their scores on the whole test for all items and all five content areas. In the light of this finding, it would have been of little use to check for correlations between the other distractor (i.e. response category 5) and test scores in the survey. While (very) easy items can distort distractor-test correlations, ultimately the issue is one of changing or reevaluating the response options in order to achieve a nice break line between low and high scorers

Table 7: The five developmental areas with 103 items and corrected distractor ("wrong")–total correlations, Merger #3, Alberta (N=52,035)

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Physical health & wellbeing (13)				
Qa2:dressed inappropriately	6.740	155.383	0.322	0.189
Qa3:Too tired	6.565	146.419	0.408	0.266
Qa4:Late	5.568	147.598	0.227	0.091
Qa5:Hungry	7.178	163.754	0.344	0.203
Qa6:Washroom	7.478	179.745	0.157	0.069
Qa7:Hand preference	7.395	173.766	0.259	0.141
Qa8:Well coordinated	6.954	153.344	0.430	0.294
Qa9:Proficient at holding pen	6.781	152.730	0.375	0.357
Qa10:Manipulates objects	7.272	161.957	0.445	0.423
Qa11:Climbs stairs	7.450	172.280	0.355	0.312
Qa12:Level of energy	7.215	160.881	0.426	0.267
Qa13:Overall physical	7.410	167.486	0.443	0.392
Qc58:Sucks thumb	7.504	181.931	0.106	0.014
Cronbach's Alpha=0.686				
Social competence (26)				
Qc1:overall soc/emotional	6.968	472.206	0.570	0.447
Qc2:gets along with peers	7.148	482.965	0.553	0.47
Qc3:cooperative	7.481	505.768	0.559	0.47
Qc4:plays with various children	7.401	502.659	0.504	0.405
Qc5:follows rules	7.499	505.887	0.587	0.492
Qc6:respects property	7.563	518.406	0.478	0.415
Qc7:self-control	7.373	494.909	0.584	0.46
Qc8:self-confidence	7.255	501.314	0.411	0.238
Qc9:respect for adults	7.596	524.386	0.424	0.401
Qc10:respect for children	7.550	516.281	0.491	0.501
Qc11:accept responsibility	7.311	491.944	0.564	0.396
Qc12:listens	7.224	482.806	0.604	0.496
Qc13:follows directions	7.397	494.785	0.613	0.535
Qc14:completes work on time	7.127	486.655	0.502	0.42
Qc15:independent	7.144	478.942	0.594	0.512
Qc16:takes care of materials	7.551	518.64	0.446	0.276
Qc17:works neatly	7.138	488.421	0.489	0.307
Qc18:curious	7.562	525.866	0.315	0.202
Qc19:eager new toy	7.624	535.199	0.200	0.56
Qc20:eager new game	7.600	531.434	0.254	0.583
Qc21:eager new book	7.473	518.197	0.344	0.275
Qc22:independent solve problems	7.087	477.233	0.578	0.372
Qc23:follow simple instructions	7.574	522.068	0.421	0.268
Qc24:follow class routines	7.344	492.389	0.588	0.444

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Qc25:adjust to change	7.451	506.179	0.509	0.322
Qc27:tolerance for mistake	7.386	509.182	0.401	0.194
Cronbach's Alpha=0.899				
Emotional maturity (30)				
Qc28:help hurt	22.563	1104.732	0.566	0.466
Qc29:clear up mess	21.949	1067.417	0.608	0.563
Qc30:stop quarrel	20.545	1028.157	0.607	0.527
Qc31:offers help	21.451	1038.129	0.656	0.57
Qc32:comforts upset	21.588	1046.459	0.641	0.659
Qc33:spontaneously helps	21.826	1063.028	0.607	0.558
Qc34:invite bystanders	21.097	1043.766	0.594	0.5
Qc35:helps sick	21.085	1031.876	0.639	0.646
Qc36:upset when left	23.197	1213.181	0.085	0.057
Qc37:gets into fights	23.479	1207.302	0.266	0.502
Qc38:bullies or mean	23.426	1200.388	0.294	0.48
Qc39:kicks etc.	23.492	1206.504	0.294	0.561
Qc40:takes things	23.494	1211.011	0.233	0.251
Qc41:laughs at others	23.527	1215.043	0.217	0.221
Qc42:restless	22.947	1149.156	0.439	0.734
Qc43:distractible	22.908	1143.178	0.461	0.638
Qc44:fidgets	22.918	1145.089	0.453	0.758
Qc45:disobedient	23.417	1193.109	0.365	0.379
Qc46:temper tantrums	23.460	1203.485	0.292	0.306
Qc47:impulsive	23.157	1164.623	0.426	0.501
Qc48:difficulty awaiting turns	23.190	1170.892	0.397	0.42
Qc49:can't settle	23.292	1177.032	0.411	0.495
Qc50:inattentive	23.035	1151.888	0.456	0.513
Qc51:seems unhappy	23.419	1202.282	0.267	0.265
Qc52:fearful	23.432	1205.998	0.236	0.684
Qc53:worried	23.425	1205.938	0.231	0.689
Qc54:cries a lot	23.454	1208.875	0.221	0.229
Qc55:nervous	23.425	1203.336	0.260	0.363
Qc56:indecisive	23.411	1204.481	0.237	0.126
Qc57:shy	22.823	1194.967	0.144	0.111
Cronbach's Alpha=0.870				
Language & thinking skills (26)				
Qb8:handles a book	40.392	1990.231	0.213	0.096
Qb9:interested in books	40.126	1946.277	0.356	0.356
Qb10:interested in reading	39.625	1877.266	0.503	0.455
Qb11:identifies letters	39.566	1849.124	0.605	0.488
Qb12:sounds to letters	38.862	1791.729	0.645	0.529
Qb13:rhyming awareness	38.477	1787.453	0.598	0.405
Qb14:group reading	39.646	1872.148	0.532	0.321

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
Qb15:reads simple words	38.030	1750.715	0.657	0.519
Qb16:reads complex words	32.839	1848.994	0.373	0.293
Qb17:reads sentences	35.428	1759.082	0.525	0.463
Qb18:experiments writing	39.792	1918.163	0.370	0.191
Qb19:writing directions	39.816	1898.851	0.471	0.263
Qb20:writing voluntarily	37.278	1791.063	0.487	0.285
Qb21:write own name	40.216	1954.188	0.364	0.19
Qb22:write simple words	39.112	1839.676	0.525	0.394
Qb23:write simple sentences	36.506	1790.298	0.460	0.366
Qb24:remembers things	38.862	1813.225	0.572	0.344
Qb25:interested in maths	39.708	1881.631	0.512	0.606
Qb26:interested in number games	39.738	1891.581	0.477	0.582
Qb27:sorts and classifies	40.019	1920.757	0.450	0.301
Qb28:1 to 1 correspondence	39.890	1893.046	0.533	0.392
Qb29:counts to 20	38.709	1803.027	0.581	0.417
Qb30:recognizes 1-10	39.156	1824.037	0.591	0.491
Qb31:compares numbers	39.415	1844.625	0.577	0.442
Qb32:recognizes shapes	39.987	1919.851	0.438	0.266
Qb33:time concepts	39.860	1912.898	0.419	0.241
Cronbach's Alpha=0.901				
Communication & GK (8)				
Qb1:effective use-English	4.500	140.389	0.745	0.589
Qb2:listens - English	4.864	162.326	0.564	0.381
Qb3:tells a story	4.252	134.798	0.721	0.555
Qb4:imaginative play	4.865	165.974	0.487	0.268
Qb5:communicates needs	4.576	145.437	0.696	0.507
Qb6:understands	4.628	150.167	0.639	0.447
Qb7:articulates clearly	4.223	145.616	0.536	0.348
Qc26:knowledge about world	4.984	171.465	0.466	0.232
Cronbach's Alpha=0.857; mean inter-item correlation=0.431				

What is your takeaway?

Distractor (wrong)-total correlations are generally high and positive. In other words, "wrong" is a positive distractor because it has all positive point-biserial correlations between individuals' scores on this particular distractor and their scores on the whole test for all items and all five content areas.

Graphical item analysis: DIF

The three most common statistics reported in an item analysis are the item difficulty, which is a measure of the proportion of individuals who responded to an item correctly, the item discrimination, which is a measure of how well the item discriminates between individuals who are knowledgeable in the content area and those who are not, and the distractor analysis, which provides a measure of how well each of the incorrect options contributes to the quality of an item. An additional analysis that is often reported is the graphical analysis, which provides a simple way of presenting differences between any pair of populations (e.g., boys and girls). One important element in the graphical analysis is the DIF, which assumes that an item should be equally difficult or easy (based on *p*-values) in any pair of populations (boys/girls) if they represent some similarity in terms of their levels of proficiency. The starting point of the graphical analysis is to compute the p-values separately for each pair of the contrasting variable.

Although sex is commonly used as a contrasting variable in DIF analyses, we used three other variables, in addition to sex: age, English/French as a Second Language (EFSL), and repeated grade or not. The *p*-values for all the items by sex, EFSL, and repeated grade or not are presented in Table 8. The *p*-values are generally higher among females, but significantly higher in 16 out of the 103 items, most (11 out of 16) of which belong to the emotional maturity area. This finding supports prior research on sex differences in socio-emotional development (Krishnan, 2011). Only one item from the language & thinking skills – Qb16: reads complex words – was found difficult for boys and girls, EFSL and non-EFSL children, and repeaters and non-repeaters.

In terms of the EFSL variable, generally speaking, non-EFSL children score higher in almost all items than non-EFSL children. However, the p-values are slightly higher among non-EFSL children than EFSL children in more than fifty per cent of the items in the emotional area. The items with higher p-values for EFSL children than non-EFSL children actually fall into a unique dimension of its own, namely, anxiety and fearfulness.[15] This may have important implications in addressing the developmental outcomes, especially among ethno-cultural groups. Repeaters, in general, score lower than non-repeaters. Only on six items, all in the language & thinking skills area, they scored the same or slightly higher than non-repeaters.

[15] A Principal Component Analysis (PCA) showed a distinct aspect of the emotional maturity area with eight items forming a component of its own, anxiety and fearfulness (Krishnan, 2010).

Table 8: The five developmental areas with 103 Items and the p-values by sex, EFSL status, and repeated or not, Merger #3, Alberta (N=52,035)

Item	All	Sex		EFSL status		Repeated grade or not	
		Male	Female	No	Yes	No	Yes
Physical health & well-being (13)							
Qa2:dressed inappropriately	89.52	91.46	90.47	91.68	86.24	91.12	86.51
Qa3:Too tired	86.24	88.04	90.28	89.01	90.07	89.34	83.20
Qa4:Late	77.65	79.27	78.99	78.50	77.24	79.32	73.53
Qa5:Hungry	94.78	95.34	95.57	95.48	95.40	95.63	90.18
Qa6:Washroom	97.81	98.40	98.61	98.61	97.80	98.51	98.11
Qa7:Hand preference	96.44	96.64	98.69	97.60	97.61	97.64	97.88
Qa8:Well coordinated	90.47	90.99	95.27	93.13	92.67	93.19	90.26
Qa9:Proficient at holding pen	65.68	65.90	80.45	72.80	73.79	73.05	73.51
Qa10:Manipulates objects	72.32	74.22	82.97	78.36	77.26	78.55	77.94
Qa11:Climbs stairs	76.29	79.35	82.90	81.08	78.38	81.12	80.27
Qa12:Level of energy	71.92	75.59	78.58	77.10	75.15	77.17	73.75
Qa13:Overall physical	73.64	76.84	80.79	78.73	76.24	78.86	76.43
Qc58:Sucks thumb	96.62	97.53	97.23	97.32	98.01	97.40	96.63
Social competence (26)							
Qc1:overall soc/emotional	66.86	68.43	76.27	72.79	68.57	72.53	64.91
Qc2:gets along with peers	71.01	71.49	78.46	75.41	70.25	75.12	69.03
Qc3:cooperative	81.36	81.23	87.70	84.85	80.63	84.58	79.46
Qc4:plays with various children	80.44	80.87	85.85	83.93	78.44	83.45	79.52
Qc5:follows rules	81.26	80.27	89.10	85.00	81.37	84.77	80.05
Qc6:respects property	88.67	87.63	93.80	90.94	88.92	90.80	86.77
Qc7:self-control	80.52	77.64	88.72	83.18	82.20	83.28	77.51
Qc8:self-confidence	73.39	75.95	80.75	78.87	74.06	78.47	73.69
Qc9:respect for adults	90.46	89.38	94.04	91.64	91.13	91.79	88.10
Qc10:respect for children	86.00	84.54	90.34	87.53	86.25	87.53	83.38
Qc11:accept responsibility	81.15	79.93	88.01	84.18	81.75	84.11	77.78
Qc12:listens	71.50	71.81	81.93	77.28	73.48	77.02	70.21
Qc13:follows directions	76.41	77.39	85.89	82.21	77.03	81.74	76.56
Qc14:completes work on time	72.83	75.47	85.04	80.52	78.21	80.26	77.86
Qc15:independent	73.13	76.19	85.87	81.56	77.04	81.06	77.80
Qc16:takes care of materials	87.92	87.47	94.05	90.85	89.26	90.83	87.04
Qc17:works neatly	73.35	72.22	85.78	78.61	79.41	78.98	76.31
Qc18:curious	87.13	89.87	91.30	91.05	85.05	90.72	86.29
Qc19:eager new toy	92.50	94.02	93.87	94.00	91.51	94.00	92.20
Qc20:eager new game	91.17	92.64	93.19	93.03	90.05	92.98	90.59
Qc21:eager new book	86.25	86.33	92.05	89.47	85.71	89.28	85.28

Table 8: The five developmental areas with 103 Items and the p-values by sex, EFSL status, and repeated or not, Merger #3, Alberta (N=52,035)

Item	All	Sex		EFSL status		Repeated grade or not	
		Male	Female	No	Yes	No	Yes
Qc22:independent solve problems	69.80	73.39	79.78	77.44	67.85	76.71	71.30
Qc23:follow simple instructions	87.82	89.43	93.67	92.06	86.61	91.62	88.60
Qc24:follow class routines	78.92	79.44	88.17	84.03	80.08	83.88	79.49
Qc25:adjust to change	83.82	84.58	90.78	87.96	83.85	87.76	83.91
Qc27:tolerance for mistake	82.77	81.75	88.42	85.69	79.76	85.18	80.85
Emotional maturity (30)							
Qc28:help hurt	66.75	63.55	76.81	71.11	62.42	70.20	68.00
Qc29:clear up mess	58.34	55.04	69.17	62.52	56.77	62.10	59.59
Qc30:stop quarrel	41.64	43.05	50.80	47.72	38.06	46.92	45.80
Qc31:offers help	49.77	49.56	63.07	57.33	47.12	56.26	55.38
Qc32:comforts upset	54.23	48.99	66.17	58.35	49.21	57.62	55.54
Qc33:spontaneously helps	55.63	53.09	66.02	59.92	53.49	59.57	56.58
Qc34:invite bystanders	46.49	46.85	54.68	51.78	41.31	50.84	46.90
Qc35:helps sick	48.99	44.89	60.35	53.30	44.56	52.67	50.13
Qc36:upset when left	86.78	89.37	88.58	88.99	90.67	89.01	88.13
Qc37:gets into fights	94.49	91.35	97.98	94.64	94.76	94.75	90.74
Qc38:bullies or mean	91.89	90.20	93.20	91.56	92.40	91.81	87.84
Qc39:kicks etc.	95.05	92.97	98.06	95.50	95.52	95.57	92.54
Qc40:takes things	94.60	94.16	96.33	95.42	94.70	95.30	93.09
Qc41:laughs at others	94.36	92.03	96.37	94.24	94.13	94.26	91.48
Qc42:restless	79.13	75.83	89.11	82.23	83.66	82.63	74.60
Qc43:distractible	77.14	76.04	87.82	82.03	82.77	82.10	73.92
Qc44:fidgets	77.38	74.57	87.37	80.64	82.93	81.14	72.90
Qc45:disobedient	90.14	88.31	94.13	91.17	91.40	91.32	86.99
Qc46:temper tantrums	94.83	94.21	96.90	95.41	96.64	95.62	93.22
Qc47:impulsive	84.94	81.02	91.84	86.11	88.31	86.56	79.91
Qc48:difficulty awaiting turns	84.48	81.88	91.01	86.22	87.13	86.52	81.96
Qc49:can't settle	87.61	86.75	94.44	90.51	90.49	90.68	86.25
Qc50:inattentive	74.67	74.00	84.65	79.38	79.21	79.49	72.14
Qc51:seems unhappy	90.31	90.44	91.73	90.94	92.02	91.19	87.38
Qc52:fearful	90.84	91.47	92.18	91.81	92.65	91.84	91.17
Qc53:worried	89.13	89.22	89.97	89.52	91.17	89.62	88.66
Qc54:cries a lot	92.51	93.71	94.55	94.16	94.72	94.13	93.89
Qc55:nervous	93.23	92.28	94.32	93.13	94.92	93.30	92.73
Qc56:indecisive	87.52	89.04	91.34	90.48	88.50	90.25	87.75
Qc57:shy	74.34	80.25	75.77	78.95	71.70	77.98	79.87

Table 8: The five developmental areas with 103 Items and the p-values by sex, EFSL status, and repeated or not, Merger #3, Alberta (N=52,035)

