Test Reliability and Stability of Children's Cognitive Functioning

Fiona. H. Spencer, PhD
Queensland University of Technology

Laurel. J. Bornholt, PhD
University of Sydney

Robert. A. Ouvrier, MD
The Children's Hospital at Westmead

Spencer, FH., Bornholt, L. & Ouvrier, RA. (2003). The test-reliability and stability of children's cognitive functioning. *Journal of Child Neurology. 18* (1). 5 - 11.

Publishers/Journal Web Page
http://www.bcdecker.com/default.aspx,
http://www.bcdecker.com/productDetails.aspx?BJID=69

Keywords
SYSTEMS screening test, test reliability, stability, children's cognitive functioning

ABSTRACT

This study addresses the test reliability of a screening test and stability of children's cognitive functioning. Children aged 5 to 8 years in western Sydney were assessed on three occasions. The first assessment provided a baseline, with the second assessment at 2-, 4-, or 12-week intervals. The final assessment was 4 weeks later. Indicators of reliability and stability suggested that a distinction can be made between test reliability and the phenomenon (cognitive functioning) stability. Cognitive functioning was assessed using the School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS). The findings have implications for indicators of reliability and stability of cognitive assessments in developmental research and clinical practice.

The major aim of this study was to investigate the notions of test reliability and phenomenon stability, applied to a new screening test of children's cognitive functioning [1]. A distinction between test reliability and phenomenon stability has been suggested by Baltes et al,[2] Bjorklund,[3] Nesselroade,[4] Nesselroade et al, [5] Wu et al[6] and researchers have provided measurable indicators for the distinction, such as Heise, [7,8] Nesselroade, [4] Valera,[9] Wiley and Wiley, [10] and Wu et al.[6]. This study used the single indicators of reliability and stability developed by Heise (1969) applied to a single test of cognitive functioning [7].

The Reliability of a Test

Reliability is defined by researchers and statisticians as an indicator that provides information about the uniformity of a test when repeated measures are conducted. Baltes et al defined reliability as the consistency of measurements that purport to measure the same thing. [2] Crocker and Algina explained that reliability determines the reproducibility of a test when scores remain consistent over time for the same test forms or alternate forms.[11] Reliability may be in the form of test-retest, split half or alternative form instrument [12]. Test-retest reliability is the most obvious form of investigating changes over time [12]. It involves the use of the same test repeated over time and is defined as the extent to which test material can be relied on to measure a characteristic consistently over time with the same test material [12]. The Standards for Educational and Psychological Testing state that test-retest reliability is 'a reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores'.[13]

The Stability of a Test

Stability is an aspect of reliability and many researchers report that a highly reliable test indicates that the test is stable over time. The definition of stability given by the Standards for Educational and Psychological Testing is 'the extent to which scores on a test are essentially invariant over time. Stability is an aspect of reliability and is assessed by correlating the test scores of a group of individuals with scores on the same test, or an equated test, taken by the same group at a later time'.[13] This definition clearly focuses on the measurement instrument and the obtained test scores in terms of test-retest stability. Many researchers report test-retest stability, for instance, Berry et al correlated an initial test with a post-test and reported stability determined for a self-efficacy questionnaire, [14] of note is that they did not report on the stability of self-efficacy as such.

The Stability of the Phenomenon Being Measured by a Test

Although stability is considered to be an aspect of reliability it may alternatively be regarded as distinct from reliability. [2] Baltes et al maintained that in developmental research a distinction should always be made between the notions of reliability and stability. They argued that reliability should refer to the repeatability of a test, whereas, stability should refer to the repeatability of what is being measured, that is, the phenomenon.[2] In arguing that stability is a quality of a person being firmly similar over time Gergen (1982) supports the distinction claim. [15] Nesselroade et al also define reliability and stability as distinct concepts. [5] They argued that the psychometric properties of a test should be viewed as distinct from the psychological processes that are measured by the test and that they should be viewed in terms of reliability and stability respectively. [5]

