To test or not to test? The selection and analysis of an instrument to assess literacy skills of Indigenous children: a pilot study.

by

John R. Godfrey, Gary Partington and Anna Sinclair

Edith Cowan University and the Education Department of Western Australia, Perth.

ABSTRACT:

This paper explains the process of selecting a standardised reading skills instrument to be used with Indigenous children in various settings in Western Australia. The selection process included the examination of a number of instruments, and consultation with educators and researchers. The instrument chosen contained items that appeared to form a basis to assess the literacy skills of Indigenous children. The test was trialed with a small sample of Indigenous children in two schools. The pilot study results were analysed and the results discussed. Implications for the evaluation of Indigenous children and educational programs are drawn.

INTRODUCTION:

The Conductive Hearing Loss Research Team consisting of researchers from Kurongkurl Katitjin, School of Indigenous Australian Studies at Edith Cowan University, Education Department of Western Australia, Catholic Education Office, Association of Independent Schools and Derbarl Yerrigan Health Service are investigating the effect of conductive hearing loss as a consequence of Otitis Media on the language development, including communication and literacy skills, of Indigenous children.

The team believes that hearing loss due to Otitis Media may affect the development of auditory discrimination and processing skills and as a consequence may reduce phonological awareness, short-term auditory memory skills, auditory sequential memory skills and thus numeracy and literacy skills. They are seeking answers to, among others, the following questions:

1. What is the relationship between conductive hearing loss and school related variables including: literacy; numeracy; attendance; behaviour of Pre-primary to Year 3 students?

5. To what extent does the implementation of new teaching strategies result in improved literacy, numeracy, reduced absenteeism and reduced behaviour problems?

The difficulty of chosing a reading test to ascertain the reading ability of Indigenous children who may have suffered Conductive Hearing Loss (CHL) proved to be a most difficult exercise. The following instruments were examined to determine their suitability; the Kimberley Standard English Vocabulary Test (Brandenburg, c.1984), the Phonological Profile for the Hearing Impaired Test (Vardy, 1991); the Western Australian Action Picture Test (Kormendy, 1988); and the The Hundred Pictures Naming Test (Fisher & Glenister, 1992). All were rejected for a multiplicity of reasons, including cultural and contextual

inappropriateness, unsuitability of language, complexity of administration, length, difficulty for assessing K to Year 3 reading skills and/or because they were considered outdated.

After careful consideration and close examination, the reading tests contained within Neil J. Waddington's (2000) Diagnostic Reading and Spelling Tests 1 & 2 (Second Edition) were chosen because these tests appeared to be uncomplicated t and the language appeared to be the most appropriate for Indigenous children in K through to Year 3. The items depicted relevant and current items to be recognised such as balls, horses, fish and the sun etc. The tests are easy to score. The use of pictures with a three option multiple choice item narrowed choices and aided statistical analysis. The correct answer was given as one of the multiple choice responses.

The test was examined by three researchers, who all agreed that the face validity of the instrument appeared suitable for assessing the reading ability in English of Indigenous children.

The Waddington (2000) reading tests were produced in parallel forms. Thus the children could be tested before and after the administration of intervention programs with tests that were constructed as closely as possible in format, question type, difficulty, discrimination and therefore reliability. Finally, Waddington's (2000) Diagnostic Reading and Spelling Tests 1 & 2 (Second Edition) booklet contains statistical data on the validity and reliability of the tests. The Kuder-Richardson 20 reliability index is reported to be 0.98 for reading test 1 and 0.97 for reading test 2. The Standard Error of Measurement (SEM) is also calculated. It is reported as ranging from plus or minus 2 months in reading age for the two parallel forms of the reading test. These statistics indicate that the tests are highly reliable for determining the reading age of young children.

The statistical data also contained graphs that indicated the trends over two decades of sampling, 1988 to 1999. The graph indicates that the results of the average chronological age are comparable across the decade. Included also were graphs of the comparisons between the sexes. The data indicates that the girls outperform the boys by 2 months on average though by age 11 the boys outperformed the girls by 3 to 4 months.