Item	All	Sex		EFSL status		Repeated grade or not	
		Male	Female	No	Yes	No	Yes
Language & thinking skills (26)							
Qb8:handles a book	99.14	99.29	99.75	99.60	98.64	99.51	99.53
Qb9:interested in books	95.35	94.90	98.50	96.85	95.55	96.71	95.58
Qb10:interested in reading	88.04	88.29	94.77	92.06	88.07	91.59	88.41
Qb11:identifies letters	86.99	89.62	92.54	92.23	87.08	91.04	91.27
Qb12:sounds to letters	77.31	81.90	86.31	85.36	74.65	84.06	84.69
Qb13:rhyming awareness	71.18	77.85	82.42	83.50	58.31	80.13	79.45
Qb14:group reading	87.81	90.17	93.59	93.08	83.42	91.90	90.37
Qb15:reads simple words	66.35	72.57	78.00	77.61	64.54	75.24	75.18
Qb16:reads complex words	16.51	21.58	24.61	24.74	13.98	23.06	23.19
Qb17:reads sentences	39.83	46.24	51.44	51.44	38.18	48.80	48.74
Qb18:experiments writing	91.10	90.58	96.33	93.63	90.89	93.44	92.81
Qb19:writing directions	89.98	92.16	95.22	94.25	91.06	93.71	92.49
Qb20:writing voluntarily	60.32	55.25	80.83	68.63	59.76	67.99	64.43
Qb21:write own name	95.81	97.03	98.35	97.96	96.76	97.67	97.70
Qb22:write simple words	80.19	83.86	89.11	87.24	81.34	86.43	86.95
Qb23:write simple sentences	51.61	55.90	64.24	61.47	51.77	59.99	60.21
Qb24:remembers things	78.74	81.98	86.11	84.79	78.45	84.30	75.43
Qb25:interested in maths	89.31	92.73	92.32	92.85	89.56	92.65	89.03
Qb26:interested in number games	89.98	93.19	92.56	93.16	89.82	92.97	90.17
Qb27:sorts and classifies	93.27	95.08	96.36	96.27	90.95	95.75	94.40
Qb28:1 to 1 correspondence	91.07	93.95	95.03	95.00	90.61	94.51	93.69
Qb29:counts to 20	74.77	81.63	83.23	83.97	75.78	82.49	80.37
Qb30:recognizes 1-10	81.03	87.06	86.85	87.86	84.35	86.97	86.65
Qb31:compares numbers	84.47	89.18	90.17	90.96	82.12	89.73	88.02
Qb32:recognizes shapes	93.16	94.68	96.12	96.42	89.13	95.39	95.33
Qb33:time concepts	90.89	93.61	94.85	95.64	83.59	94.27	92.59
Communication & GK (8)							
Qb1:effective use-English	69.15	84.99	91.35	92.80	43.18	88.48	80.99
Qb2:listens - English	74.31	91.68	95.97	96.35	66.42	94.00	90.73
Qb3:tells a story	64.02	79.42	87.58	88.31	39.43	83.91	74.21
Qb4:imaginative play	73.39	47.16	96.29	95.25	73.32	93.58	90.03
Qb5:communicates needs	69.92	86.28	92.18	91.86	61.73	89.52	82.71
Qb6:understands	70.23	87.48	93.00	93.11	60.31	90.52	83.99
Qb7:articulates clearly	66.83	79.57	88.50	86.49	58.33	84.41	75.83
Qc26:knowledge about world	83.24	95.70	97.17	97.62	83.27	96.51	94.26

Notes:
1. Items with very high *p*-values for girls than boys (>10 point difference) are highlighted in pink shades and the item, Qc57 with a higher p value for boys than girls is highlighted in green shade.
2. The one item that is difficult for boys and girls, EFSL and non-EFSL children, and repeaters and non-repeaters is shaded in orange.
3. Items with higher p-values for EFSL than non-EFSL children are highlighted in blue shades.
4. Items with almost similar or higher *p*-values for repeaters than non-repeaters are highlighted in violet shades.

For DIF, the total sample is split into four groups, 1 denoting the groups with the lowest scores and 4 with the highest scores on the basis of the total scores in each area. In each group, the proportion of correct responses is computed and plotted against the group number for three contrasting variables, namely age, sex and EFSL status, as shown in Figures 3, 4, and 5. We highlight below only those items that stood out in differentiating the two groups. An important backdrop to this discussion is, of course, we rely on the p-values only.

An important thing to notice is that in the lowest of all groups, two items (Qa6: washroom and Qa10: manipulates objects) in the physical health & well-being area, six items (Qc42: restless; Qc43: distractible; Qc44: fidgets; Qc47: impulsive; Q48: difficulty awaiting turns; and Qc57: shy) in the emotional competence area, and an item (Qb20: writing voluntarily) in the language & thinking skills area have dissimilar *p*-values for boys and girls. There are three items (Qa12: level of energy; Qb16: reads complex words; and Qb20: writing voluntarily) in group 3 with dissimilar *p*-values for boys and girls. Only one item, namely Qb20 (writing voluntarily) shows some dissimilarity in *p*-values among group 2.

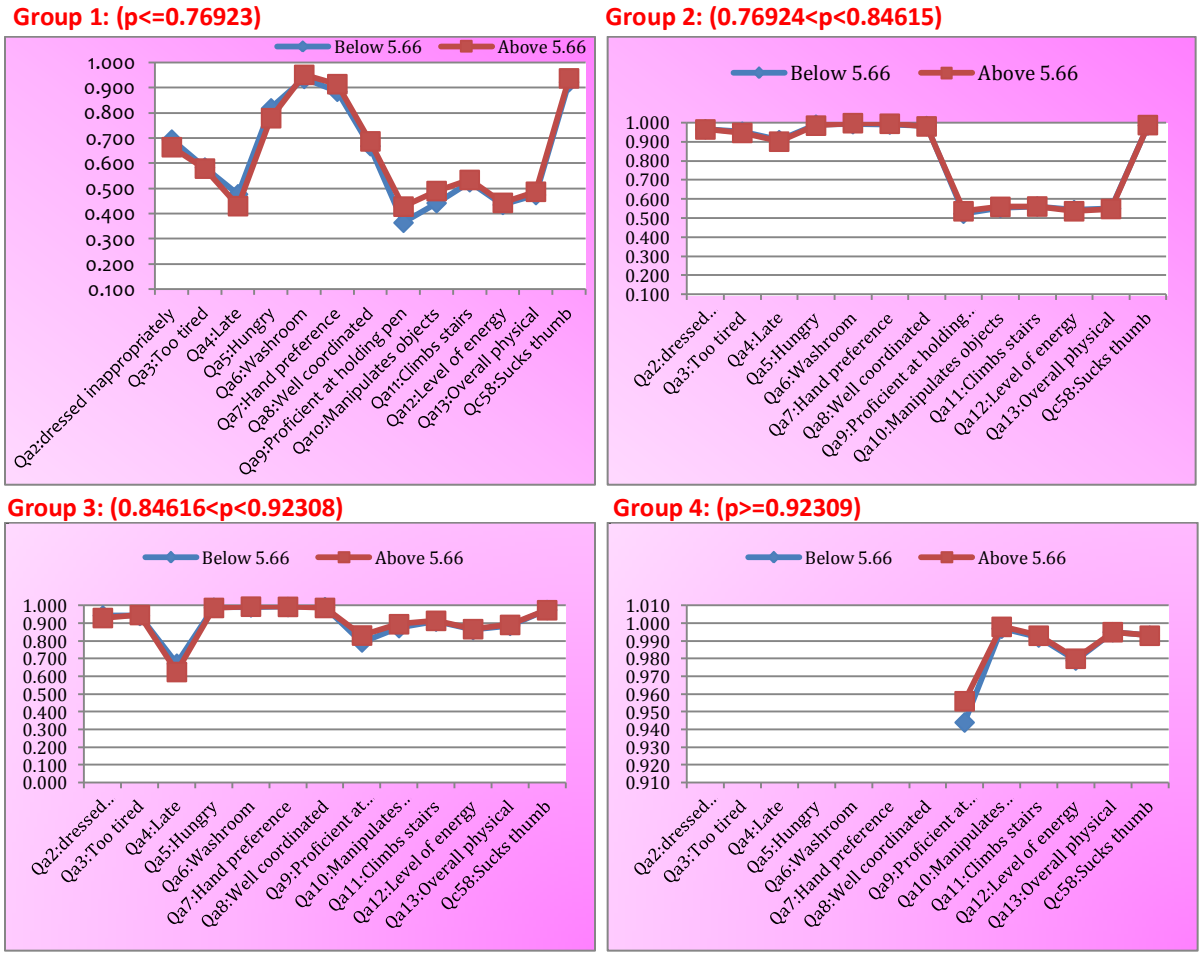
In terms of EFSL status, Qb13 (rhyming awareness) and Qb33 (time concepts) in the language & thinking skills area show differences in *p*-values between EFSL and non-EFSL children among low scorers. Another highly relevant finding is that, among the lowest scoring group, all items in the area of communication & general knowledge favor non-EFSL children. Although there are ways of testing the differences in *p* values statistically, an obvious driver in future discussion is the cultural fairness of the questions in the communication & general knowledge area.[16]

[16] It is acknowledged here that the *p* values inherently contain error and we simply cannot expect very similar values in each group, whatever the contrasting variables employed in DIF.
The low scorers' performance in the communication & general knowledge area suggests the importance of looking at both the vertical and horizontal aspects of development. For example, how an item, say "knowledge about world" would affect children at the bottom of the performance level differently from those at the top? And, how it affects across groups, such as EFSL and non-EFSL children?

Regardless of the child's gender or linguistic backgrounds, more than two-thirds of the entire sample got a large majority of questions correct. Stated simply, even among low scorers, a number of items are equally easy for boys and girls and for EFSL and non-EFSL children with same levels of proficiency. In fact, the presence of a large number of very high *p*-values and implausible distractors in items make most items artificially much easier than it ought to be, and probably not providing much discrimination between high scorers and low scores.

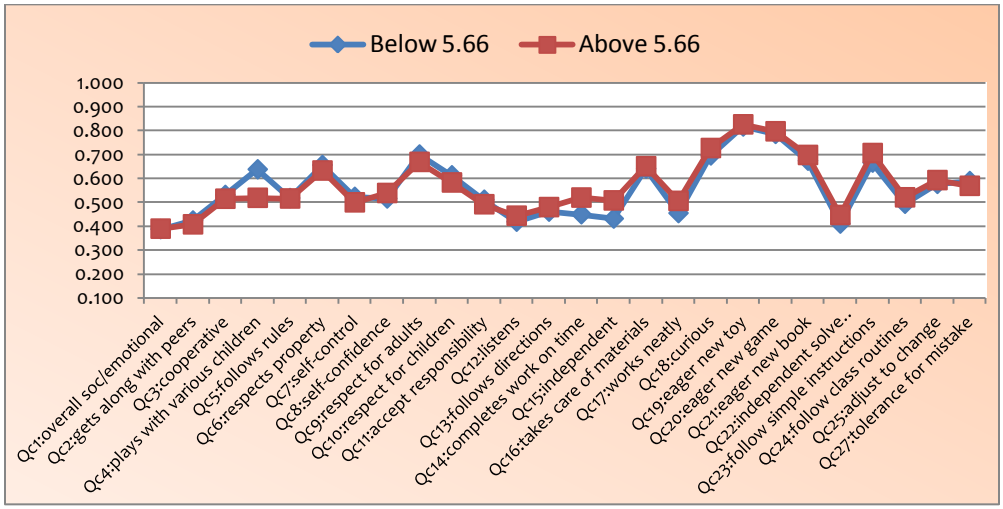
Figure 3: DIF analysis by age of child (based on quartiles)

A: Physical health & well-being

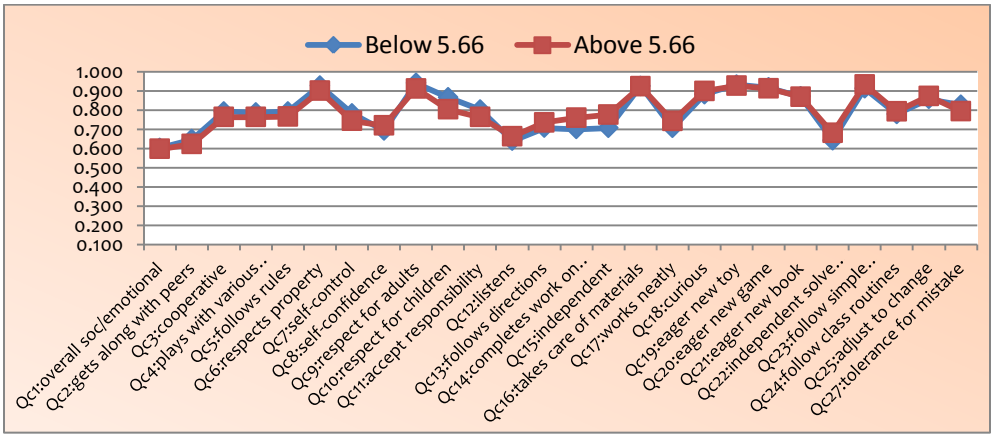


B: Social competence

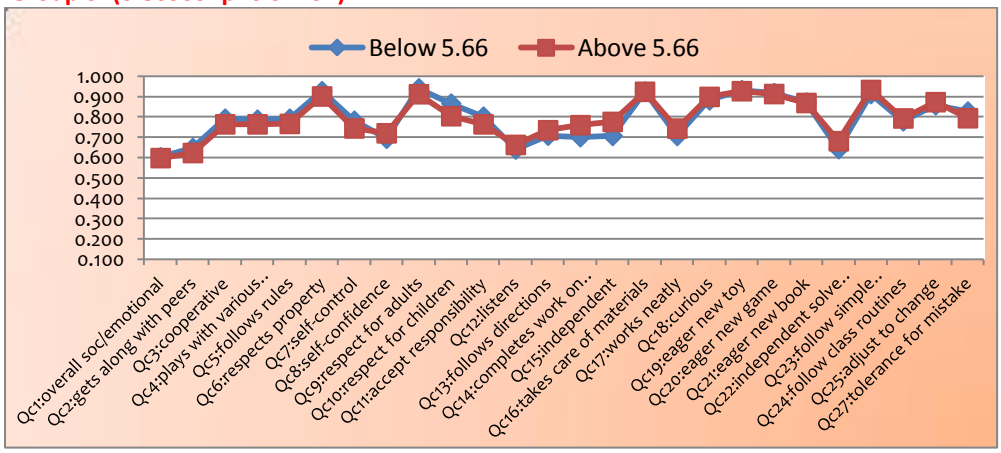
Group 1: (p<=0.71154)



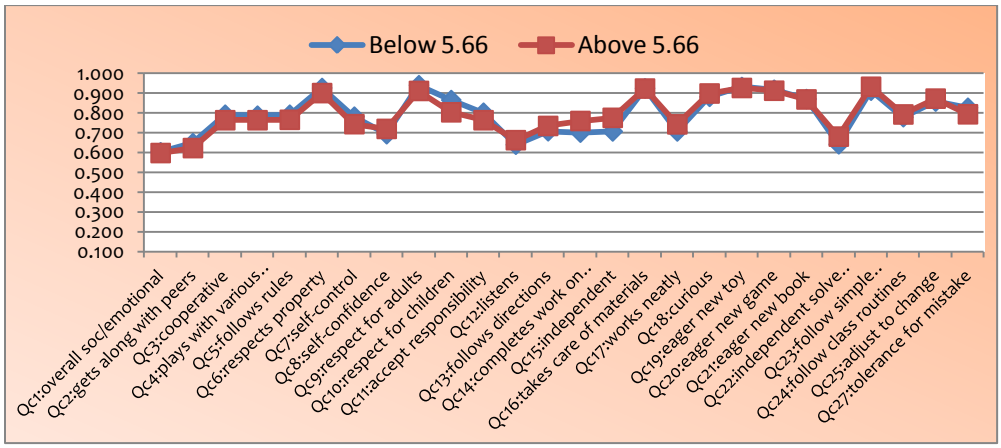
Group 2: (0.71155<p<0.86538)



Group 3: (0.86539<p<0.94231)

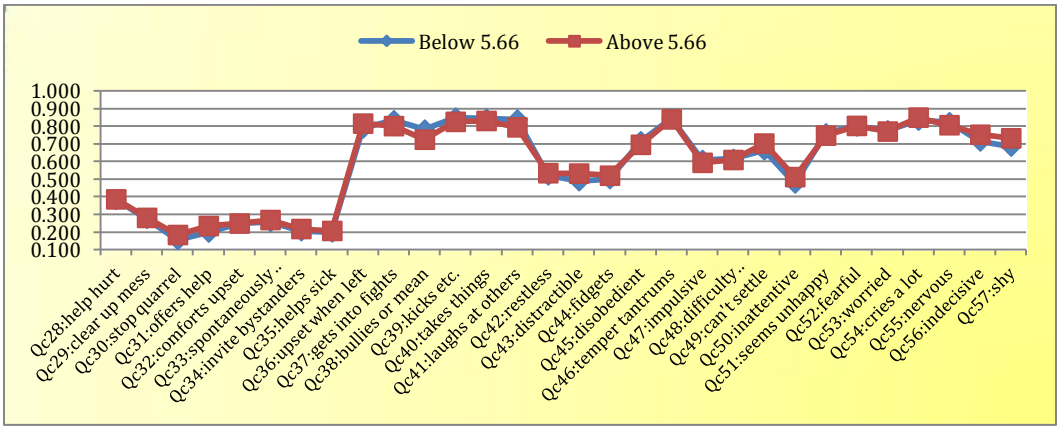


Group 4: (p>=0.94232)