In a recent paper, Wu et al clarified the theoretical position and the practice of distinguishing between reliability and stability. They state "The important distinction between reliability of the measurement instrument … and stability of the phenomenon being measured over time … is, unfortunately, obscured by the practice of using test-retest correlations to estimate reliability" [6]. Although, it is clear form the definitions of reliability and stability that a test-retest design is required to determine the test reliability and stability of psychological processes. [2] Most test-retest correlations use of the Pearson product moment correlation coefficient, the Spearman rho coefficient was more appropriate for the current study. In a review of literature, Ottenbacher reported that in 30 studies the most commonly reported reliability statistic was the Pearson product moment correlation. [16] However, Bjorklund argued that in the investigation of children's intelligence the test-retest correlation should be referred to as "the extent to which different children maintain their rank order over time in comparison with their peers".[3] The Spearman correlation coefficient involves the values of each of the variables being ranked from smallest to largest, and the Pearson correlation coefficient is then computed on the ranks.[17]

Indicators of Reliability and Stability

It is argued that to show the distinction between the notions of reliability and stability, there should also be a distinction between measurement methods for the two notions of reliability and stability. Nesseloade et al, in recommending that reliability and stability be measured with two distinct methods, used separate measures of test reliability and construct stability. [5] They contended that the reliability of an instrument should be measured by the correlation between two similar test forms, and that stability is shown with the test-retest correlation between two same test forms measuring the phenomenon of interest (in their study example, anxiety was measured). [5] Wu et al and Nesselroade conducted structural modeling to determine different measures of reliability and stability with various test forms.[4,6] Tisak and Tisak suggest that, in longitudinal designs utilizing multiple measurement forms, 'normal stability' can only be labeled as so with the use of latent curve models (LCM) and latent state-trait models (LST) integrated into one LC-LSTM model.[18] They argue that stability is defined "from the decomposition of latent trajectories into sets of basis curves and individual saliences ".[18] and the maintenance of rank-order over time.[18] They suggest three measures coefficients for reliability – systematic, static and dynamic - for a test at each measurement interval. Although this method may be appropriate, in many clinical situations only one test form is utilized and available for determining reliability estimates.[18]

An alternate approach to utilizing multiple indicator models is to utilize only one test form. Heise, in arguing for a separation between reliability and stability, used path analysis to develop two separate measurement equations based on the test-retest correlations of one test form. [7] These equations included an indicator for test reliability and a indicators for stability. He asserted that test-retest correlations could be used in equations to produce more accurate measures, accounting for measurement error.

Heise's indicator for reliability was recommended as a more accurate estimation of reliability than general test-retest correlation coefficients [7]. The equation makes use of computed test-retest coefficients from a three-wave design with Time A to Time B (r_{AB}), Time B to Time C (r_{BC}) and Time A to Time C (r_{AB}) as shown in formula 1, which is "free from temporal change effects".[7]

$$r_{xx} = r_{AB}r_{BC}/r_{AC} \quad (1)$$

Stability based on suggestions by Heise (1969) involves the calculation of three indicators.[7] He determined these indicators through path analysis and recommended that they were more appropriate in determining stability results than test-retest correlation coefficients. As in the reliability equation indicated previously, the three stability equations make use of computed test-retest coefficients from a three-wave design with Time A to Time B (r_{AB}), Time B to Time C (r_{BC}) and Time A to Time C (r_{AC}). Formulas 2, 3 and 4 account for measurement error and prevent underestimation of stability.[7]

$$s_{AB} = r_{AC}/r_{BC} \quad (2)$$
$$s_{BC} = r_{AC}/r_{AB} \quad (3)$$
$$s_{AC} = r_{AC}^2/r_{AB}\, r_{BC} \quad (4)$$

Heise argued that the subsequently computed reliability was a true measure of reliability and that the computed stability was a measure of the amount of change occurring during intervals. [7] Heise's methods are not well known, particularly in psychological research and in test development.[9]

In contrast to Heise's model,[7] Wiley and Wiley completed path analysis computations and developed three separate measures that could be used to determine a test's reliability at Time A, Time B and Time C.[10] In an application of the two models, Wiley and Wiley argued that their model provided higher stability results than Heise's model.[7,10] A closer examination of the reported results showed that these differences were very small. For example, reliability calculated with example data using Heise's model [7] was .878 and with Wiley and Wiley's model [10] the same example data was Time A = .862, Time B = .878 and Time C = .899. Wiley and Wiley contended that their model provided a more thorough investigation of reliability because it showed a regular increase with time.[10] With Heise's model the stability for the example data for Time A & Time B was .941, Time B & Time C = .941 and Time A & Time C = .886. [7] Wiley's model showed stability results for the same example data of .950, .930 and .884 respectively.[10] Wiley and Wiley argued that Heise's model [7] underestimated stability results, and therefore their proposed model was more appropriate.[10] The stability results for the two models were actually very similar. The differences between the results of the two models ranged from .002 to .011 only.

