Numeracy Test Item Readability During Transition from Pre-School to School

Judith A. Mousley
Deakin University

A key test used in Australia to assess the mathematical knowledge of young children uses illustrations of objects such as coins and three-dimensional shapes. This study explored the effects of giving 104 kindergarten children, aged 4-5 years, the questions with either moveable objects or illustrations. It was found that children who were categorized by their teachers as having "higher levels of numeracy" scored well on test questions using either illustrations or objects, while children who were categorized as having "lower levels of numeracy" scored higher with objects than with illustrations. This result could have implications for consideration of test item readability in relation to graphicacy.

The Research Context

As in other countries, there is increasing interest in provision in Australia of accessible but challenging mathematics education for 3- to 5-year-old children and growing emphasis on assessment to inform program planning, transition to school, and intervention programs (Doig, McCrae, & Rowe, 2003). While most evaluation in pre-schools uses observation, a test used by some experts to assess school readiness is *I Can do Maths* (Australian Council for Educational Research [ACER], 2000). The ACER developed this kit initially for use in the major national project, *Curriculum and Organisation in the Early Years of Schooling* (de Lemos, 2000), which investigated relationships between school entry age, school structures, and later learning outcomes.

The *I Can do Maths* teachers' guide (Doig & de Lemos, 2000) includes norm tables and scales for children in pre-Grade 1 as well as profile templates, diagnostic maps, and advice about preparation of descriptive reports. Two tests are available: Levels A and B. Test A is suitable for school-entry children, with the some of its questions being drawing on "Curriculum Level 1" material (ACER, 2000, p. 22). The introduction to the teachers' guide says that the purpose of the test kit is:

> … to inform teachers and parents about children's development in numeracy in the early years of schooling … [resulting in] descriptive and normative reports of children's performance in number, measurement and space [geometry], and not simply a score, so that planning a teaching program appropriate to an individual child's needs is made easier. (Doig & de Lemos, 2000, p.5)

Although the Level A test booklet was designed for use with children in their first year of schooling, it is also used in pre-school settings. Such resources are used to assess the readiness of children to proceed from pre-school (called "kindergartens" in Victoria and throughout this paper) to school. Pre-school Field Officers are available in all regions to evaluate children and give professional advice to the teachers and parents about the transition and children's readiness. Many of them test children's literacy and numeracy using commercially available tests along with a battery of other instruments that focus on social, emotional, physical, language and literacy development.

The questions in *I Can Do Maths* are designed so that they may be presented to half- or whole-class groups of children. Typically, teachers read the questions aloud while the children use the supplied booklets to mark their responses. For example, the children may be asked to "Put a tick under the 10-cent coin" (Q3) and to "Put a cross on the shortest snake" (Q5). The pencil-and-paper nature of the test and the use of images raise the question of how this implementation process affects individual children's performance.

In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2). Palmerston North, NZ: MERGA.

Whether *I Can Do Maths* and similar tests from other countries are to be used to assess children's readiness for school or progress in school or to inform curriculum design and program planning, it is important to know which specific modes of assessment best measure very young children's knowledge of mathematical words and concepts. The nature and wording of the questions is significant, as are illustrations and other aids to comprehension. This paper focuses on the latter aspect: illustrations.

If the format of illustrations in whole-class assessment instruments affects results, and particularly for some groups of children, then at least this needs to be understood so that teachers can be advised to probe further with the children most likely to need individual assessment, using other materials. Thus the questions of (a) whether two-dimensional representations affect comprehension, and (b) if so, whether the effect is common across the range of children, are vital. This is especially applicable in the context of the school-readiness testing that is now being used in many countries.

The research questions for this study were "Does replacing the two-dimensional representations in the Level A *I Can Do Maths* test with real objects improve young children's test results?" and "If so, are children with different levels of achievement affected to the same extent?"