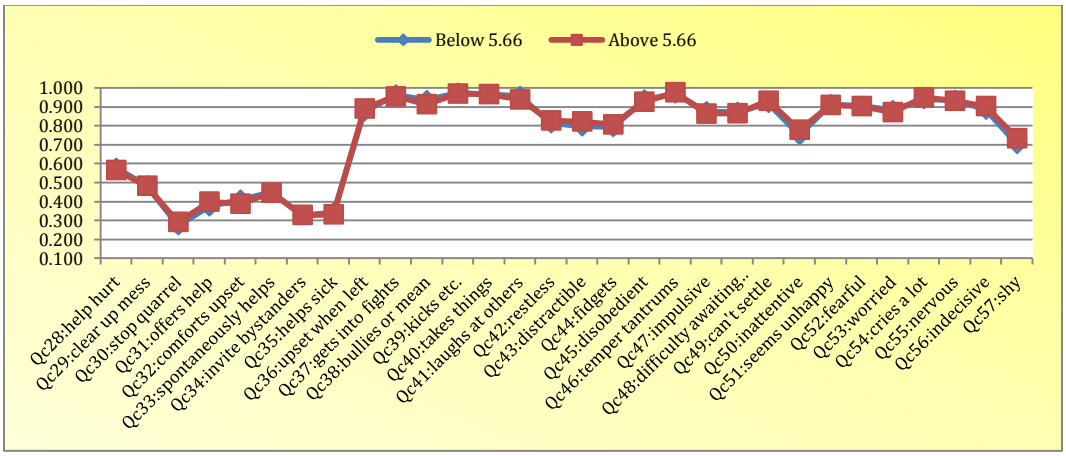


C: Emotional maturity

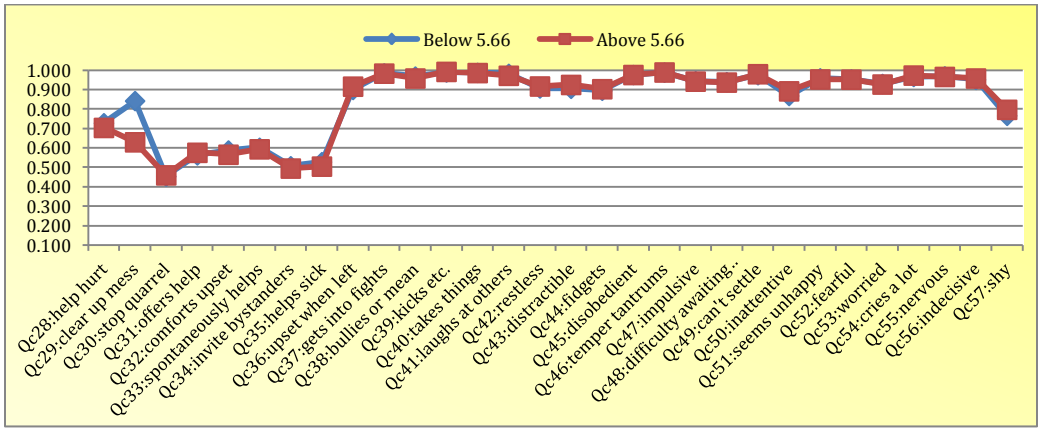
Group 1: (p<=0.7000)



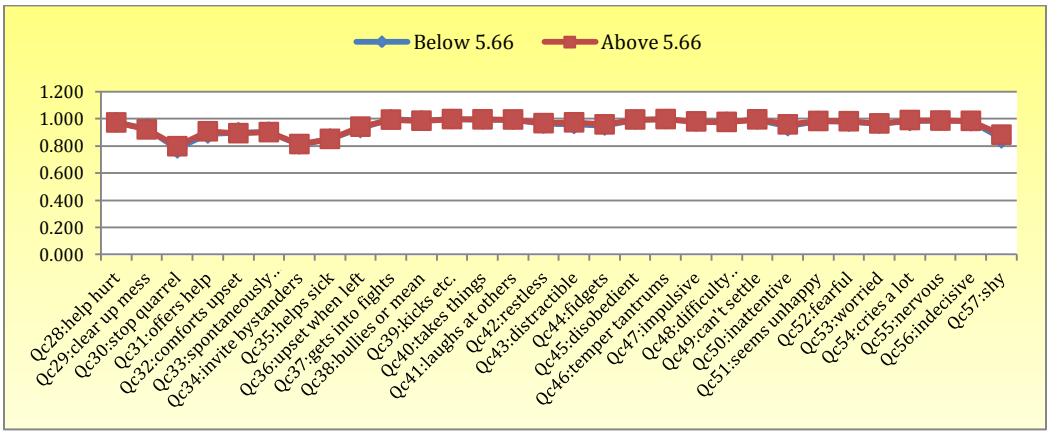
Group 2: (0.70001<p<0.80000)



Group 3: (0.80001<p<0.86667)

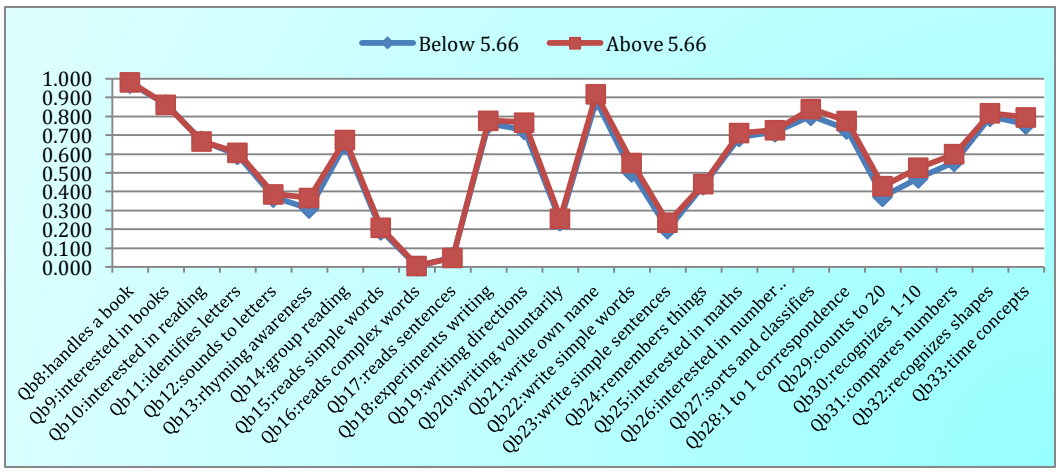


Group 4: (p>=0.86668)

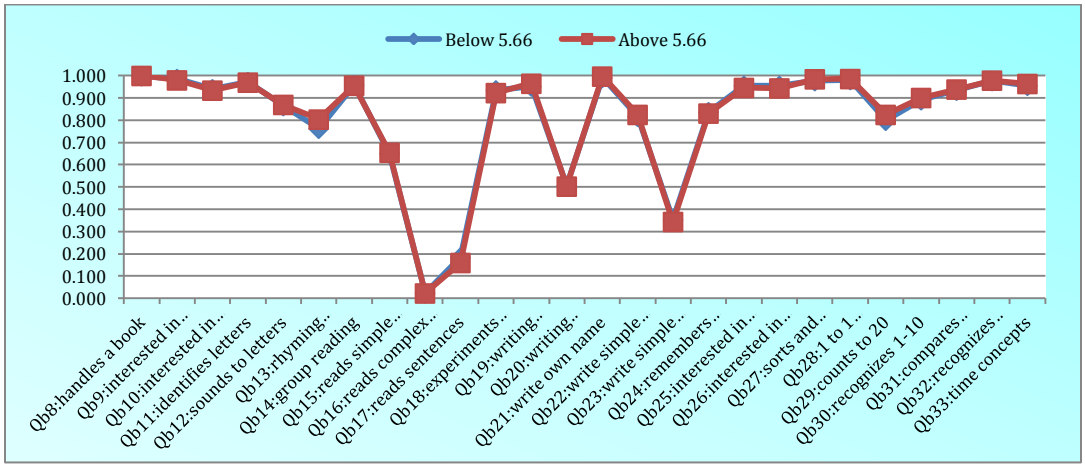


D: Language & thinking skills

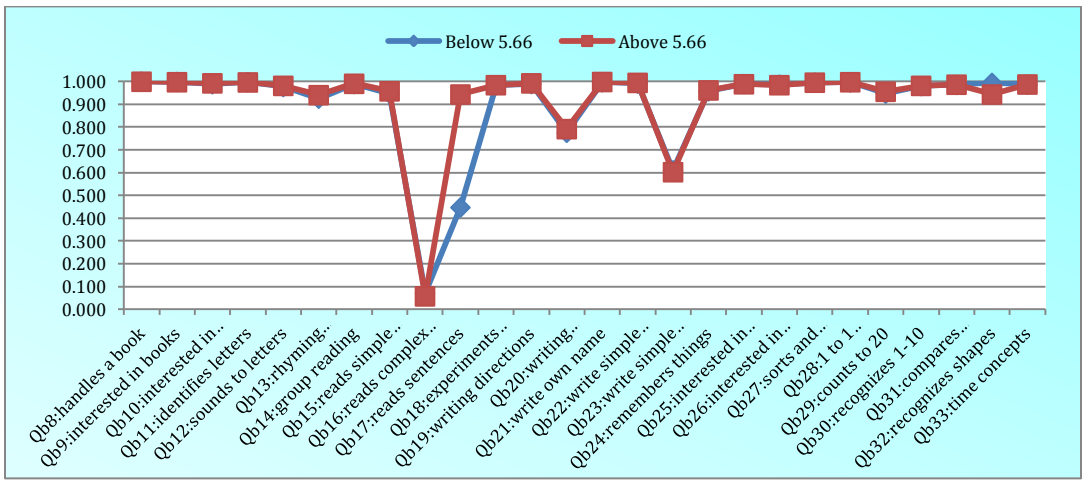
Group 1: (p<=0.73077)



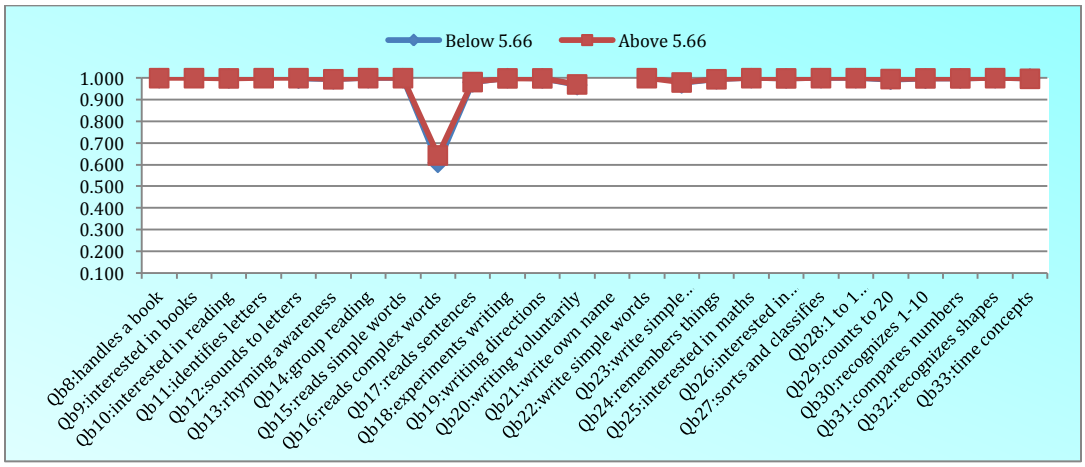
Group 2: (0.73078<p<0.87500)



Group 3: (0.87501<p<0.92308)

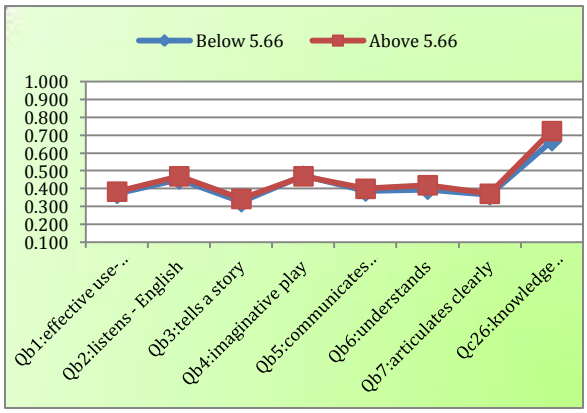


Group 4: (p>=0.92309)

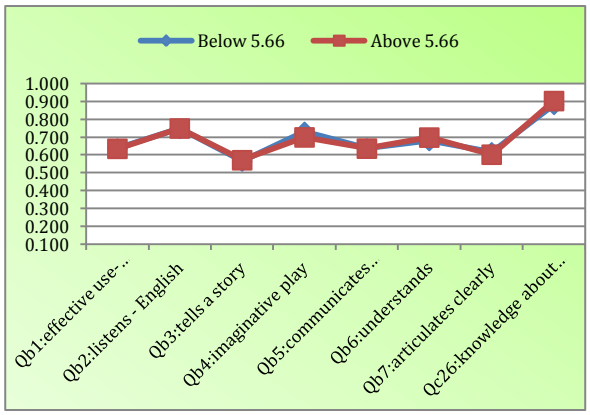


E: Communication & general knowledge

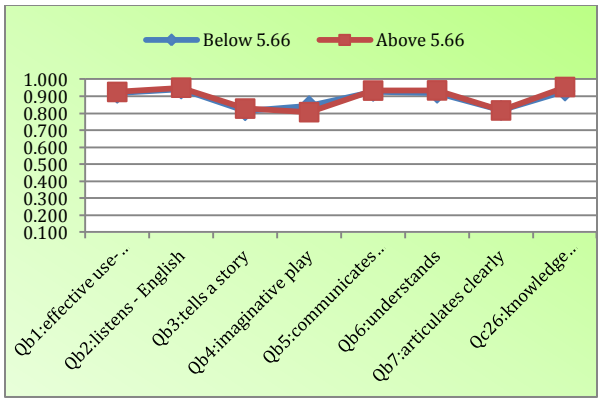
Group 1: (p<=0.56250)



Group 2: (0.56251<p<0.75000)



Group 3: (0.75001<p<0.93750)

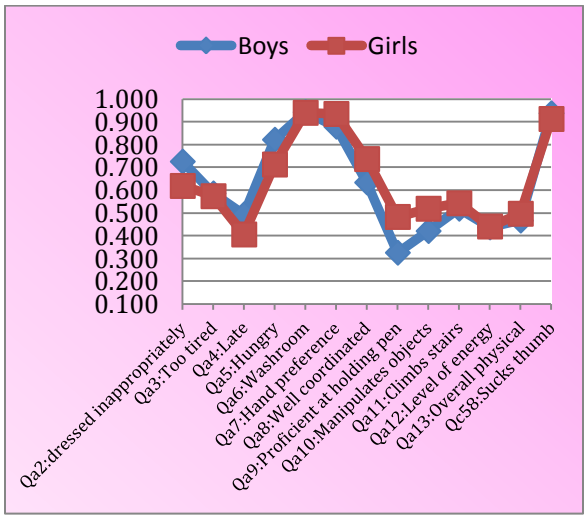


Group 4: (p>=0.93751) All items constant

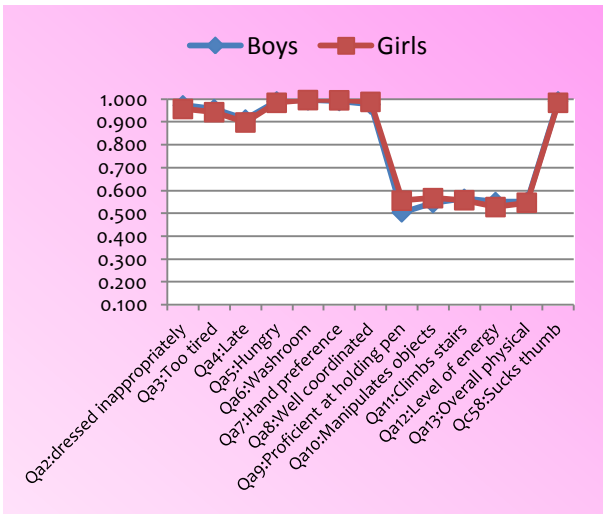
Figure 4: DIF analysis by sex of child

A: Physical health & well-being

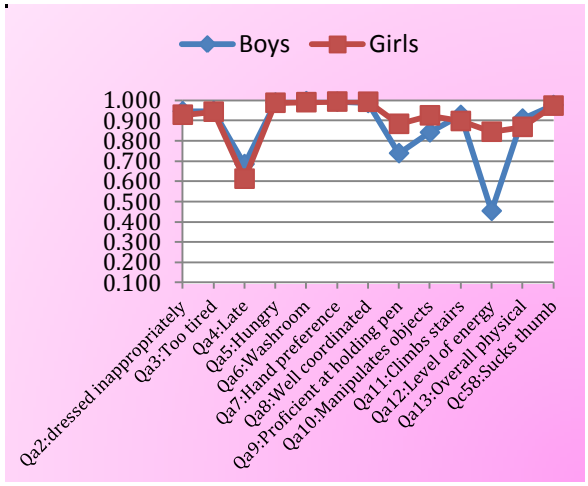
Group 1: (p<=0.76923)



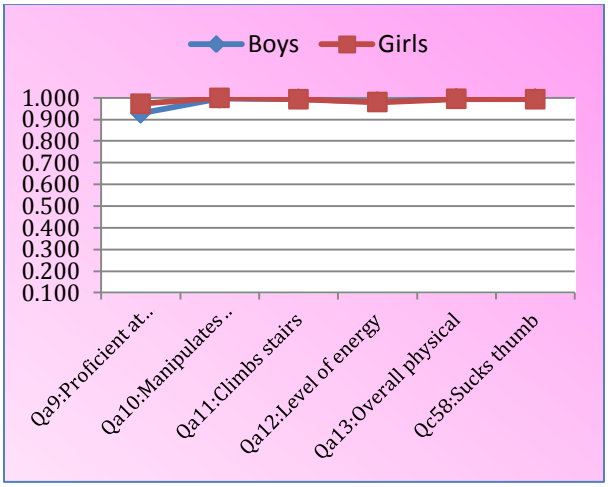
Group 2: (0.76924<p<0.84615)



Group 3: (0.84616<p<0.92308)

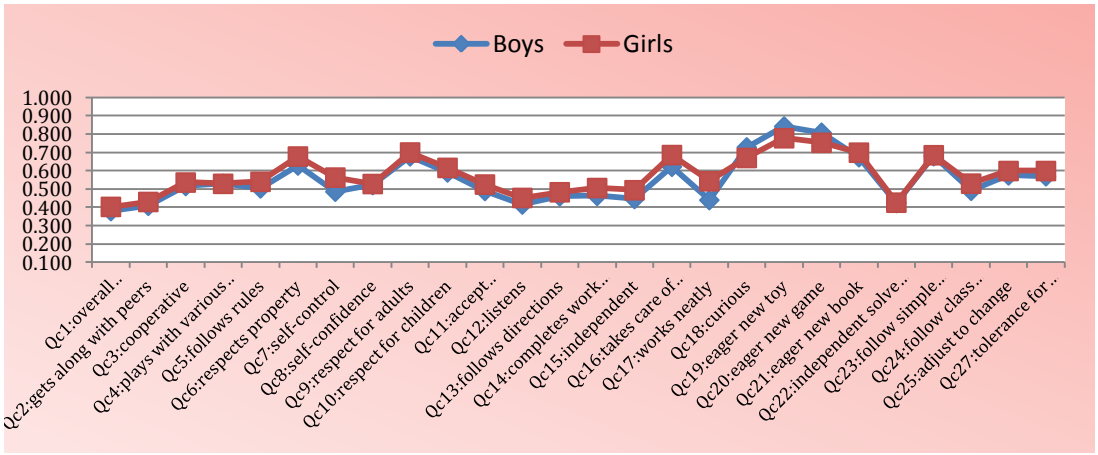


Group 4: (p>=0.92309)