In a demonstration of the separation of reliability and stability indicators, Valera [9] illustrated Heise's [7] model and the model proposed by Wiley and Wiley[10]. Valera found similar reliability coefficients between the two models.[9] The model assumptions were reported to be the major difference between the models.[9] Particularly, Heise's model assumed that reliability across time remains constant,[7] whereas Wiley and Wiley's model [10] assumed the opposite. Heise [8] contended that the Wiley and Wiley model [10] was appropriate and preferable for longitudinal data analysis only. He also affirmed that "the Wileys' argument is irrelevant for the situations that social scientists encounter".[8]

The Heise model [7] has been utilized in the current study. It uses one test form, provides two distinct measurement methods that account for error from a short-term three-wave test-retest design, which would enable establishment of a test's reliability with one equation, and three equations of stability to assist with the investigation of change over time.

Cognitive Functioning

In the current study cognitive functioning is the phenomenon being investigated. Cognition involves activity of the brain in the form of thinking [19] or, in other words, processing information. Thinking is involved in many aspects of cognitive requirements used in everyday life [20] such as calculation, language, reading and writing. Since the mental activity of cognitive functioning includes thinking about events, objects and concepts,[3] cognitive functioning is considered to be the process of mental activities, which relates to cognition that is in action at the time processing takes place. Bjorklund regarded cognition as referring to the processes by which knowledge is acquired and manipulated and as a reflection of the mind.[21]

Richardson considered that cognitive functioning involves mathematical, spatial and verbal abilities.[22] In the current study, cognitive functioning involves activity of the brain in the form of thinking and is tested through the retrieval and manipulation of information with the School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS).[23]

Summary of Current Study

This study contributes to the literature by providing an applied analysis of the separation between the indicators of reliability and stability for a single test. The main aim of the study is to investigate the distinction between reliability and stability concepts by illustrating that they can be measured separately. The study will investigate the resulting reliability and stability for groups of children based on the testing time interval, gender and age groups. The separate equations, indicating reliability and stability, proposed by Heise,[7] allow a distinction to be made statistically between the two indicators. Reliability is referred to as the consistency of a test over time. [2] Whereas stability is the quality of a person tested (for example in the current study cognitive functioning is measured) being similar over time.[15]

METHOD

Research Design

The investigation focused on the assessment of the SYSTEMS reliability in measuring cognitive functioning and the stability of cognitive functioning over time. The study incorporated the use of a three-wave design (i.e., three measurement occasions). In accordance with this design, Coleman [24] and Heise [7] demonstrated that a three-wave test-retest design is a more appropriate design than a traditional two-wave design.

The retest intervals were established for specific purposes. Assessment at Time A established a baseline indicator. Time B varied the test-retest interval for three groups (B1 = 2 weeks later, B2 = 4 weeks later and B3 = 12 weeks later). Prior to testing, children were allocated to one of the three Time B interval groups. The Time B2 interval of 4 weeks was suggested by Anastasi and Urbina as an appropriate time interval for children.[12] Time B3 of 12 weeks was incorporated to test children's upper time interval limit.

A standard interval of 4 weeks from Time B to Time C was used for the final retest interval for all children. This retest interval of 4 weeks was in accordance with Anastasi & Urbina's argument for retesting periods for children.[12]

Sampling Technique

A sample of children at Time A (N = 399) was a part of a project at the Children's Hospital at Westmead (CHW) entitled "The development of a screening test of mental functioning to be used with children by neurologists" (National Health & Medical Research Council (NHMRC) Project). As an extension of the CHW project the current project utilized a randomly selected group of students (N = 135) from the CHW sample to be involved in follow-up testing at Time B and Time C.

All government primary schools within a 20km radius of the CHW, Sydney, were included in a sample list. A number of these schools were randomly selected for inclusion in the study (N = 50) and sent a letter explaining the research. Of these, six of the schools agreed to participate. The Occupation and Education Index of the Socio-Economic Indicators for Area (SEIFA) was utilized to ensure that children were sampled from a wide range of schools according to the socio-economic areas [25] with the current sample mean within the ABS average range. The response rate was 49%, which is considered satisfactory.