Moreover Waddington's (2000) Reading Tests 1 & 2 contained data and a graph of a sample of 204 Indigenous children (2.7% of the 7611 children tested in 1999). Waddington (2000, p. 83) claimed that: "on average, this group were 7.8 months behind the average for their age group in reading . . ." A comparison was made between the results for Indigenous children from 1988 and 1999. In 1988 the Indigenous sample (2.4% of the 2575 students tested in 1988) was on average 19.4 months behind non-Indigenous children in reading. Waddington (2000, p. 83) claims that: "the 1999 results indicate a pleasing 250% increase in the literacy levels of indigenous Australians over the 11 year period."

Waddington also compared students from Non-English speaking backgrounds (NESB):

> Out of the 7611 students in the 1999 sample, 656 (8.6%) were identified by their teachers as being from non-English speaking backgrounds. On average, NESB students performed 0.3 of a month above the average for their age for reading … It appears that this group is making very significant literacy advances in spite of their respective backgrounds (Waddington, 2000, p. 83).

Unfortunately Waddington's analyses of these two sub-groups leave a number of crucial issues unanswered. For example he does not disclose the full details of either the NESB or Indigenous group. The sample of NESB students may have included some Indigenous students. Also the samples of both the NESB and Indigenous students is small and thus the

reported trends are open to question. However, the trends are positive rather than negative and therefore "pleasing" (Waddington, 2000, p. 83).

RESULTS OF PILOT STUDY:

Two schools were chosen for the Pilot Study, one a remote Independent Aboriginal school in the Fitzroy valley of the Kimberley region and the other a rural Aboriginal school in the Goldfields region of Western Australia. The chronological age of the children from the Kimberley school ranged from 5 years 6 months to 11 years 3 months and they were familiar with three languages types. The chronological age of the children from the Goldfields school ranged from 5 years 11 months to 9 years 10 months, most spoke English as their first language. All were considered by their teachers to be at a reading age of approximately 6 years. Most had a history of suffering from Conductive Hearing Loss in infancy and at some time during their schooling.

The total sample consisted of 15 children, 9 from the Goldfields school and the other 6 from the Kimberley school.

The test was administered on both occasions by the same researcher in the same room as the other children and on one occasion with the teacher present. Most of the children were tested with the first 24 items that contained pictures and required a multiple choice response. The results and analysis were calculated with the aid of the EdStats computer program (Knibb, 1995).

The average of the total scores was 11.5 and the standard deviation of 4.7. The Cronbach Alpha reliability coefficient was calculated as 0.84 while the Pearsons' correlation between the two halves of an odd-even items split produced a co-efficient r of 0.93 and after the Spearman-Brown correction was applied a Split-half reliability coefficient of 0.96. The SEM of the total scores was 0.92 which would produce a variation in reading age of approximately plus or minus one month. These results are consistent with those reported by Waddington (2000). He calculated using the Kuder -Richardson 20 (KR20) technique that Reading Test 1 has a reliability coefficient of 0.98 and a SEM of plus or minus 2 months.

Table 1. Test Statistics

Cronbach Alpha 0.84

Pearsons r 0.93

Split-half Reliability 0.96

Totals Mean 11.5

Totals Standard Deviation 4.7

Standard Error of Measurement 0.92

---

a. Norm Referenced Test:

The results were analysed as Norm Referenced Test (NRT) data with the assistance of the EdStats programme (Knibb, 1995). The data produced the following Discrimination Indices (DI), Difficulty Indices (Diff) and Item Contribution Indices (ICI) (see Table 2).