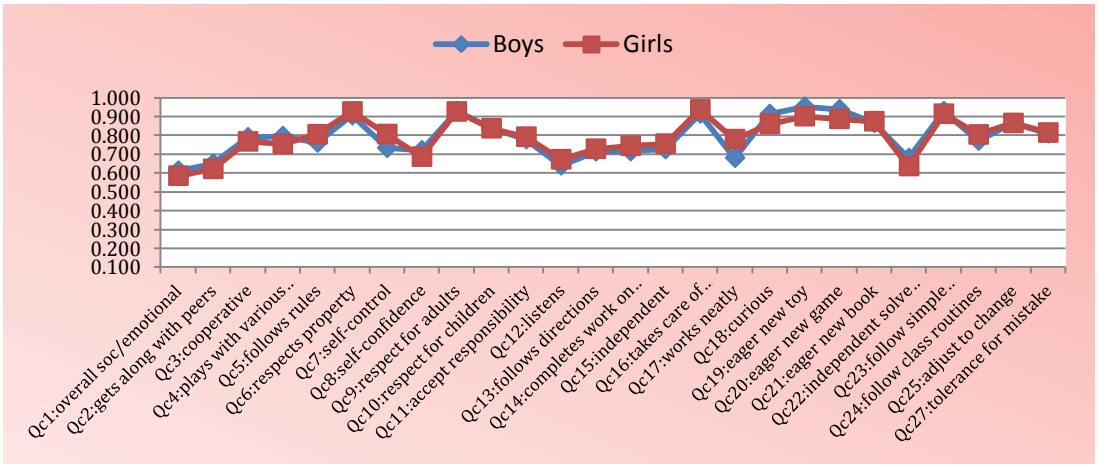


B: Social competence

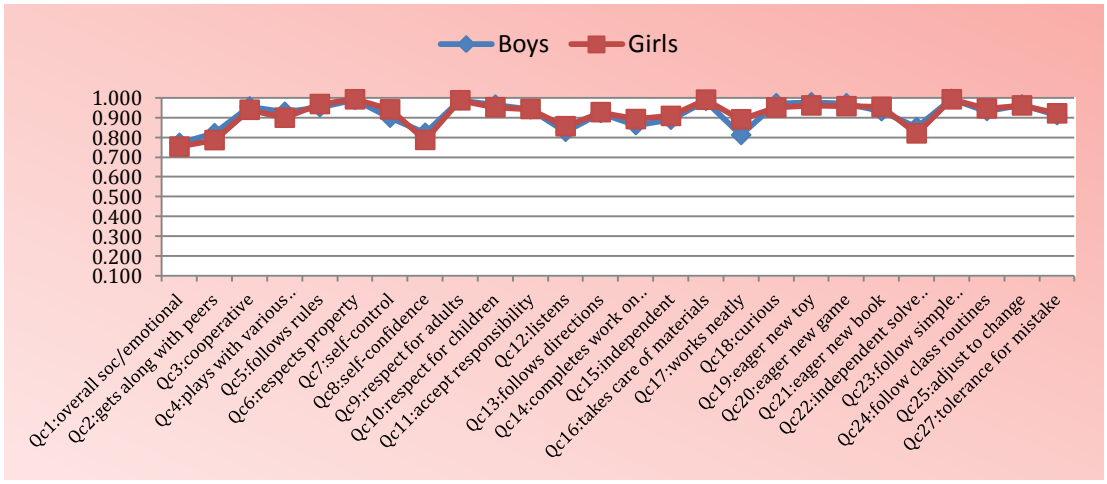
Group 1: (p<=0.71154)



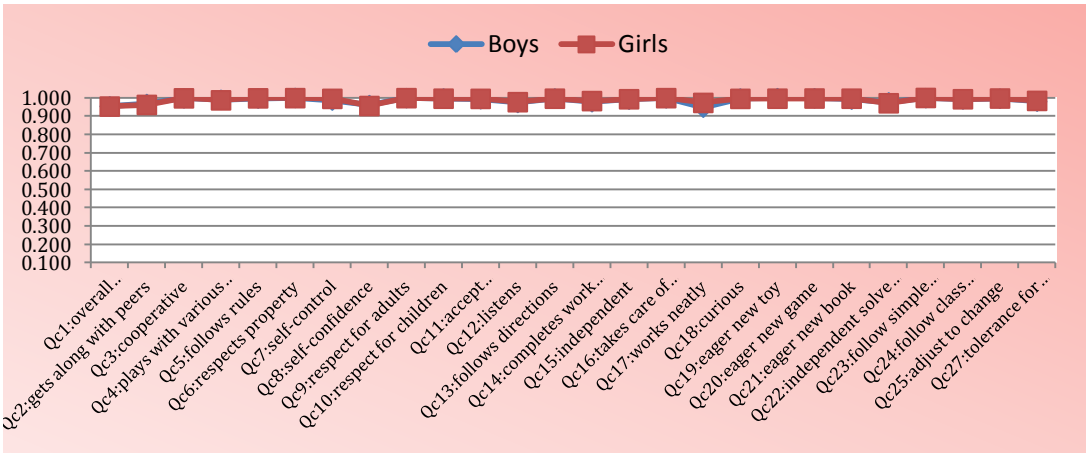
Group 2: (0.71155<p<0.86538)



Group 3: (0.86539<p<0.94231)

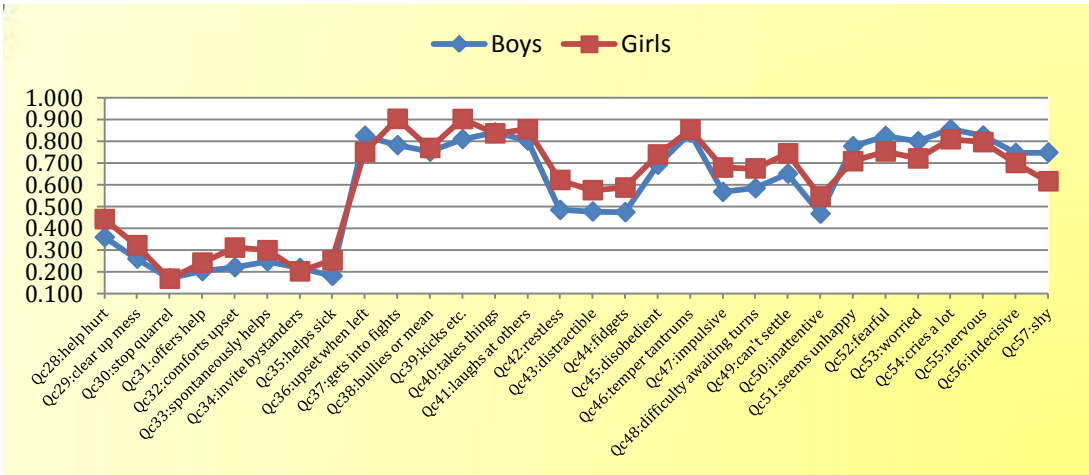


Group 4: (p>=0.94232)

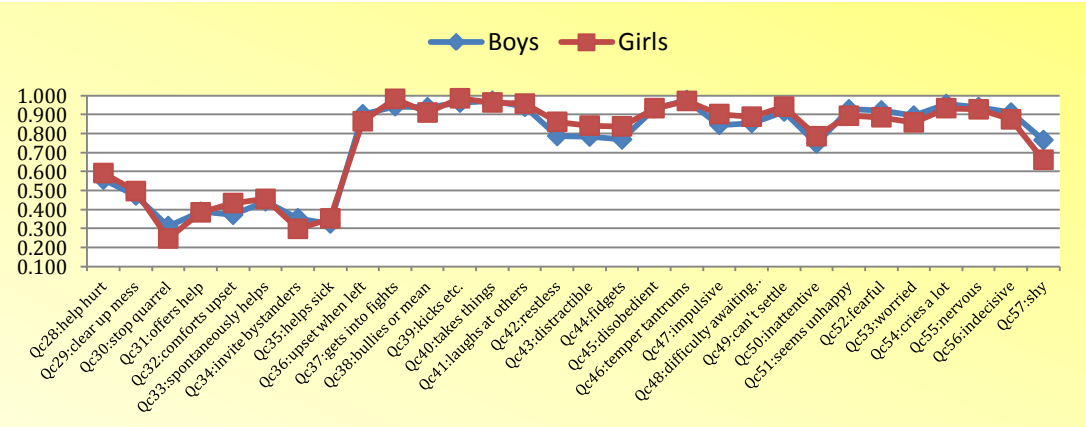


C: Emotional maturity

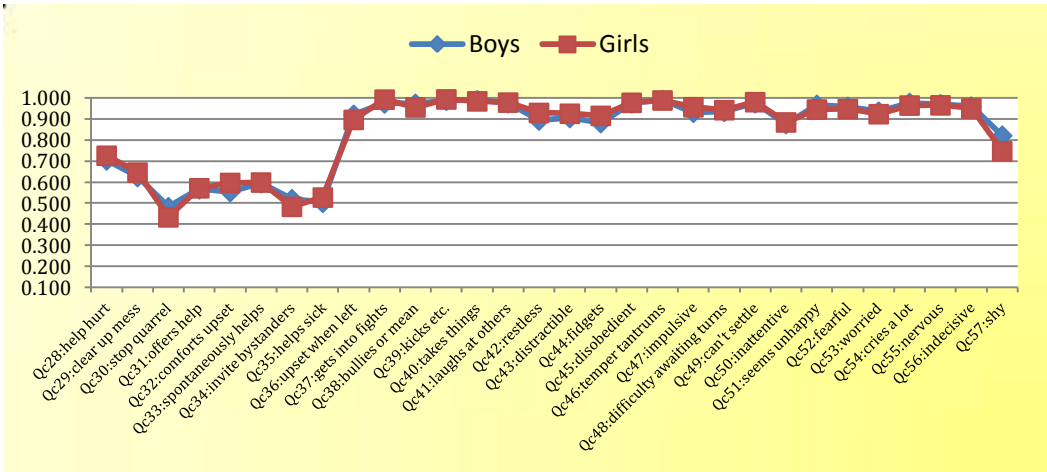
Group 1: (p<=0.7000)



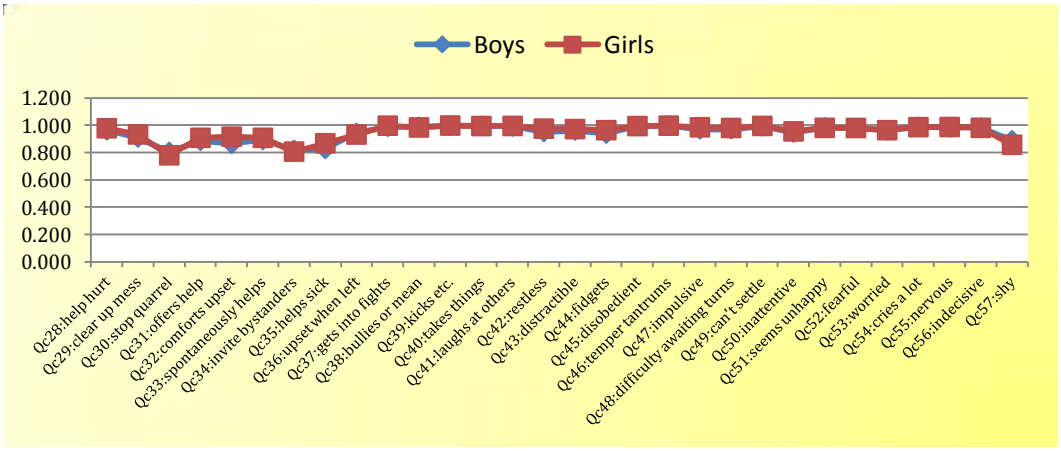
Group 2: (0.70001<p<0.80000)



Group 3: (0.80001<p<0.86667)

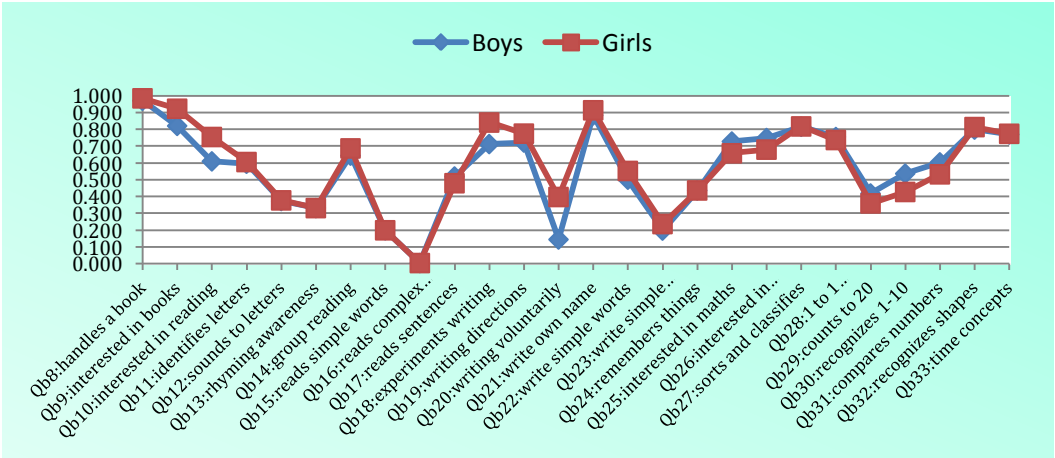


Group 4: (p>=0.86668)

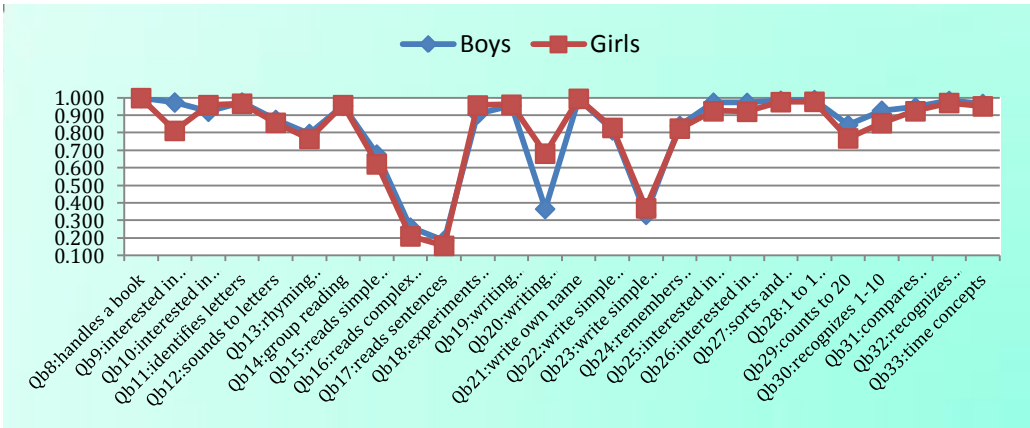


D: Language & thinking skills

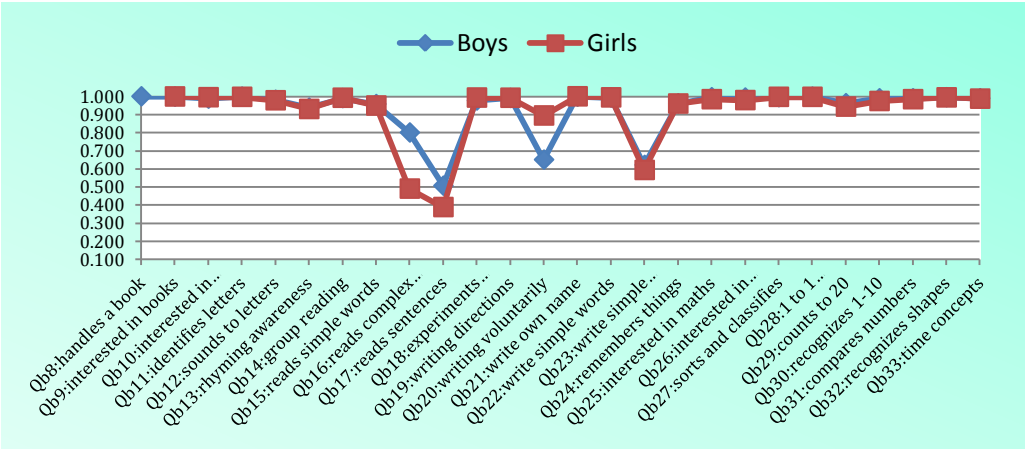
Group 1: (p<=0.73077)



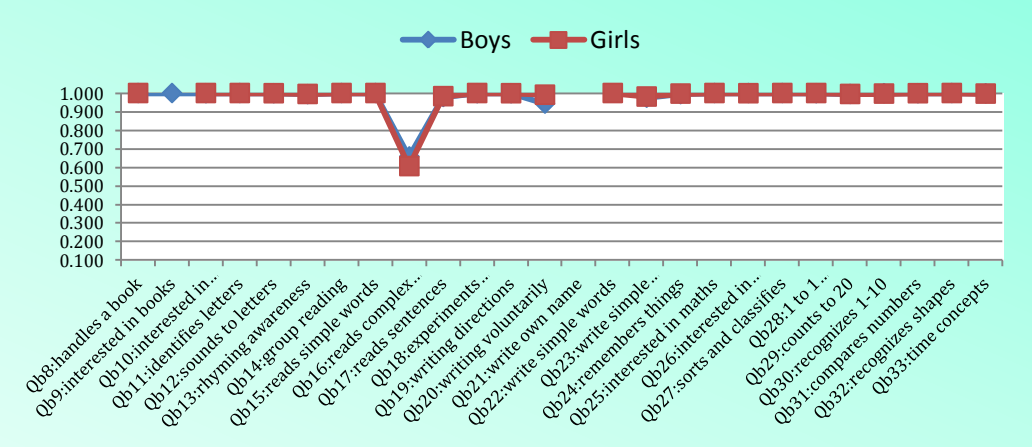
Group 2: (0.73078<p<0.87500)