Research Sample

The children (N = 135) were aged between 5 to 8 years (up to 8 months within their age group at Time A). The groups of children were younger (5 and 6 year olds, n = 64) and older (7 and 8 year olds, n = 71). The results of previous studies utilizing the screening test show a ceiling effect at age 10.[26] Therefore, children older than 8 were not included in the current study. An equal number of boys (50%) and girls (50%) were tested to account for any gender differences.

The first two testing sessions (Time A and Time B) included 135 children. Only 134 children were tested at Time C because one child moved away from a school between Time B and C. Regression analysis enabled a SYSTEMS Time C score to be estimated for this child.

The sample of 135 retested children had a 100% power analysis. This depended on a half standard deviation difference between test scores, significant at a level of 0.05.[27] The time interval groups (B1 = 2 weeks, B2 = 4 weeks and B3 = 12 weeks) with sample sizes of 45, 48 and 42 respectively, had a sample power of 87% with a half standard deviation difference between test scores, significant at a level of 0.05 [27].

Procedures

The current study was approved by the appropriate organisations as an extension of a project at the CNW project. Following sample selection and permission procedures, children were tested in a one-to-one interview situation at their school during class time in an office, quiet room or area. The standard testing procedures for interviewing each child on every occasion incorporated evenly paced items, without feedback to the child about correct or incorrect responses. The testing environment included a desk and two chairs, one for the researcher and the other for the child.

Materials

The School-Years Screening Test for the Evaluation of Mental Status is a cognitive screening test for children based on the adult MMSE.[28] Ouvrier et al developed the SYSTEMS as an alternative to the MMSE for use with children.[1]

The SYSTEMS was designed to provide an indicator of a child's cognitive functioning, where cognitive functioning was taken to mean cognitive manipulation as well as general information and skills. [1] The test includes activities incorporated from theoretical and empirical research and clinical experience. [1] Items cover areas of attention, memory, mental manipulation, drawing and language, as well as reading, spelling, arithmetic and autobiographical information. [1] The test was found to be highly correlated with the *Stanford-Binet Intelligence Test*, 4th Edition [29] (r = .88) administered to school children, and related to the *Differential Ability Scales* [30] results (r = .75) administered to neurological clinic patients.

The SYSTEMS is reported to be internally consistent (alpha = 0.92) for five to eleven year olds. [1,23] SYSTEMS is also internally consistent for 5, 6, 7 and 8 year olds (alphas from 0.64). [1] SYSTEMS has been found to be unbiased by gender (f = 0.75, p = ns) and socio-economic indicators for areas [25] (f = 0.16, p = ns). [1] SYSTEMS was found to have high inter-rater reliability (r = .94) and test-retest reliability (r = .94). [1]

For the administration of the screening test (SYSTEMS) on each testing occasion the following instructions were given to the child by the researcher:

"Hi, my name is..... I am here today to ask a few questions and do some other activities. I'd like you to answer as well as you can. If you can't answer any just let me know".

The test was designed to be administered verbally, as set out on the SYSTEMS form, in an interview situation. The child's worksheet was given to the child only when administering the corresponding questions as indicated on the test form (i.e., not at the beginning of the assessment). If a child asked for a question to be repeated the test administrator asked and determined whether the child did not hear the question. If the child had not heard the question it was repeated; otherwise the test administrator asked the child to attempt the answer. If the answer was incorrect it was scored as such. The child was permitted to have only one try at the copying section.[23] Each of the 46 items was scored on a dichotomous scale with "1" indicating a correct response and "0" indicating an incorrect response. As detailed in the testing manual, total scores were summed with 46 as the maximum score. [1,23] The researcher argued that "… SYSTEMS is a valuable clinical tool to assist in the decision about the need for further cognitive assessment".[1]

RESULTS

Reliability and Stability Results

Test-retest correlation coefficients. To determine Heise's [7] indicators of reliability and stability test-retest correlations coefficients first need to be determined. When cognitive functioning scores over time were compared using the Spearman rho ranked correlation, the results were high between testing occasions (Time A to Time B $\underline{rs}_{AB}$ = .94, Time B and Time C $\underline{rs}_{BC}$ = .94 and Time A to Time C $\underline{rs}_{AC}$ = .93) (see Table 1). When correlation coefficients were computed for groups of children based on age, gender and time interval, the three strongest correlations were for children in the 4 week interval group ($\underline{rs}_{BC}$ & $\underline{rs}_{AC}$ = .96) and for children in the 12 week interval group ($\underline{rs}_{BC}$ = .96). Correlations were not as strong ($\underline{rs}_{AC}$ = .81) for older children from Time A to Time C.