Table 2: Norm Referenced Analysis Results

Item	DI	Diff	ICI
1	0.45	0.87	18
2	0.32	0.87	13
3	0.77	0.47	54
4	0.39	0.67	39
5	0.41	0.87	16
6	0.34	0.73	27
7	0.09	0.93	2
8	0.41	0.60	37
9	-0.12	0.67	-12
10	0.45	0.87	24
11	0.48	0.60	43
12	0.17	0.40	10
13	0.10	0.13	2
14	0.53	0.40	34
15	0.56	0.47	39
16	0.19	0.13	4
17	0.47	0.33	25

18 0.55 0.20 17

19 0.87 0.47 68

20 0.39 0.20 13

21 0.32 0.13 7

22 0.04 0.20 1

23 0.67 0.20 21

24 0.27 0.07 3

Mean: 0.38 0.48 21

The DI's in Table 2 indicate that the correlation between the scores on the item and the total scores is positive for all items except item 9. The DI's for items 7 and 22 are low.

The ICI is an indication of the contribution of the item to the test as a whole with regard to reliability of the instrument. The difficulty and discrimination of the item are used to determine the ICI value. "Items with ICI's less than 0 should be considered for modification or removal. Items with ICI's more than 20 are desirable" (Knibb, 1995). The ICI for item 9 is negative while for items 7, 13, 16, 22 and 24 it is low.

b. Criterion Referenced Analysis:

The results were further analysed as Criterion Referenced Test (CRT) data with the assistance of the EdStats programme. The data produced the DI and Diff results as listed in Table 3. The Mastery level was set at 50% level of mastery as an artibritary level to enable an analysis of the suitability of the items. The analysis indicated that the items, as a mastery test, were operating satisfactory with an average discrimination at 0.34. However, the discrimination indices for items 9, 12 and 13 were a cause of concern. These three items would need to be revised to increase the reliability of the instrument. Item 9 involves recognition of the first letter of the word that agrees with a picture of a bird. While items 12 and 13 require recognition of the word for 'pig' and 'flag' respectively.

Notwithstanding the DI's of these three items the Waddington (2000) Reading Test 1 appears on these results to be a discriminating, reliable instrument for assessing the mastery of the English reading skills and sub-skills.

Table 3: Criterion Referenced Analysis

Item	DI	Diff
1	0.25	0.87
2	0.25	0.87
3	0.73	0.47
4	0.63	0.67
5	0.25	0.87
6	0.23	0.73
7	0.13	0.93
8	0.48	0.60
9	-0.18	0.67
10	0.25	0.87
11	0.48	0.60
12	0.05	0.40
13	0.02	0.13
14	0.59	0.40
15	0.73	0.47
16	0.29	0.13
17	0.45	0.33
18	0.43	0.20
19	1.00	0.47
20	0.16	0.20
21	0.29	0.13
22	0.16	0.20
23	0.43	0.20
24	0.14	0.07

Mean: 0.34 0.48

---

c. Rasch Model Analysis:

The Rasch measurement model (Rasch, 1980) is ideally suited to measure concepts such as reading skills (Andrich & Godfrey, 1978-9). The EdStats computer programme was used to check that the responses from this instrument fit the Rasch measurement model according to the criteria described by Wright and Masters (1982) and Wright (1985). It calculates the student skill on the scale that is required for the student to have a 50 per cent chance of gaining a correct response to an item. These skills/behaviours are calculated in log odds (logits) on a scale ordered to represent the increasing skill/behaviour needed to answer each category. Skill/behaviour items for which the students do not use the categories consistently are not considered to fit the model and are discarded. This analysis using the EdStats program was used as a preliminary check on the items to ensure the instrument measures a uni-dimensional trait.

The EdStats computer program used to analyse this data performs:

> . . . Rasch analysis using Andrich's 1978 (Andrich, 1978a; 1978b) rating scale model. Values are estimated using the UCON algorithm (Wright & Masters, 1982) . . . . This item fit is the standardised t Fit statistic recommended by Wright and Masters (1982) . . . The pattern of results for items values greater than 2 or less than -2 is not consistent with the item responses fitting the Rasch model. These items should be modified or excluded from the measurement model (Knibb, 1996, pp. 49-51).