Group 3: (0.87501<p<0.92308)

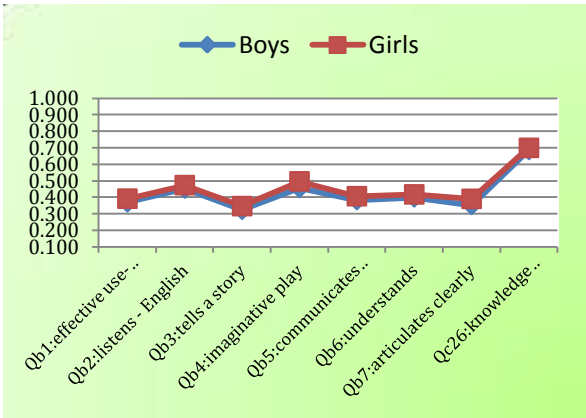


Group 4: (p>=0.92309)

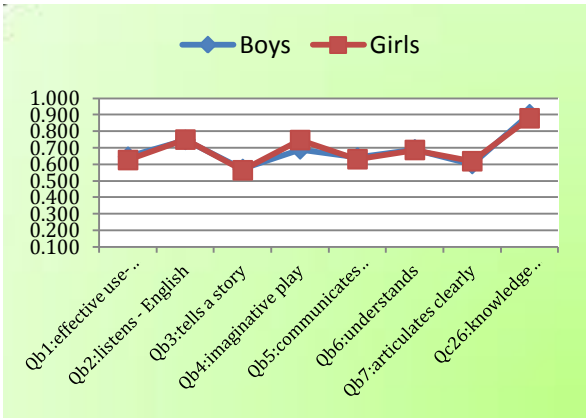


E: Communication & GK

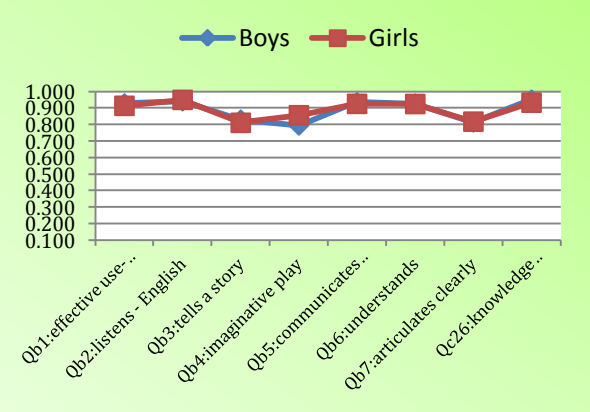
Group 1: (p<=0.56250)



Group 2: (0.56251<p<0.75000)



Group 3: (0.75001<p<0.93750)



Group 4: (p>=0.93751)

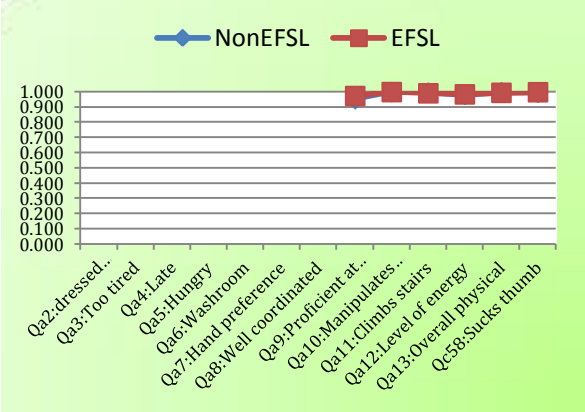
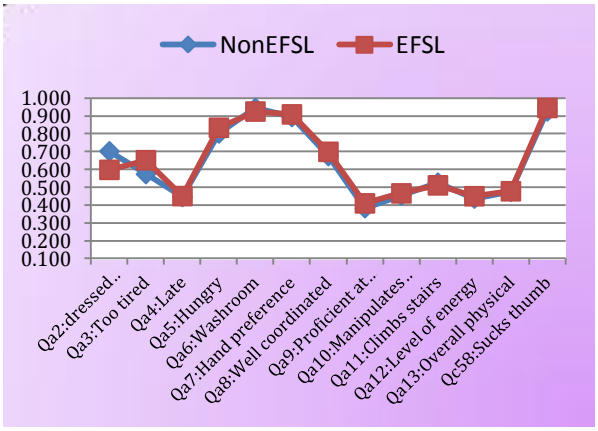


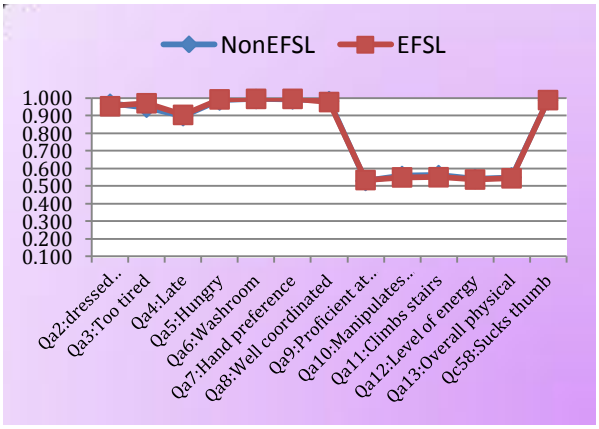
Figure 5: DIF analysis by EFSL status of child

A: Physical health & well-being

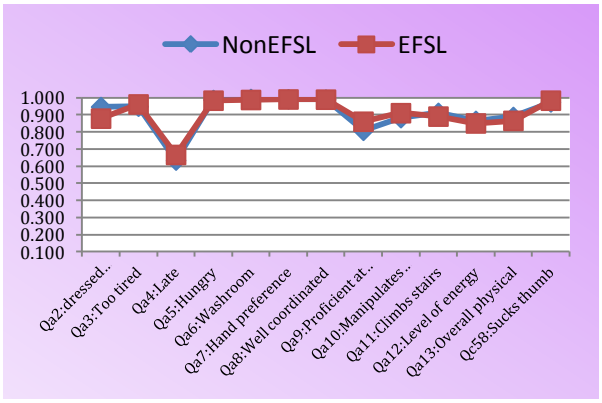
Group 1: (p<=0.76923)



Group 2: (0.76924<p<0.84615)



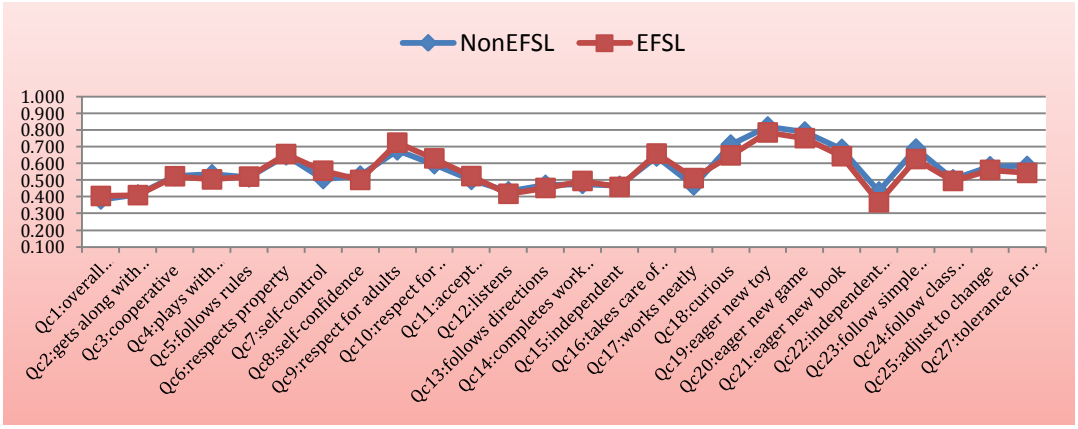
Group 3: (0.84616<p<0.92308



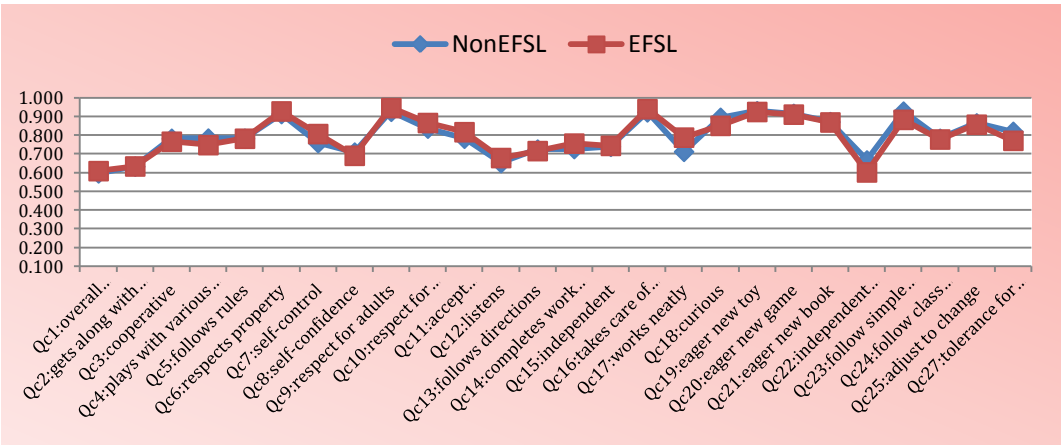
Group 4: (p>=0.92309) All items constant

B: Social competence

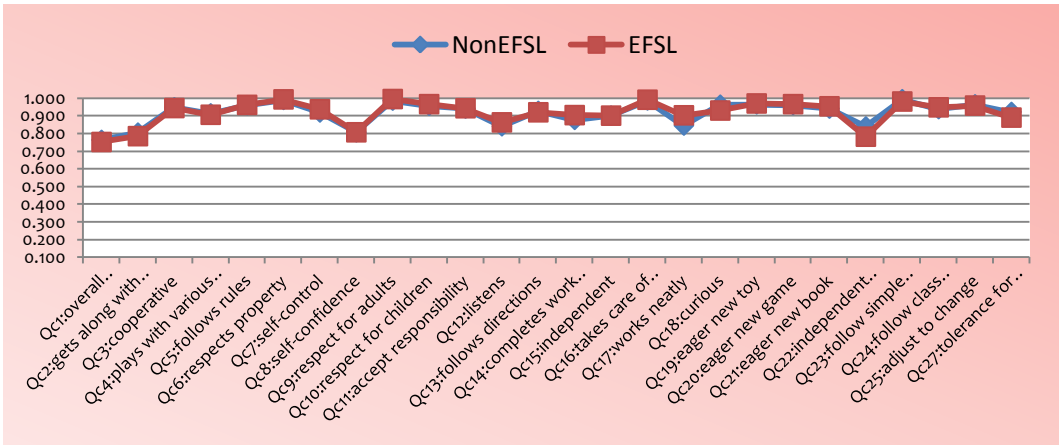
Group 1: (p<=0.71154)



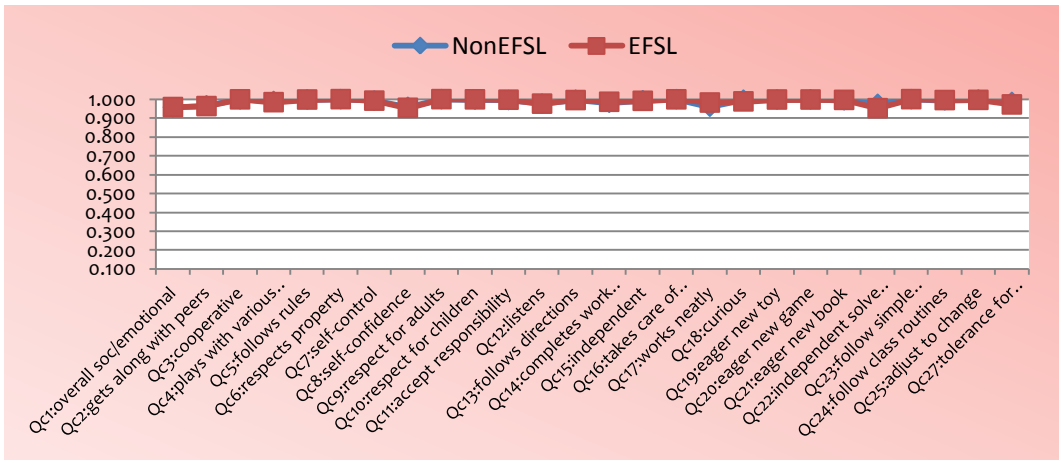
Group 2: (0.71155<p<0.86538)



Group 3: (0.86539<p<0.94231)

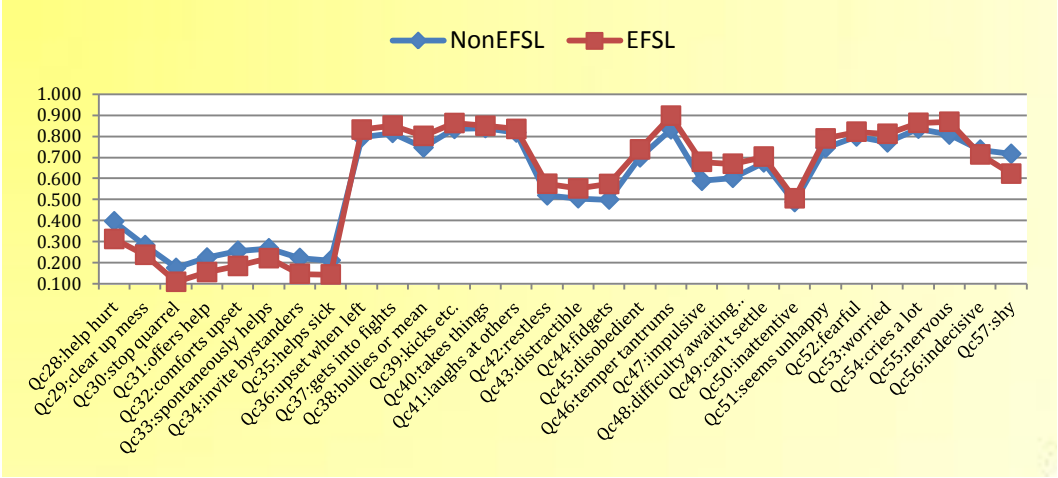


Group 4: (p>=0.94232)

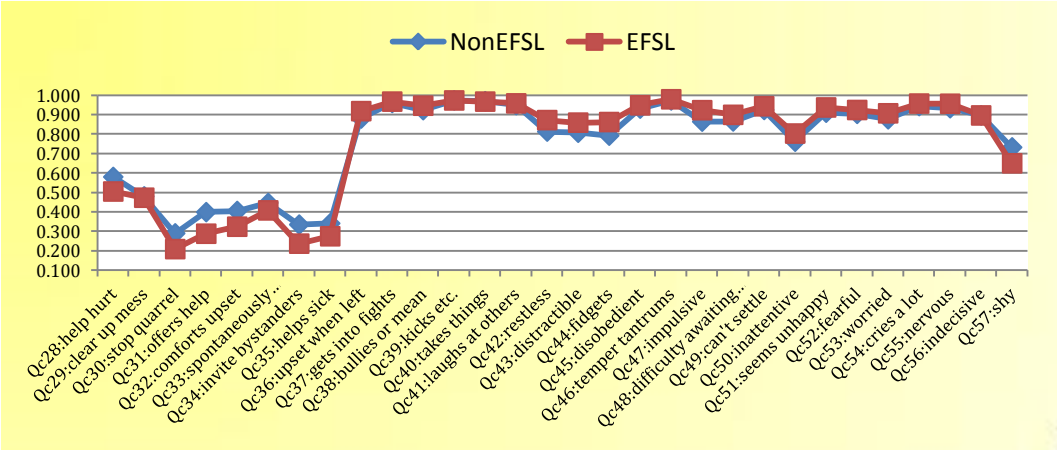


C: Emotional maturity

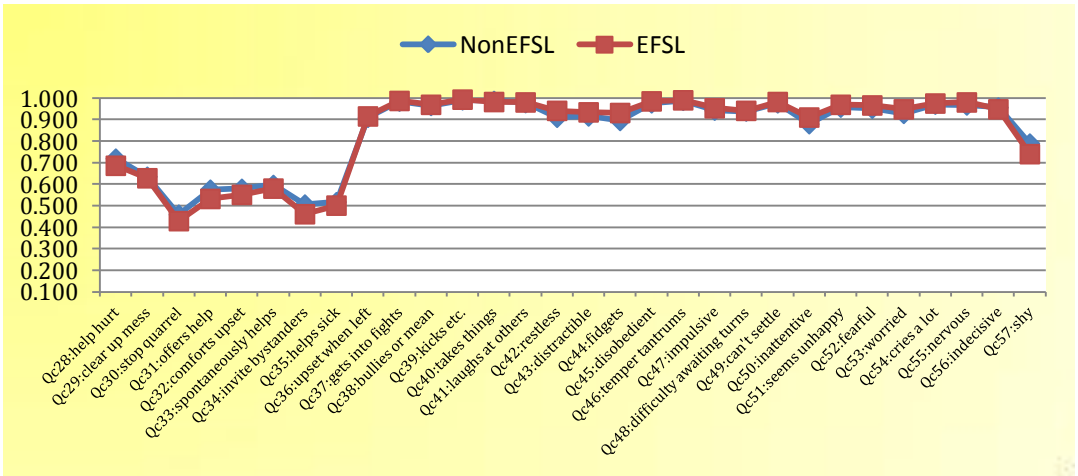
Group 1: (p<=0.70000)