Table 1
Cognitive Functioning Correlation Coefficients

Groups of Children	$\underline{r}_{sAB}$	$\underline{r}_{sBC}$	$\underline{r}_{sAC}$
Overall	0.94, n = 135	0.94, n = 135	0.93, n = 135
Younger	0.88, n = 64	0.93, n = 64	0.86, n = 64
Older	0.84, n = 71	0.83, n = 71	0.81, n = 71
Boys	0.94, n = 67	0.95, n = 67	0.93, n = 67
Girls	0.94, n = 68	0.92, n = 68	0.94, n = 68
B1	0.92, n = 45	0.89, n = 45	0.93, n = 45
B2	0.95, n = 48	0.96, n = 48	0.96, n = 48
B3	0.95, n = 42	0.96, n = 42	0.92, n = 42

Cognitive functioning test reliability. Using Heise's [7] reliability measure, the overall test reliability in Figure 1 was found to be strong ($\underline{r}_{xx}$ = .95). When test reliabilities were computed for groups of children based on age, gender and time interval, the strongest was for children in the 12 week interval group ($\underline{r}_{xx}$ = .99) and the weakest test reliability (although still strong) was for older children ($\underline{r}_{xx}$ = .86).

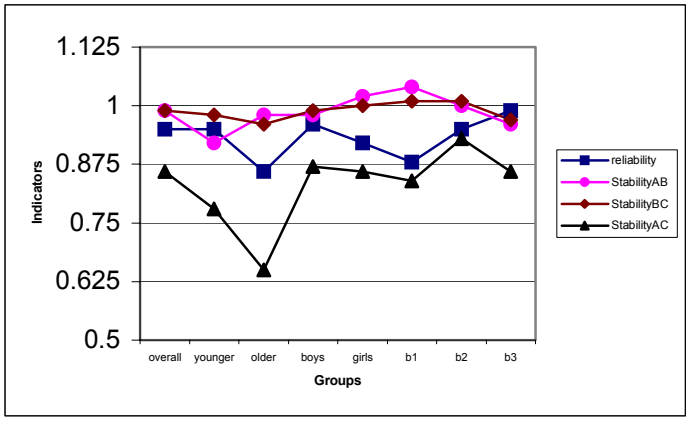

Figure 1
Test Reliability and Cognitive Functioning Stability

Cognitive functioning stability. The stability of cognitive functioning over time, measured with Heise's three stability measures, [1] was found to be strong, as reported in Figure 1. Overall, cognitive functioning did not change from the first testing session to the second (s_{AB} = .99), or from the second to the third (s_{BC} = .99), but more change was noted with less shared variance from the first to the third testing session (s_{AC} = .86). When cognitive functioning stabilities were computed for groups of children based on age, gender and time interval, the strongest was for children in the 2 week interval group (s_{AB} = 1.04) and the weakest stability result (although still strong) was for older children (s_{AC} = .65).

Comparison Results. In comparing the test reliability and cognitive functioning stability results Figure 1 shows that reliability and $stability_{AC}$ show a similar pattern over the groups examined, with reliability slightly higher than $stability_{AC}$. $Stability_{AB}$ and $stability_{BC}$ appear to show a similar pattern, although the $stability_{AB}$ for younger children was slightly less than their at $stability_{BC}$ result.

DISCUSSION

This study examined the distinction between test-reliability and phenomenon (cognitive functioning) stability through determining the SYSTEMS reliability and cognitive functioning stability. The study presented and assessed children's cognitive functioning over time and examined groups of children based on age, gender and test-retest intervals.

The study showed that the SYSTEMS was a reliable test over time and that cognitive functioning in children was stable between each initial and subsequent testing session, such as from Time A to Time B and from Time B to Time C, measured with Heise's model. [7] More change in cognitive functioning was noted over time from the first testing session to the last, that is, from Time A to Time C, indicated by lower stability results. This implies that cognitive functioning did not change over short periods of time, although the more notable changes occurred over the total testing period (a maximum of 4 months for some children).