The t Fit values established by Wright and Masters (1982, pp. 99-102) of a range of plus 2 or minus 2 as a check on item fit to the model is used as a guide in this analysis.

Table 4: Rasch Model results:

Item Diff Item Fit t

---

1 -2.352 -0.469

2 -2.352 0.122

3 -0.211 -1.173

4 -0.962 0.466

5 -2.352 -0.022

6 -3.173 0.427

7 -3.173 0.427

8 -0.539 0.176

9 -0.926 2.451

10 -2.352 -0.469

11 -0.539 0.253

12 0.589 1.474

13 2.417 0.324

14 0.589 -0343

15 0.211 -0.460

16 2.417 0.518

17 0.979 0.114

18 1.885 -0.896

19 0.211 -2.513

20 1.855 0.134

21 2.417 0.046

22 1.855 0.896

23 -0.163 -0.339

24 3.242 0.279

Only two items , items 9 and 19 of the twenty four items used in the analysis did not fit the pattern of results for items values greater than 2.0 or less than 2.0 which is consistent with the item responses fitting the Rasch model.

DISCUSSION:

On the basis of the administration of the Waddington's (2000) Reading Test 1 and the statistical analysis of the results of a small sample of Indigenous children it would appear that the Waddington (2000) test is suitable as an indicator of the reading skills of Indigenous children. The DI, Diff and ICI statistics indicate that most items are functioning at a

reasonable level in regard to difficulty and discrimination. The Cronbach Alpha and Split-Half reliability coefficients are high to very high indicating that the test is a highly consistent measure of reading ability based on this small sample. The Rasch Model analysis indicates that a uni-dimensional latent trait of reading skills is assessed by the instrument.

Notwithstanding the above, the item analysis indicates that a number of items need to be revised. In particular Item 9 possibly needs to be removed from the instrument while items 13, 16, 21, 22 and 24 require close examination to determine how they can be improved. A larger sample may indicate that these items, while discriminating at a minimal level, need no revision.

IMPLICATIONS:

Unfortunately the administration of the Waddington (2000) tests to Indigenous children produces a wide divergence of opinion. These differences of opinion may be based on the location of various schools. For example at a meeting in a remote school district those responsible for the educational welfare of Indigenous children in the district were clearly opposed to the test being administered to Indigenous children. These strong opinions were due to perceptions that the test contained numerous inappropriate, culturally biased items. A researcher received the following reception when the Waddington tests were introduced into a discussion of the assessment of the reading skills of Indigenous children.

> One of the . . . Education Office staff asked if Waddington was the proposed reading instrument. When I agreed, it was like opening the floodgates of condemnation. I felt as if I had been ambushed. The . . . staff collectively rounded on me and enumerated the sins of Waddington: it was culturally inappropriate; it didn't provide diagnostics; it would never be used in this district! (G. Partington, personal communication, July 26, 2001).

On the other hand a few days later the same researcher in Perth received the following reception in a metropolitan school:

> The principal was unequivocal in her support for Waddington. Coming after the rejection at the previous meeting, this was a surprise. She stated that the school had results compiled centrally for all students in the school and administered the test as a matter of course. There was no consideration that it might be inappropriate. On the contrary, they regarded it as an important instrument for the assessment of students (G. Partington, personal communication, July 26, 2001).

Indigenous community leaders are concerned that their children are frequently subjected to numerous assessments. Unfortunately the problem will not dissipate; it is a feature of modern society to assess most areas of behaviour and achievement. If the assessment and the associated instruments are culturally appropriate, valid and reliable then Indigenous communities and parents should welcome such evaluation programs. Indeed, carefully constructed evaluation programs have been used to support programs that have aided the education of Indigenous children (Cataldi & Partington, 1998). However Drew (2000) claims that:

> To remove all cultural variables would have the effect of lowering the validity of the test with respect to the domain it purports to measure. The inference from this is that the test would thereby fail to detect problems requiring amelioration. This is complex argument. On the one hand, people from different cultural backgrounds (including Aboriginal and Torres Strait

> Islanders) are expected to perform within the context of the dominant cultural groups. If tests are able to detect 'deficits' in performance then , on the face of it, their use would be advantageous. On the other hand, if the deficits are systematically linked to cultural variables, they may serve to perpetuate myths and stereotypes, which in turn may lead to increased marginalisation, discrimination and exclusion (p. 326).