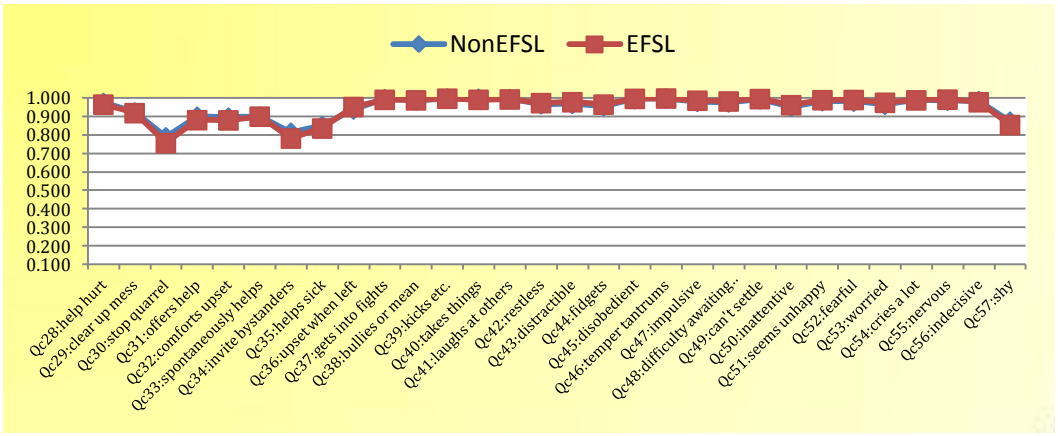
Group 2: (0.70001<p<0.80000)



Group 3: (0.80001<p<0.86667)

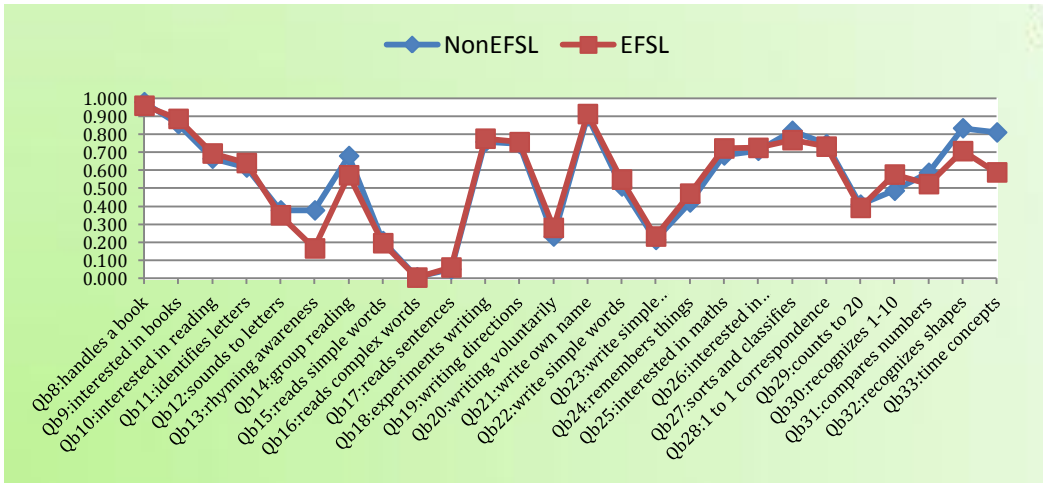


Group 4: (p>=0.86668)

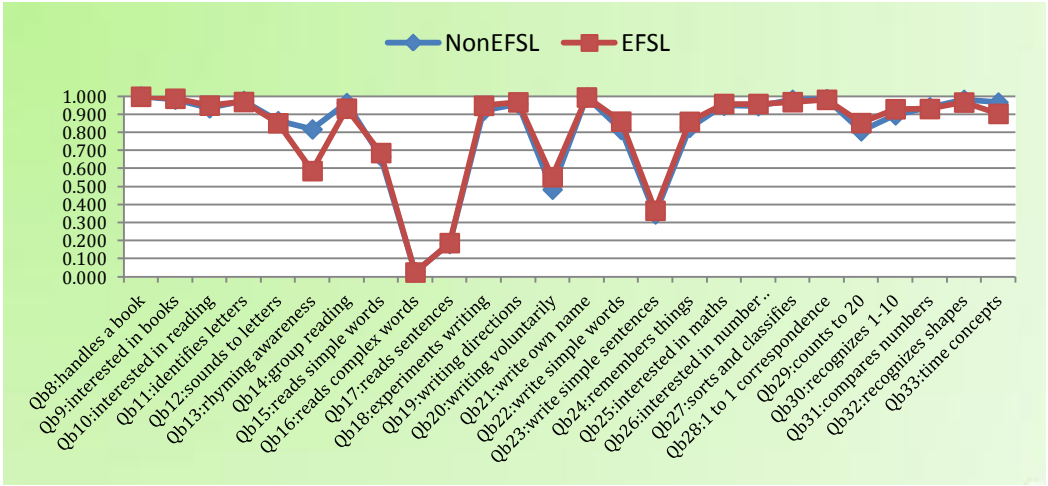


D: Language & thinking skills

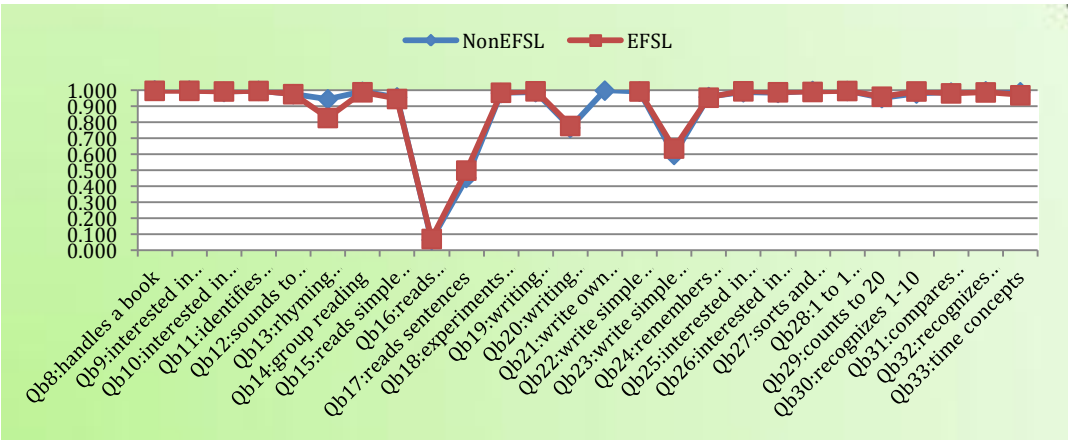
Group 1: (p<=0.73077)



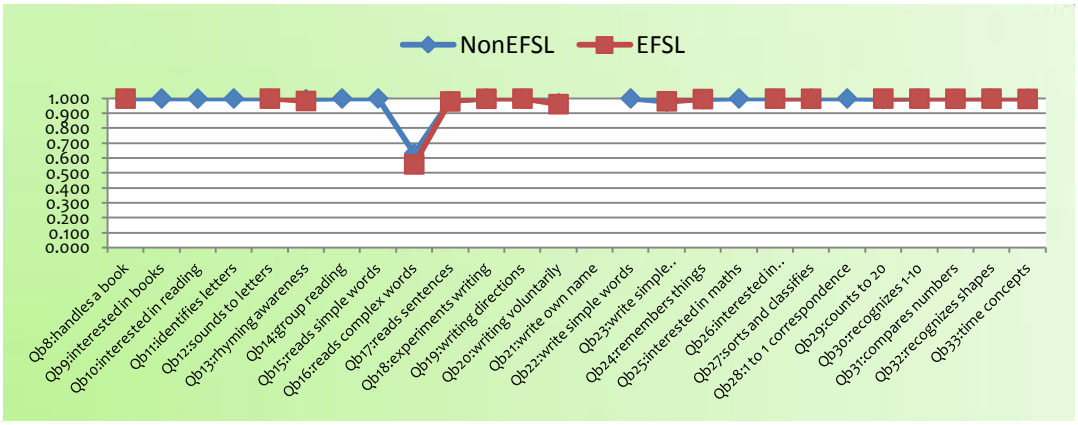
Group 2: (0.73078<p<0.87500)



Group 3: (0.87501<p<0.92308)

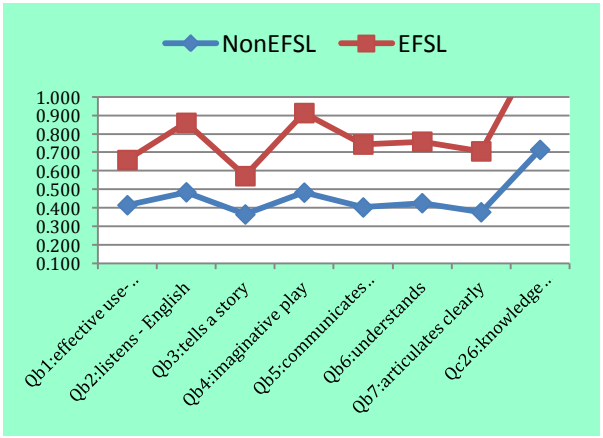


Group 4: (p>=0.92309)

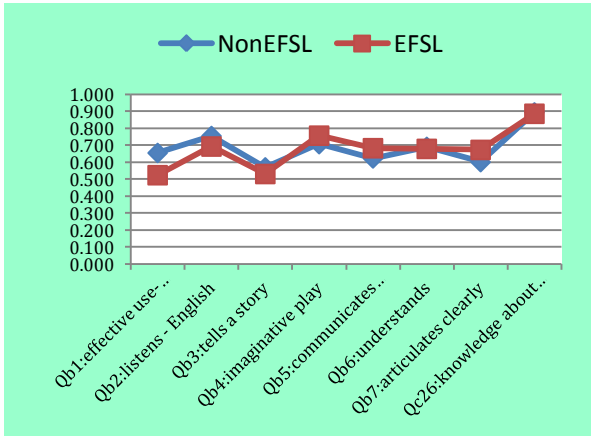


E: Communication & GK

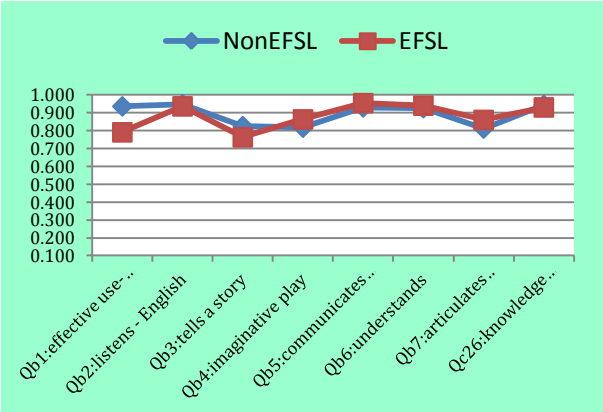
Group 1: (p<=0.56250)



Group 2: (0.56251<p<0.75000)



Group 3: (0.75001<p<0.93750)



Group 4: (p>=0.93751) All items constant

Reliability

If test scores are to be used to make inferences about proficiency, they must be both reliable and valid. Reliability refers to the ability of a test to produce stable and consistent results. It is a prerequisite for validity, although not all highly reliable tests can be valid. It is quite possible for a tool to measure the wrong thing but show a high level of consistency. Our focus here is on reliability, not validity because it is an important first step in determining the validity.

It is clear from the definitions stated earlier that reliability coefficients can very well be affected by at least three important factors: group homogeneity, test length, and test difficulty. From our earlier discussions on the relationship between observed score, true score, and error, we can also say that the magnitude of reliability depends on variations among individual performers. The more the variability, the higher is the reliability. The more homogeneous the test content, the higher the reliability coefficient. The more questions/items, the higher is the reliability. Generally, the more the number of moderately difficult items, the better the reliability. The number of response options could also influence reliability; the more options, the higher is the reliability.

Of the four general classes of reliability estimates researchers use, test-retest (coefficient of stability), parallel forms (coefficient of equivalent forms), inter-rater and internal consistency, our examination of reliability is focused on the internal consistency reliability. More specifically, we focus on split-half reliability, item-total correlations, Cronbach's alpha, and Spearman-Brown formula, a formula that is used to correct the correlations between, say total scores on odd items and the scores obtained on even items to the full test.

This is also an occasion to review the statistics involved in assessing the internal consistency reliability. Much more than Cronbach's alpha is needed to evaluate internal consistency, as described below. The overall assessment centers around three ideas: (1) split-half reliability, (2) item-total correlations, (3) Cronbach's alpha, and (4) Spearman-Brown formula. Why use split-half reliability? It is a test for internal consistency and is a useful measure when it is impractical to assess reliability based on two test administrations or when we are stuck with a single test administration. Why use item-total correlations? How can we possibly add unrelated things into a total? That is, if an item is uncorrelated with other item, it does not contribute to the internal consistency reliability of the total score. Put simply, reliability increases with high item-total correlations. For example, if the reliability is 0.49, then the estimated correlation is 0.70 (=sqrt (0.49)). Why use Cronbach's alpha? Alpha, as we noted earlier is a lower-bound estimate of reliability under the assumption that all items are with uncorrelated errors. We had already noted that it can be used for any mixture of binary (true/false) and partial credit items (true/sometimes/false). If the items are measuring the same construct (i.e., not multi-dimensional), they should show identical responses, although alpha itself does not detect dimensionality issue. Why use Spearman-Brown formula? In the split-half reliability method, we will get the correlation of scores between the two halves by using the usual Pearson correlation formula. But it only estimates the reliability of each half of a test. It is necessary to use a statistical correction to estimate reliability of the whole test. Thus, by using Spearman-Brown formula, we can adjust or re-evaluate the correlation we obtained. If we assume that the longer the test, the more reliable it is, the test that has been shortened from the split-half method could underestimate the reliability of the whole test. The formula can also be applied to see approximately how many items we need for desired reliability under CTT.

In Table 9, the item-total and reliability statistics are presented based on the split-half model. If the 13-item physical health & well-being area is cut to a 6-item one by keeping only the items that have high point-biserial correlations (that is, taking out Qa2, Qa3, Qa4, Qa5, Q6, a7, & Qa8), reliability will be 0.865. In the physical health & well-being area, Qa6 and Qc58 are the two most problematic items because they have the lowest point-biserials. This would suggest that they behave very differently from other items or they simply are measuring another content area. Also, the reliability is highest when Qa4 is deleted. Even though the item has a high *p* value (0.79), its point-biserial being low, it appears to have no quality. The Spearman-Brown coefficient is 0.523, making it difficult to trust the Cronbach's alpha value of 0.782 for the whole set (Table 6). Other highly problematic items that do not fit the content area, based on point-biserials, include Qc36 and Qc57 in emotional maturity and Qb8 in language & thinking skills. Items that are very easy and answered correctly by an

overwhelming majority of children will have poor point-biserial correlations, and such items require more thorough psychometric analysis.

Table 9: The five developmental areas with 103 items and item-total and reliability statistics, split-half method, Merger #3, Alberta (N=52,035)

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's Alpha if item deleted
Physical health & wellbeing (13)				
Qa2:dressed inappropriately	104.11	289.637	.288	.781
Qa3:Too tired	104.29	277.405	.377	.772
Qa4:Late	105.28	280.284	.214	.802
Qa5:Hungry	103.67	299.942	.298	.777
Qa6:Washroom	103.37	318.154	.138	.786
Qa7:Hand preference	103.46	311.444	.224	.782
Qa8:Well coordinated	103.90	285.227	.403	.769
Qa9:Proficient at holding pen	105.89	255.109	.581	.749
Qa10:Manipulates objects	105.34	257.305	.671	.741
Qa11:Climbs stairs	105.10	265.166	.634	.746
Qa12:Level of energy	105.49	256.358	.663	.741
Qa13:Overall physical	105.32	258.323	.698	.739
Qc58:Sucks thumb	103.49	317.321	.132	.786
Cronbach's Alpha: Part 1 (7 items) =0.549; Part 2 (6 items): 0.865; S-B coefficient=0.523				
Social competence (26)				
Qc1:overall soc/emotional	212.57	1838.988	.670	.952
Qc2:gets along with peers	212.31	1844.318	.682	.952
Qc3:cooperative	211.36	1861.963	.737	.951
Qc4:plays with various children	211.47	1868.690	.670	.952
Qc5:follows rules	211.35	1859.320	.757	.951
Qc6:respects property	210.75	1898.615	.689	.952
Qc7:self-control	211.50	1860.114	.699	.951
Qc8:self-confidence	211.97	1883.097	.548	.953
Qc9:respect for adults	210.65	1911.968	.654	.952
Qc10:respect for children	211.07	1886.586	.678	.952
Qc11:accept responsibility	211.42	1853.920	.718	.951
Qc12:listens	212.13	1841.524	.712	.951
Qc13:follows directions	211.65	1845.053	.762	.951
Qc14:completes work on time	211.79	1854.856	.647	.952
Qc15:independent	211.71	1843.785	.700	.951
Qc16:takes care of materials	210.74	1897.944	.689	.952
Qc17:works neatly	211.92	1860.756	.620	.952
Qc18:curious	210.75	1923.426	.548	.953
Qc19:eager new toy	210.42	1961.898	.420	.954
Qc20:eager new game	210.51	1950.876	.455	.954