The research has an impact on understanding the definitions of psychometrics related to retesting concepts such as reliability and stability. The results indicated that older children had the weakest test-reliability and weakest Time A to Time C stability (although still at acceptable levels). This result may be indicating the nature of older children, in that their cognitive functioning is changing more rapidly than younger children. While the cognitive functioning of older children appears to be changing rapidly the test's reliability was still at an acceptable level at .86 for these children. Final conclusions were established about the nature of cognitive functioning in children. Specifically that cognition is less stable and is perhaps developing more rapidly for older children than younger children.

The research made a significant contribution to the study of the applied separation of a test's reliability and phenomenon stability. The two indicators based on Heise's model [7] revealed different results for measures of reliability and stability for a single test, specifically the strongest reliability result was found for children in the 12 week interval group and the strongest stability result was found for children in the two week interval group. Although the reliability and stability$_{\underline{AC}}$ showed a similar pattern of results, as shown in figure 1, reliability results were higher than stability$_{\underline{AC}}$, results for all groups investigated and a large difference was found between the two measures for older children. These distinct results demonstrate that the test reliability and phenomenon stability should be regarded as separate concepts, as previously argued by Baltes et al, [2] Bjorklund,[3] Heise, [7,8] Nesselroade, [4] Nesselraode et al, [5] Valera, [9] Wiley and Wiley,[10] and Wu et al.[6]

The implications of the research extend to areas of psychological, medical and educational testing of children's cognitive functioning. The research has shown the importance of applying and separating a test's reliability and psychological phenomenon stability in a three-wave test-retest study design with the use of a single test. The reliability results have implications for the usefulness of the test.

The investigation of stability enabled conclusions to be drawn about children's cognitive functioning over time. In particular, cognitive functioning was stable over short time intervals. However over the total testing period, from the initial testing to the final testing, cognitive functioning stability tended to decrease, indicating more change. This finding also has implications for the use of the test to indicate general cognitive development as well as recovery (or decline) in cognitive functioning following injury or illness, and to monitor general cognitive functioning in evaluating interventions in clinical settings.

These research findings need to be considered by researchers interested in investigating change over short time intervals, in particular when cognitive assessments are sensitive in terms of reliability and stability. The main point here is that reliability and stability are separate concepts and should be calculated separately.

Future research may need to investigate the SYSTEMS test's reliability and cognitive functioning stability over extended periods of time in longitudinal studies, such as 6 and 12-month intervals. The reliability and stability model of Wiley and Wiley [10] could then be incorporated. Another suggestion would be to model a study on the suggestions of Nesselroade et al with reliability calculated from the correlation between two similar test forms and stability calculated from the correlation between two same test (such as, test-retest) forms. [5] Perhaps Nesselroade's structural model approach could be incorporated, whereby stability and reliability are determined from one structural equation with the use of two or more test forms administered over a number of time intervals. [4] Alternatively, Tisak and Tisak's distinct measures for reliability and stability for use with longitudinal designs utilizing multiple measurement forms could be further analyzed.[18] Finally, it is recommended that in the practice of medical, psychological and educational assessment the terms test-reliability and phenomenon stability need to be considered as separate constructs.