To succeed in 21st century Australian Indigenous children need to be participants in these educational evaluation processes.

Indigenous people have the right and indeed the responsibility to complain and seek to redress unfair, unreliable and invalid assessments made of Indigenous children. Children in educational settings will continue to be assessed on a range of variables and with numerous techniques. For example, teachers are continually making judgements that affect the educational welfare of children, both Indigenous and non-Indigenous, under their care. The complaints of Indigenous people should not be directed in the first instance at the use of standardised instruments but most bitterly at the unfair assessments made of Indigenous children by school personnel without the aid of reliable and valid instruments (see Godfrey, Partington, Richer, & Harslett, 2001; Godfrey, Partington, Harslett, & Richer, 2001, in press).

In spite of the lack of crucial information regarding of both the Indigenous and NESB sub-groups Waddington (2000) has indicated some steps to emulate by producing a comparison of the results of reading test surveys for 1988 and 1998. He has revealed that by comparing the results of the two sub-groups of Indigenous children that the reading age of Indigenous on the Waddington (2000) reading test 1 increased by 250% over the decade. Evaluators should follow the example of Waddington in this regard by adhering to procedures such as: ensuring that a Pilot study of the any instrument is conducted before using it on the wider Indigenous community; comparing the results collected over time from valid and reliable instruments to ensure the long term reliability of the results; and using the results of valid and reliable instruments to compare various groups within Australian society in order to assist educationally and socially those sectors that are disadvantaged. Merely measuring ourselves by ourselves, and comparing ourselves by ourselves is not wise.

CONCLUSION:

The National Strategy for the Education of Aboriginal and Torres strait Islander Peoples; 1996- 2002 (Ministerial Council on Education, Employment, Training and Youth Affairs, 1995) realises the importance of assessment to Indigenous education programs. It lists as one the strategies for both Early Childhood Education and Schooling; "Formalise assessment procedures, strategies and instruments which appropriately reveal Aboriginal and Torres Strait Islander children's achievement" (Strategies 5.2.6.e & 5.2.6.s).

Cataldi and Partington (1998) reinforce this recommendation by describing a case study of testing of student literacy and numeracy at Lajamanu school which is situated 600 kilometres south west of Katherine. They explain the need for the assessment program at Lajamanu:

> the lack of restraint, understanding and helpfulness on the part of some non-Indigenous teachers was a problem. The complaints, many and varied, from teachers were welcomed in the regional office and led to criticism of the [Warlpiri language bilingual] program from department officers. . . . From its beginning in 1982 the program received strong opposition, particularly from many in the Education Department who tried to stop it.

> It was absolutely necessary to determine whether or not the Warlpiri program at Lajamanu School was producing results, and to demonstrate what those results were, beyond even reasonable doubt (pp. 324-5).

Cataldi and Partington (1998) acknowledged the difficulty of such a course of action but they claim:

> . . . it is essential for a body of students with a non-standard experience of education to have accurate and informative records of their achievement and progress. . . . The success at Lajamanu showed that, given the right processes and content, Indigenous students can succeed. This may have been unpalatable to others in the system. . . . In a way the teachers at Lajamanu were right about the relationship between testing and survival. The only tangible public record of the Lajamanu School Bilingual Program will be this account of the testing program conducted with the children (pp. 329-331).

In short, it is essential to work through the problems associated with various types of tests and assessment programs in general to ensure that accurate and valid instruments and assessment programs are established and maintained to allow Indigenous and non-Indigenous educators to be well informed of the achievements of Indigenous children and Indigenous educational programs "beyond even reasonable doubt" (Cataldi & Partington, 1998, p. 325).