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's Alpha if item deleted
Qc21:eager new book	210.89	1916.621	.531	.953
Qc22:independent solve problems	212.14	1840.741	.685	.952
Qc23:follow simple instructions	210.65	1908.099	.665	.952
Qc24:follow class routines	211.43	1855.201	.720	.951
Qc25:adjust to change	211.04	1878.657	.691	.952
Qc27:tolerance for mistake	211.31	1887.193	.597	.952
Cronbach's Alpha: Part 1 (13 items) =0.934; Part 2 (13 items)=0.906; S-B coefficient=0.890				
Emotional maturity (30)				
Qc28:help hurt	235.18	1781.772	.625	.911
Qc29:clear up mess	235.94	1760.515	.649	.910
Qc30:stop quarrel	237.45	1763.037	.614	.911
Qc31:offers help	236.51	1746.195	.677	.909
Qc32:comforts upset	236.40	1758.682	.643	.910
Qc33:spontaneously helps	236.20	1761.429	.649	.910
Qc34:invite bystanders	237.09	1769.826	.614	.911
Qc35:helps sick	236.92	1756.141	.639	.910
Qc36:upset when left	233.32	1924.661	.176	.917
Qc37:gets into fights	232.75	1902.192	.425	.914
Qc38:bullies or mean	233.04	1884.704	.446	.914
Qc39:kicks etc.	232.66	1905.937	.434	.914
Qc40:takes things	232.69	1908.409	.407	.914
Qc41:laughs at others	232.80	1905.803	.405	.914
Qc42:restless	233.96	1815.375	.570	.912
Qc43:distractible	234.01	1807.919	.591	.911
Qc44:fidgets	234.11	1812.869	.568	.912
Qc45:disobedient	233.09	1861.571	.564	.912
Qc46:temper tantrums	232.65	1905.729	.425	.914
Qc47:impulsive	233.57	1831.830	.570	.912
Qc48:difficulty awaiting turns	233.57	1837.298	.553	.912
Qc49:can't settle	233.15	1851.858	.564	.912
Qc50:inattentive	234.27	1809.166	.598	.911
Qc51:seems unhappy	233.10	1886.691	.422	.914
Qc52:fearful	233.03	1905.816	.330	.915
Qc53:worried	233.25	1903.350	.315	.915
Qc54:cries a lot	232.80	1913.611	.331	.915
Qc55:nervous	232.88	1906.500	.350	.915
Qc56:indecisive	233.20	1880.998	.436	.914
Qc57:shy	234.38	1922.296	.135	.919
Cronbach's Alpha: Part 1 (15 items)=0.896; Part 2 (15 items)=0.862; S-B coefficient=0.705				
Language & thinking skills (26)				
Qb8:handles a book	209.61	1990.231	.213	.901
Qb9:interested in books	209.87	1946.277	.356	.899
Qb10:interested in reading	210.37	1877.266	.503	.897

Item	Item-total statistics			
	Scale mean if item detected	Scale variance if item deleted	Corrected item-total correlation	Cronbach's Alpha if item deleted
Qb11:identifies letters	210.43	1849.124	.605	.895
Qb12:sounds to letters	211.14	1791.729	.645	.893
Qb13:rhyming awareness	211.52	1787.453	.598	.894
Qb14:group reading	210.35	1872.148	.532	.896
Qb15:reads simple words	211.97	1750.715	.657	.893
Qb16:reads complex words	217.16	1848.994	.373	.901
Qb17:reads sentences	214.57	1759.082	.525	.897
Qb18:experiments writing	210.21	1918.163	.370	.899
Qb19:writing directions	210.18	1898.851	.471	.897
Qb20:writing voluntarily	212.72	1791.063	.487	.898
Qb21:write own name	209.78	1954.188	.364	.900
Qb22:write simple words	210.89	1839.676	.525	.896
Qb23:write simple sentences	213.49	1790.298	.460	.899
Qb24:remembers things	211.14	1813.225	.572	.895
Qb25:interested in maths	210.29	1881.631	.512	.897
Qb26:interested in number games	210.26	1891.581	.477	.897
Qb27:sorts and classifies	209.98	1920.757	.450	.898
Qb28:1 to 1 correspondence	210.11	1893.046	.533	.897
Qb29:counts to 20	211.29	1803.027	.581	.895
Qb30:recognizes 1-10	210.84	1824.037	.591	.895
Qb31:compares numbers	210.58	1844.625	.577	.895
Qb32:recognizes shapes	210.01	1919.851	.438	.898
Qb33:time concepts	210.14	1912.898	.419	.898
Cronbach's Alpha: Part 1 (13 items) =0.823; Part 2(13 items) =0.831; S-B coefficient=0.871				
Communication & GK (8)				
Qb1:effective use-English	54.35	298.563	.850	.917
Qb2:listens - English	53.91	314.386	.792	.922
Qb3:tells a story	54.78	294.442	.838	.918
Qb4:imaginative play	54.19	322.195	.695	.929
Qb5:communicates needs	54.29	301.028	.844	.918
Qb6:understands	54.19	304.610	.821	.919
Qb7:articulates clearly	54.64	304.994	.724	.927
Qc26:knowledge about world	53.07	345.464	.562	.937
Cronbach's Alpha: Part 1 (4 items)=0.897; Part 2(4 items) =0.852; S-B coefficient=0.928				

What is your takeaway?

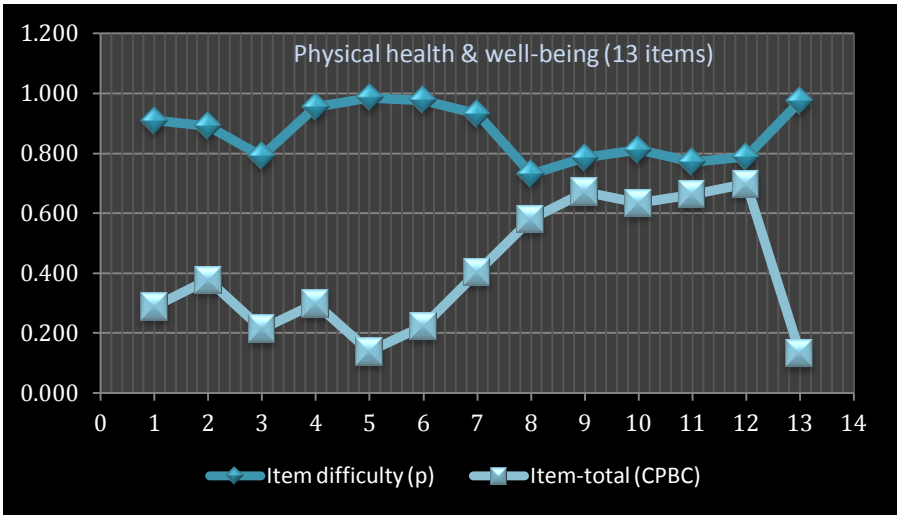
When it is impractical or undesirable to assess reliability with two tests or two administrations of the same test, a sub-type of internal consistency reliability, namely split-half reliability comes to the rescue. It divides the test into half, find the correlation between the two, and adjust the correlation mathematically using the Spearman-Brown formula to estimate the correlation of the whole test. The very low item-total correlations and the reliability coefficient values suggest the removal of at least five items (Qa6: washroom; Qc58: sucks thumb; Qc36: upset when left; Qc57: shy; and Qb8: handles a book) from the test to improve the quality of the whole test.

More on graphical analysis: Item difficulty and discrimination coefficients

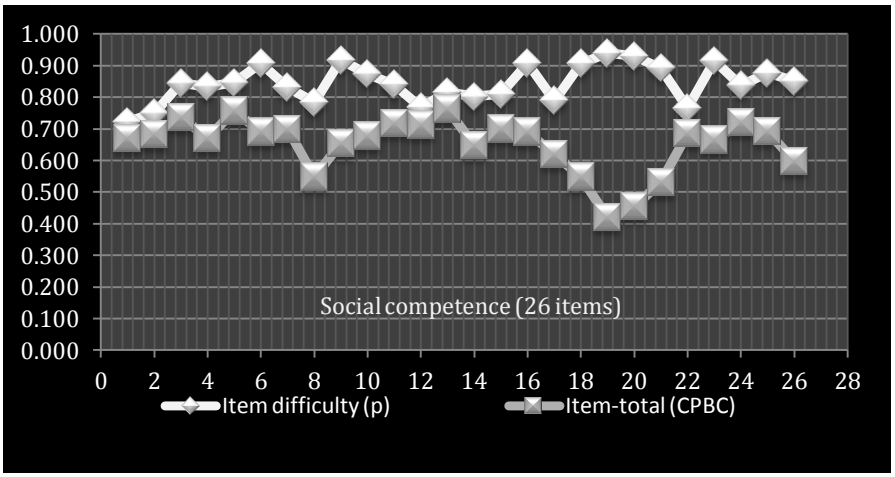
To compare the patterns of item difficulty and item discrimination across items, scatter plots were created (see Figure 6). The x-coordinate is represented by the p-value of the item and the y-coordinate the item-test correlation. Two items with graphical representation near each other have approximately the same p-value and the same discrimination. It is also a check to find out the extent to which the formation of parallel forms in split-half scenario has been successful. For example, in the physical health & well-being area, items, Qa8 to Qa13 form a pair, where item content, its format (with the exception of perhaps, Qa8), proportion endorsing the item (*p*-value) and discrimination (Corrected item-total correlation) are almost similar. The items that show some kind of imbalance in terms of them having conflicting difficulty and discrimination indices in other areas include: Qc8 to Qc10, Qc17 to Qc21, and Qc27 in the social competence area; Qc36 to Qc49 and Qc51 to Qc57, in the emotional maturity area, Qb8 to Qb11, Qb14, Qb18 to Qb22, and Qb24 to Qb33 in the language & thinking skills area, and Qc26 in the communication and general knowledge area.

Figure 7 gives a graphical view of behaviour of 10 items, chosen based on their difficulty levels and item-total correlations. They all have below thresh-hold level item-total correlations, indicative of some problem in those items. These items may have some kind of psychometric imbalance with respect to content, wording, or some other characteristic that impacts the correlation between the item and the content area. Psychometric imbalance, probably cannot be taken as a deciding factor in test construction, however, the graphical representation gives some additional insight into constructing parallel forms from a single administration because a necessary condition for parallelism is that the contents should be comparable.

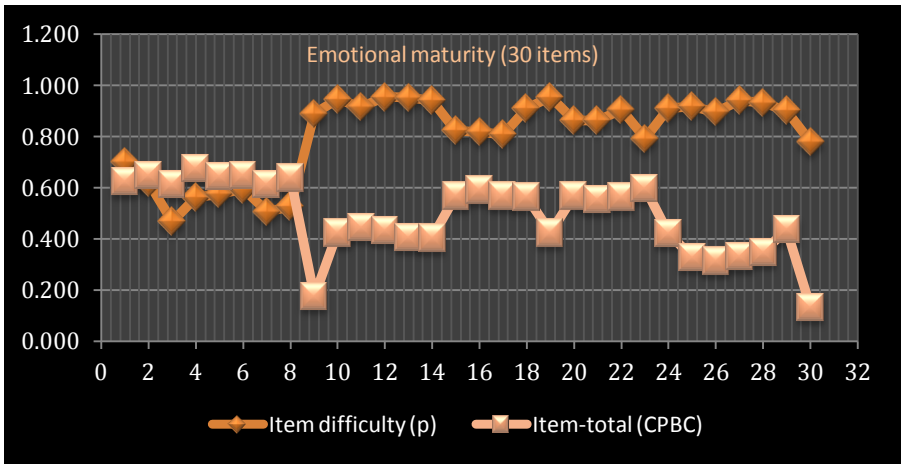
Figure 6: Graphical representation of the relationship between item difficulty and discrimination (item-total correlations)



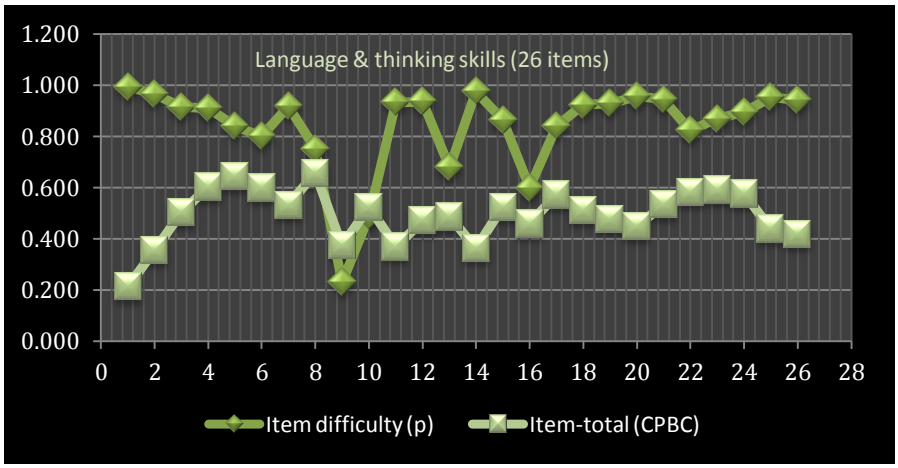
Numbers 1 to 7 and 13 represent the items, Qa2 to Qa7 and Qc58. Difficulty and discrimination do not go hand in hand for these items.



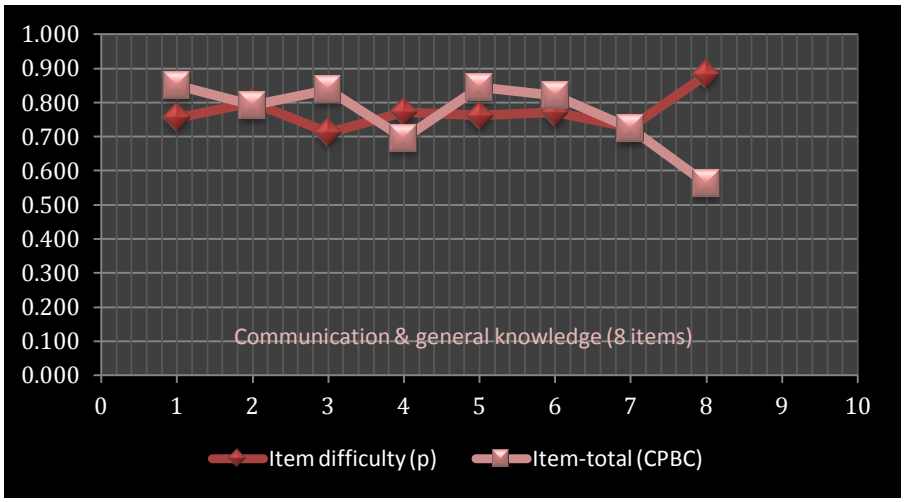
Numbers 8 to 10, 17 to 21, and 26 represent the items, Qc8 to Qc10, Qc17 to Qc21, and Qc27. Difficulty and discrimination do not go hand in hand for these items.



Numbers 9 to 22 and 24 to 30 represent the items, Qc36 to Qc49 and Qc51 to Qc57. Difficulty and discrimination do not go hand in hand here.

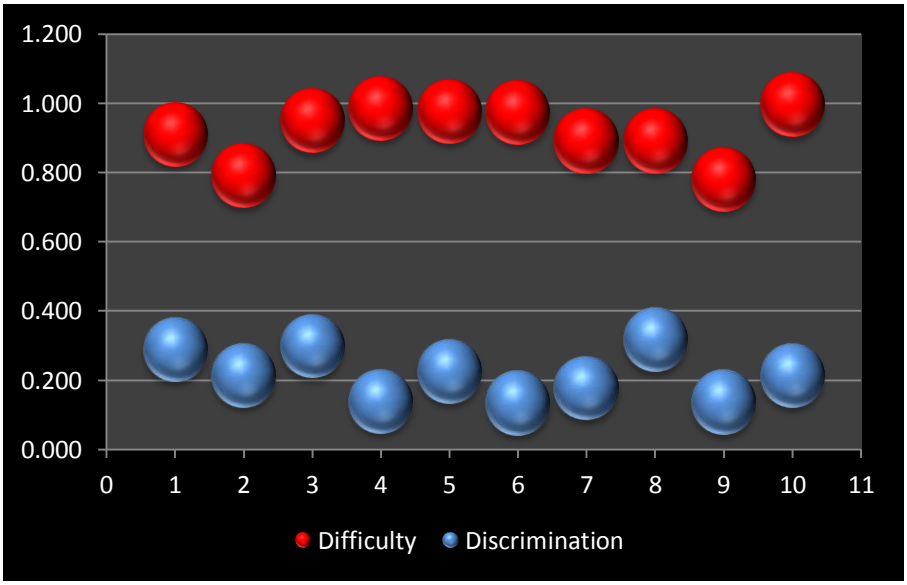


Numbers 1 to 4,7, 11 to 15 and 17 to 26 represent the items, Qb8 to Qb11, Qb14, Qb18 to Qb22 and Qb24 to Qb33. Difficulty and discrimination do not go hand in hand for these items.



Number 8 represents the item, Qc26. Difficulty and discrimination do not go hand in hand for the item.

Figure 7: Graphical representation of selected items with conflicting difficulty and discrimination indices



What is your takeaway?

What is the relationship between an item's difficulty level and its discrimination? Ideally, those who had the highest test scores were more likely to get the items correct than those with low overall scores. Here, many items are very poor in discriminating; those who scored most highly on the test overall were not likely to get the item correct, whereas those with low overall scores were likely to get the item correct. Examples include: Qa2 to Qa7, Qc58 in the physical health & well-being area and Qc26 in the communication and general knowledge area.

Standard Error of Measurement (SEM)

From our earlier discussion of SEM, we know that given a fixed value of sample standard deviation, the higher the reliability, the smaller the standard error of measurement. In Table 10, the SEM values and confidence intervals, calculated within the CTT framework are shown for each area. The social competence area demonstrated excellent internal consistency (0.954) with a SEM value of 0.3738 or the social competence area is the most reliable among the five developmental areas.

The results should be interpreted with some caution, however. A potential problem with reliability is that it is a property of scores. That is, a given coefficient alpha, 0.954 indicates that, with the sample studied on that particular occasion, the observed scores have a particular proportion of their variance that can be attributed to true scores (DeVellis, 2006). It does not guarantee that scores for the same set of 26 items given to a group, say to a group of German children under somewhat different circumstances would yield the same coefficient alpha. Coefficient alpha, as we had noted earlier, is determined, among other things, by two quantities, the test length and the strengths of correlations among the test questions, and consequently a change in either of these could impact the alpha value.

Table 10: Standard Error of Measurement (SEM) for all five areas of development

	Observed score	SD of observed score	Reliability	SQRT(1-Reliability)	SEM
Physical health & well-being	8.690	1.4018	0.782	0.4669	0.6545
Social competence	8.441	1.7428	0.954	0.2145	0.3738
Emotional maturity	8.086	1.4640	0.915	0.2915	0.4268
Language & thinking skills	8.428	1.7296	0.901	0.3146	0.5442
Communication & general knowledge	7.726	2.5171	0.933	0.2588	0.6515

Cross-area comparisons of perfect scorers

In comparing the developmental areas in terms of response options, proportionately more children scored perfect 10s in the communication & general knowledge than other areas (Table 11). This is not surprising because of the small number of items within the area. There were none in the emotional maturity area who scored perfect scores in all the items. It is interesting to note that none scored zeros or 10's in all five areas (not shown in the Table), although three scored 5's in all areas.