References

1. Ouvrier R, Hendy J, Bornholt L, Black F. The School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS). Journal of Child Neurology 1999;14:772-80.
2. Baltes PB, Reese W, Nesselroade JR. Life-span developmental psychology: Introduction to research methods. Monterey, CA: Brooks/Cole; 1977.
3. Bjorklund DF. Children's thinking: Developmental function and individual differences. 2nd ed. Sydney, NSW: Brooks/Cole; 1995.
4. Nesselroade JR. Adult personality development: Issues in assessing constancy and change. In: Rabin AI, Zucker RA, Emmons RA, Frank S, editors. Studying Persons and Lives. New York: Springer Publishing Company; 1990. p. 41-83.
5. Nesselroade JR, Pruchno R, Jacobs, A. Reliability vs. stability in the measurement of psychological states: An illustration with anxiety measures. Psychologische Beitrage 1986;28(1-2):255-64.
6. Wu X, Hart CH, Draper TW, Olsen JA. Peer and teacher sociometrics for preschool children: Cross-informant concordance, temporal stability, and reliability. Merrill-Palmer Quarterly 2001;47(3):416-43.
7. Heise DR. Separating reliability and stability in test-retest correlation. American Sociological Review 1969;34(1):93-101.
8. Heise DR. Comment on "The estimation of measurement error in panel data". American Sociological Review 1970;35:117.
9. Valera JB. Two path analytic models of test-retest reliability: Examples from social psychology. Philippine Journal of Psychology 1991;24(1):49-58.
10. Wiley DE, Wiley JA. The estimation of measurement error in panel data. American Sociological Review 1970;35:112-17.
11. Crocker LM, Algina J. Introduction to classical and modern test theory. New York: Holt, Reinhart & Winston; 1986.
12. Anastasi A, Urbina S. Psychological Testing. 7th ed. New Jersey: Prentice-Hall International; 1997.
13. American Education Research Association APA, and The National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC:: American Educational Research Association; 1999.
14. Berry JM, West RL, Dennehey DM. Reliability and validity of memory self-efficacy questionnaire. Developmental Psychology 1989;25(5):701-13.
15. Gergen KJ. From self to science: What is there to know? In: Suls J, editor. Psychological Perspectives on the Self. London: Lawrence Erlbaum.; 1982. p. 129-49.
16. Ottenbacher KJ. An examination of reliability in developmental research. Journal of Developmental and Behavioral Pediatrics 1995;16(3):177-82.
17. Statistical Package for the Social Sciences. Statistical Package for the Social Sciences (SPSS). Version 10.0.5 ed. Chicago, IL: SPSS Inc.; 1999.
18. Tisak J, Tisak MS. Permanency and ephemerality of psychological measures with application to organizational commitment. Psychological Methods 2000;5(2):175 - 98.
19. Ellis HC, Hunt RR. Fundamentals of human memory and cognition. 4th Ed ed. Dubuque, Iowa: WC Brown.; 1989.
20. Strub RL, Black FW. The mental status examination in neurology. 3rd ed. Philadelphia, PA: F.A. Davis; 1993.
21. Bjorklund DF. Children's thinking: Developmental function and individual differences. 3 ed. Sydney, NSW: Brooks/Cole; 2000.
22. Richardson JTE. Introduction to the study of gender differences in cognition. In: P.J. Caplan. M. Crawford. J.S. Hyde, Richardson JTE, editors. Gender difference in human cognition. New York: Oxford University Press; 1997. p. 3-29.
23. Ouvrier R, Hendy J, Bornholt L, Black F. The Administration and Scoring Manual for the School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS). Westmead, NSW: The New Children's Hospital; 2000.
24. Coleman JS. The mathematical study of change. In: H.M. Blalock (Jr), Blalock AB, editors. Methodology in social research. New York: McGraw-Hill.; 1968. p. 428 - 78.

25. Australian Bureau of Statistics. Socio-economic indices for areas SEIFA. Canberra, Australian Capital Territory: Australian Government Printing Service.; 1990.
26. Ouvrier RA, Goldsmith RF, Ouvrier S, Williams DC. The value of the Mini Mental State Examination in childhood. A preliminary study. Journal of Child Neurology 1993;8:145-48.
27. Bach LA, Sharpe K. Sample size for clinical and biological research. Australian and New Zealand Journal of Medicine 1989;19:64-68.
28. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:189-98.
29. Thorndike RL, Hagen EP, Sattler JM. The Stanford Binet Intelligence Test. 4th ed. Marrickville, NSW: Harcourt Assessment Company & the Psychological Corporation; 1986.
30. Elliot CD. The Differential Ability Scales. Marrickville, NSW: Harcourt Assessment Company and the Psychological Corporation; 1990.

Acknowledgments:
We acknowledge the encouraging support given by the Children's Hospital at Westmead and are grateful to the insightful children, the parents, teachers and principals involved with the project. This research was part of a Doctoral thesis completed by Dr Fiona Spencer (Nee: Black) and supervised by Dr Laurel Bornholt and Clinical Professor Robert Ouvrier. Funding for the PhD was through a Public Health Postgraduate Research Scholarship from the National Health and Medical Research Council, Australia. This project was approved by the Children's Hospital at Westmead Ethics Committee, the University of Sydney Ethics Committee and the NSW Department of Education and Training.