Notwithstanding the above, it is essential to heed the concerns of Drew (2000, pp. 326- 332) and Partington and McCudden (1992, pp. 255- 272). The latter conclude that: "even if it is claimed that tests are culture free, teachers and test administrators should view them with scepticism, for there is no such thing as a culture free test" (Partington & McCudden, 1992, p. 272). Notwithstanding Drew (2000, p. 326) claims that: "it has been shown that even when test items, purported to be biased, are removed, overall scores for individuals from different cultural groups did not differ appreciably."

One simple procedure to check the unsuitability of instruments is to administer and analyse a pilot study. This analysis of the results of the small sample of Indigenous students who were tested with the Waddington (2000) Reading Test 1 may assist with the ongoing debate regarding Waddington tests in particular and testing Indigenous children with standardised tests in general. It may assist to isolate more effectively those areas of reading skills of concern to Indigenous children. Evaluators should accept that assessments in their various formats are necessary and comparisons should be made between sections of Australian society to highlight both deficiencies and achievements.

REFERENCES:

Andrich, D. (1978a). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement, 38, 665-680.

Andrich, D. (1978b). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

Andrich, D., & Godfrey, J. (1978-9). Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: an empirical investigation using a latent trait model. Reading Research Quarterly, XIV (2), 182-200.

Brandenburg. P. (c. 1984). Kimberley Standard English Vocabulary Test. Available from M. Kormendy of Edith Cowan University, Perth.

Cataldi, C., & Partington, G. (1998). Beyond even reasonable doubt: Student assessment. In G. Partington (Ed.), Perspectives on Aboriginal and Torres Strait Islander education (pp. 309-332). Katoomba: Social Science Press.

Drew, N. (2000). Psychological testing with Indigenous people in Australia. In P. Dudgeon, D. Garvey & H. Pickett (Eds.), Working with Indigenous Australians: A handbook for Psychologists (pp. 325 - 333). Perth. Guanda Press.

Fisher, J. P., & Glenister, J. M. (1992). The Hundred Pictures Naming Test. Hawthorn: Australian Council for Educational Research.

Godfrey, J., Partington, G., Harslett, M., & Richer, K. (2001, in press). Attitudes of Aboriginal students to schooling. Australian Journal of Teacher Education.

Godfrey, J., Partington, G., Richer, K., & Harslett. M. (2001). Perceptions of their teachers by Aboriginal students. Issues in Educational Research, 11(1), 1-13.

Knibb, K. (1995). EdStats. Version 1.0.5. Available from Edith Cowan University, Mount Lawley.

Knibb, K. (1996). EdStats User's Guide. Mount Lawley: Mathematics, Science & Technology Education Centre, Edith Cowan University.

Kormendy, M. (1988). Western Australian Action Picture Test. Available from author at Edith Cowan University, Perth.

Ministerial Council on Education, Employment, Training and Youth Affairs. (1995). National Strategy for the Education of Aboriginal and Torres Strait Islander Peoples; 1996- 2002. (P. Hughes, Chairperson). Canberra: Department of Employment, Education, Training and Youth Affairs.

Partington, G., & McCudden, V. (1992). Ethnicity and Education. Wentworth Falls: Social Science Press.

Rasch, G. (1980). Probabilistic models for intelligence and attainment tests (expanded edition), Chicago: The University of Chicago Press.

Vardy, I. (1991). Phonological Profile for the Hearing Impaired Test. Perth: Iris Vardy.

Waddington. Neil J. (2000). Diagnostic Reading and Spelling Tests 1 & 2 (Second Edition). Strathalbyn, S.A: Waddington Educational Resources.

Wright, B. D. (1985). Additivity in psychological measurement. In E. E. Roskam (Ed.), Measurement and Personality Assessment (pp 101-112.). Amsterdam: Elsevier Science Publishers B. V..

Wright, B., & Masters, G. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.