Table 11: Perfect scorers vs. others, Merger #3, Alberta (N=52,035)

	All 0's^	All 5's*	All 10's#	Total
Physical health & well-being	1	278	16292	16,571 (28.53%)
Social competence	2	333	10833	11,168 (19.23%)
Emotional maturity	0	6	na	6 (0.01%)
Language & thinking skills	16	na	8725	8,741 (15.05%)
Communication & general knowledge	255	1926	19424	21,605 (37.19)
Total@	274	2543	55274	58,091 (100.00%)

^: 0 for each and every item within the area; *: 5 for each and every item within the area; and #: 10 for each and every item within the area. That is, if a child scored different combinations of 0, 5, & 10, they are not included.
na: not available for the emotional maturity area and not applicable for the language & thinking skills area.
@: The totals may include some children scoring the same in other areas, for example, a child may score all 10s in the physical area and he/she can be counted again if he/she scored 10's in the social competence area.

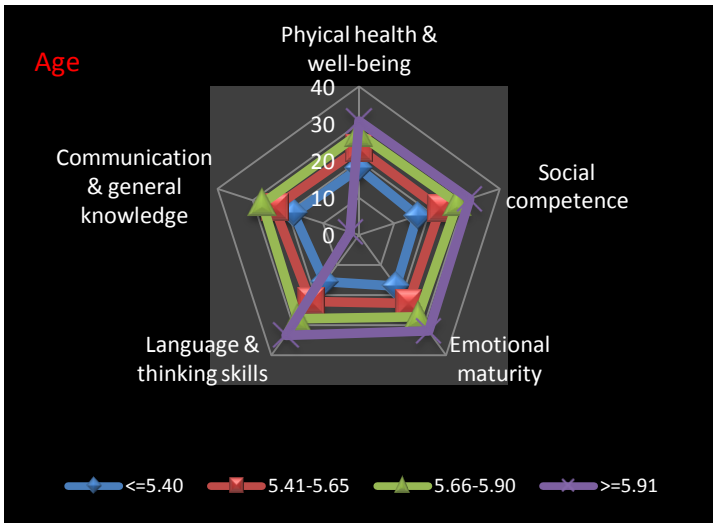
What does the data tell us about differences in background characteristics of perfect scorers? Figure 8 shows the differences between the four age groups (quartiles), boys and girls, Canadians and First Nations, non-E/FSL and E/FSL children who were perfect scorers (10s in all items within the area). Percentages of children who got all correct varied between age groups; the older the children, the better the likelihood of getting all items correct in an area, except for the communication & general knowledge area.

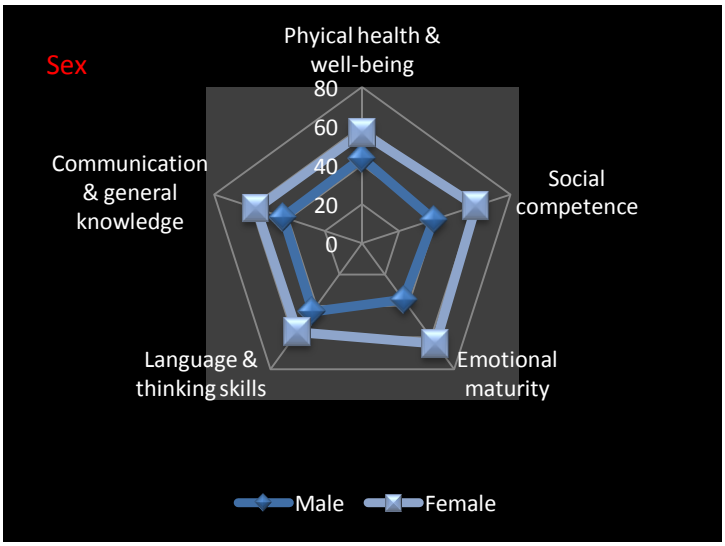
Girls are proportionately more than boys to get all items correct. The gaps get wider in the areas of social competence and emotional maturity. A conclusion which can be drawn from our results is that the largest share of overall difference between boys and girls in each item

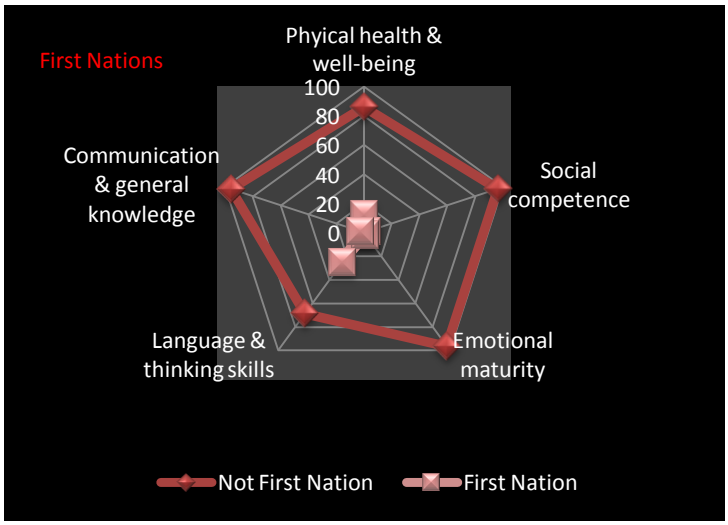
is accounted for by differences in social and emotional development, as girls outperform boys, again confirming our earlier findings.

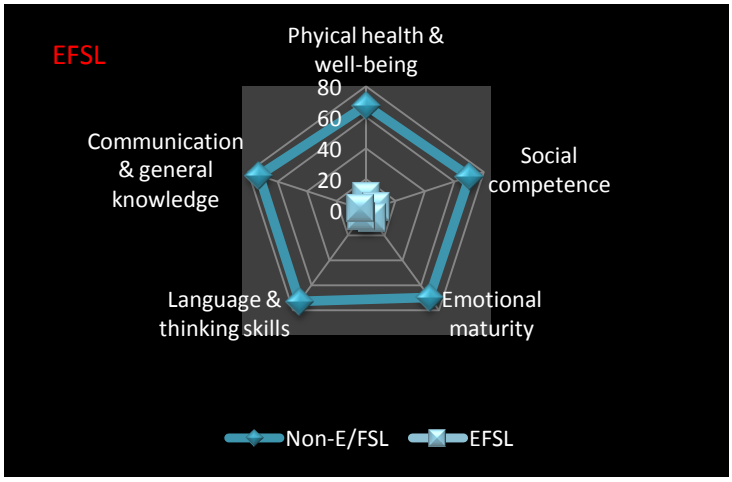
The tendency for First Nations children to perform worse than their Canadian counterparts should be read with some caution, as there is a clear under-representation of children of Aboriginal origins. We omitted this variable from our DIF analysis for the same reason. A similar picture emerged for E/FSL status. The results suggest that although there are differences between non-E/FSL children and E/FSL children of perfect scorers, the differences become more pronounced among low scorers (Figure 5).

Figure 8: Cross-area comparisons of perfect scorers by age, sex, First Nations status, and E/FSL status









What is your takeaway?

Test takers when grouped into four homogeneous groups based on the proxy for the levels proficiency (*p* -values), even among low scorers, a number of items are equally easy for both boys and girls and for EFSL and non-EFSL children. Among the lowest scoring group ($p<0.563$), **all items** in the area of communication and general knowledge favor non-EFSL children. A puzzling question is: is the presence of implausible distractors making most items artificially much easier than it ought to be, while contributing to little or no discrimination between high scorers and low scores?

There are age and sex differences among perfect scorers; perfect scorers comprised of proportionately older children and girls. No definitive conclusions can be drawn based on children's Aboriginal or E/FSL status because of their small numbers in the sample of perfect scorers.

Discussion and Conclusions

Several methodological challenges have been encountered in the course of this work. Some have been overcome, but some remain. First, various opinions about the suitability of CTT as a measurement device are making researchers concerned about the best approach to psychometrically evaluate a test, especially for a non technical reader. Item difficulty indices, generated through CTT and IRT are very comparable, and may give similar values under the two approaches (Macdonald & Paunonen, 2002; Sohn, 2009). We had no way of providing support for this claim because the scope of this study was limited to CTT. Second, procedures associated with CTT vary from investigator to investigator. A series of analytical procedures were attempted in the present study to provide information from different angles in order to assist researchers and tool developers in making decisions on the potential future of EDI application. Finally, there are suggestions that that we must pay attention to both incorrect and correct responses to each item because one of the many purposes of item analysis is to detect the anomalies (Wainer, 1989). However, medium level responses were not analyzed in our study.

Many issues, especially those that are fundamental to CTT remain as long as explanations are intended for unidimensional or multidimensional models. Future studies using diverse sample sizes and different data sets could add more insights into many unanswered questions. More specifically, the graphical analyses to detect DIF did not show considerable variation across different groups. It is also recommended to use an effect size because the effect size can detect differences better in large samples. Some highlights from our analyses are noted below.

High levels of proficiency levels and low levels of discrimination: The difference between top and bottom performers, on average, tended to be much less pronounced. Of all the 103 items in the EDI, 82.5 per cent of them were either too easy or easy. Kindergartners in Alberta displayed high levels of proficiency in all developmental areas, with the proportions giving correct responses ranging from 23.07 (Qb16: reads complex words) to 98.50 (Qa6: washroom). Several items displayed conflicting results when relationships between p-values (measured as the proportion of test takers endorsing the correct response) and the total scores with that item removed were examined. In particular, seven items (Qa4: late; Qa6: washroom; Qa7: hand preference; Qc58: sucks thumb; Qc36: upset when left; Qc57: shy; and Qb8: handles a book) showed very low discrimination and basing on their reliability

coefficients, we suggest removing these items in order to increase the overall quality of the test.

Surprisingly, for all items, the response option “wrong” came out as a positive distractor as indicated by point bi-serial correlations between item scores on the distractor and the scores on the whole test. If distractors are selected by only a handful of individuals, it means the content area was well understood by an overwhelming majority, causing the distractors to behave like the key (correct response). Typically, instruments must deal with a broad range of items with room for distractors.

There appears to have an implicit bias toward selecting the correct response option, potentially skewing some of the items and losing the capacity to discriminate between low scorers, medium scorers, and high scorers. This has possibly caused the ability levels well within the (very) superior ranges of performance and likely an overestimate of ability for those in the average and high average ranges.

Factors behind high vs. low proficiency levels: The graphs depicting four groups of individuals, classified according to their proficiency levels and characteristics, such as age, sex and EFSL status, showed no indication of noticeable differences among the two age groups (based on median age), sexes and the two EFSL groups, in terms of their proficiency levels. A possible explanation may be the presence of implausible distractors. However, EFSL children who are low-scorers appear to be in double jeopardy; all items in the area of communication and general knowledge seem to favor non-EFSL children of similar standing.

By examining the age-sex composition of children, it was possible to control for potential bias which may have been introduced at a macro-level. For example, if 70 percent of children were under the age of five years and six months, serious caution should be exercised in interpreting the results based on area scores. It is advisable to study in-depth the influence of age and setting of cut-off scores should take this into account.

At a provincial level, deciding where the cut-off values for each area should be is no easy task. Until now, the practice has been to adopt the national cut-off values to decide *a priori* what the values/scores should be for each area of development. However, this does not seem to make much sense. It is clear from a cross-area comparison of perfect scorers, categorized according to four age groups (based on quartiles) that the average scores do vary between age groups; younger children are at a disadvantage than their older counterparts in all but one area of development.

Internal consistency: When each area of development was divided into two groups with more or less equal number of items and the correlation between the two parts was calculated, the results confirmed the earlier finding that some items need a careful

rethinking in terms of their inclusion in the test. More specifically, at the very least, five items (Qa6: washroom; Qc58: sucks thumb; Qc36: upset when left; Qc57: shy; and Qb8: handles a book), if not more, need to be reevaluated for their quality and contribution to the overall reliability of the test.

Communication & general knowledge area: Is it a linguistically appropriate area? Acquiring a language other than mother tongue involves learning the culture that is transmitted through the language. Preserving one's mother tongue preserves one's own identity, while acquiring and being proficient in additional languages promotes one's own life chances and enhances his/her understanding of the society at large. Where ever a cultural group feels threatened or in the fear of losing their cultural heritage and sense of belongingness, meaningful and effective education in their own language can have positive impacts, not only linguistically, but also socially and psychologically.

Among the lowest scoring group, all items in the area of communication & general knowledge favor non-EFSL children. Thus, the generalizability of findings from this developmental area must be questionable. In other words, children from minority language communities placed in mixed classrooms with native speakers of English in which instruction is provided in English could score well below the national norms.

Key conclusions: While more evidence from methodologically different, longitudinal research with heterogeneous samples of children is need, this study provides a basis for EDI researchers and developers to draw some tentative conclusions of a general nature, as follows:

- If children are growing up with mother tongue, other than English or French, educational provisions need to support them if they are expected to do better in the communication and general knowledge area of development. Programs need to be targeted especially to those who are already scoring low in other areas as well.
- It follows from the point made earlier that a culturally appropriate or contextually responsive assessment tool is required rather than a 'one size fits all', based on the emergence of a new dimension, anxiety and fearfulness.
- Age-adjusted cut-off values are recommended. It is possible that some of the survey questions (e.g., Qb16: read complex words) may not be relevant, in particular to the youngest age group (e.g., <=5.40) and EFSL children.
- There is evidence for support for the tool, but our analysis showed inconsistencies in the difficulty-discrimination relationship.
- There is internal consistency from a statistical point of view, but a well-balanced structure (certain areas have three times more items than others) that include

better items (not more items) is needed to account for substantial variation within the underlying construct.

- Inconsistent scaling may have contributed to some kind of confusion or bias, especially in the physical health & well-being area. This area has seven dichotomous and six polytomous items with half of them turned out to be bad items.
- There may be many confounding issues, more importantly, length of survey and question wording. Two areas, physical health & well-being and emotional maturity have a mix of negatively- and positively-worded questions, and that may have caused some confusion for some teachers. There is a possibility that some teachers become disengaged with the content of the questions and answered generally for the whole content area if there is inconsistency in the wording of questions.[17]

[17] See Colosi (2005) for a comparison of the results of two surveys with negatively and positively worded questions.

References

Briggs S.R. & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality,* 54,106-148.

Cohen, A. (2009). The multidimensional poverty assessment toll: Decision, development and application of a new framework for measuring rural poverty. Rome: International Fund for Agricultural Development.

Colosi, R. (2005). Negatively worded questions cause respondent confusion. Retrieved January 14, 2013, from http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000508.pdf

Council of Europe (2004). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg: Language Policy Division, Council of Europe.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cureton, E. E. (1957). The Upper and Lower Twenty-seven Per Cent Rule. *Psychometrika* 22,293-296.

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, 44(11), S50-S59.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement,* 58 (3), 357-381.

Forer, B. & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research*, 103(2), 231-265.

Ghiselli, E. E., Campbell, J. P. & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Jossey Bass.

Hollis, L. & Krishnan, V. (2012). Two machine learning methods for identifying vulnerable communities from early childhood development outcomes. Early Child Development Mapping Project (ECMap) Alberta, Community-University Partnership, Faculty of Extension, University of Alberta.

Hymel, S., Le Mare, L., & McKee, W. (2011). The Early Development Instrument (EDI): An examination of convergent and discriminant validity. *Social Indicators Research*, 103(2), 267-282.

Janus, M., Brinkman, S. A., & Duku, E. K. (2011). Validity and psychometric properties of the Early Development Instrument in Canada, Australia, United States, and Jamaica. *Social Indicators Research*, 103(2), 283-297.

Hambleton, R. K. & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Kaftandjieva, F. (2004). *Standard setting.* Section B of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg: Language Policy Division, Council of Europe.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, California: Sage Publications Inc.

Krishnan, V. (2010). A comparison of Principal Components Analysis and Factor Analysis for uncovering the Early Development Instrument (EDI) domains, Early Child Development Mapping (ECMap) Project Alberta, Community-University Partnership (CUP), Faculty of Extension, University of Alberta, Edmonton, Alberta.

Krishnan, V. (2011). Teachers' assessment of preschoolers' social and emotional competence: Does sex of children matter in developmental outcomes? Paper presented at the World Academy of Science, Engineering, and Technology (WASET), Paris, July 27.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Macdonald, P. & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921-943.

Nelson, L.R. (2008). Rasching an achievement test. *Thai Journal of Research Methodology and Cognitive Science,* 6(2), 20-40.

Nunnally, J. C. (1978). *Psychometric theory* (2nd Ed.). New York: McGraw-Hill.

Pallant, J. (2007). *SPSS Survival Manual.* New York: McGraw-Hill Education.

Pope, G. (2009). Item analysis analytics part 1: What is Classical Test Theory? Retrieved October 5, 2012, from http://blog.questionmark.com/item-analysis-analytics-part-1-what-is-classical-test-theory.

Sohn, Y. (2009). A comparison of methods for item analysis and DIF using classical test theory, item response theory, and generalized linear model. Unpublished Master of Arts Thesis, University of Georgia, Athens, Georgia.

Stage, C. (1998). A comparison between item analysis based on item response theory and classical test theory: A study of the SweSAT Subtest READ. *Educational Measurement* No 30. Umeå, Sweden: University of Umeå, Department of Educational Measurement. Retrieved October 5, 2012, from www.umu.se/edmeas/publikationer/index_eng.html.

Stage, C. (2003). Classical test theory or item response theory: The Swedish experience. *Educational Measurement* No 42. Umeå, Sweden: University of Umeå, Department of Educational Measurement.

Thorndike, R.L. (1982). Educational measurement: Theory and practice. In D. Spearitt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theory* (pp. 3-13). Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations.

Varma, S. (2008). Preliminary item statistics using point-biserial correlation and *p*-values. Retrieved October 7, 2012, from http://www.eddata.com/resources/publications/EDS_point_Biserial.pdf.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26,* 191-208.

Acknowledgements

The author is responsible for the choice of the method and the presentation of the results contained in this report and for the opinions expressed therein, which are not necessarily those of the University of Alberta or Alberta Education and does not commit these organizations. The author acknowledges the contributions of Dr. Susan Lynch, Director of the ECMap Alberta Project in providing all that is needed to insure completeness and accuracy in the report. The support of the entire ECMap team members is also appreciated.