Literature Review

The importance of finding out what mathematics children know and use then building on this in learning programs in kindergarten and the first school year has been stressed by Ball, (1990), Ginsberg et al (2006), Perry (2000), and Sarama and DiBiase (2004), amongst others. Most early numeracy assessment is undertaken through observation (e.g., Lidz, 2003; Twaddel, 2000), one-to-one clinical interviews (McDonough, Clarke, & Clarke, 2002; Pearn, 1998; Wright, Martin & Stafford, 2006; Gervasoni, 2000), portfolios to record development of specific mathematical concepts (Stenmark, 1991), and the recording of narrative evaluations such as "learning stories" (Perry, Harley, & Dockett, 2006, p. 1). However, because of their relative convenience and ability to provide information about the class as a whole as well as about individual children, broader-based paper-based testing is common, even in the very early years (Lidz, 2003).

Wildy, Louden and Bailey (2001) have reported evidence of support among practitioners for the use of entry-level assessment programs across Australia, and this is happening in many countries in relation to literacy and numeracy. For example, UK teachers now use standardised "performance indicators" (Tymms, 1999, p. 1), with children pointing to illustrations on a computer screen in the PIPS On-entry Baseline Assessment (CEM, 2006), a test that has been translated into Dutch, French, German, Thai, Urdu, Bengali, and Cantonese. Of course, there are forms of entry-level assessment other than standardised tests, such as that arising from Victoria's Early Years Numeracy Project (Clarke et al., 2002) and New Zealand's School Entry Assessment (Ministry of Education, 1996), and many involve the use of some illustrations. Further, there is a range of early childhood diagnostic and rating scale instruments (e.g., Lidz, 2003) as well as research instruments that include illustrations.

Literacy is recognised as a factor in early childhood assessment. The teacher's guide for *I Can Do Maths* recognises this: "All questions are read to children to avoid performance being affected by reading factors" (Doig & de Lemos, 2000, p. 5). However, typical mathematics questions for young children also require "graphicacy" (Anning, 2003, p. 1). Graphicacy means competency with the interpretation of illustrations and

representations, including handling shape recognition, size differences, perspective, (Kress, 1997), and the spatial orientation and occlusion of illustrations (Cox, 1991).

While there is ample research on school-age children's ability to interpret graphs, diagrams, charts and tables in mathematics worksheets and texts (see summary by Department for Education and Employment, UK, 1998), there seems to be no relevant research involving children aged under 6 on their ability to interpret illustrations. However, Anning (2003) and Cox (1991) each pointed to the need for research into how children's performance with manipulatives compares with their performance with diagrammatic representations. This research makes one contribution to this gap in empirical findings.

Methods

The research compared the number of correct answers to "original" and "modified" questions. Original questions had illustrations and modified questions used manipulatives. The original set of questions comprised the 20 easiest questions from the *I Can Do Maths: Level A* test. The modified set used almost the same words and made the same numeracy demands but used real objects selected from the wealth of toys and materials readily available in the kindergarten. For example, one *I Can Do Maths* question reads "Put a tick on [the illustration of] the cone" and "a cross on the cylinder". As the children were not required to write in the research trials, the original question for the research asked the child to "Point to the cone" then to "Show me cylinder" as they were shown the row of pictures of 3-dimensional shapes photocopied from the *I Can Do Maths: Level A* booklet. The modified question used the same words but the child pointed to a set of wooden 3D shapes. Thus two sets of 20 equivalent questions were constructed, one with illustrations requiring manipulatives. The 40 questions were alternated to make 2 equivalent tests (Tests 1 and 2), as shown in Figure 1, so that all children were asked both original and modified questions.

Question	Test 1	Test 2
1	Modified	Original
2	Original	Modified
3	Modified	(etc.)

Figure 1: The alternating format of test questions.

Convenience sampling was the basis of the selection of the research venues, as the 3 kindergartens and 2 long day care centres selected were relatively accessible during 2007 and 2008 and involved in a larger research project. The subjects were 104 children (50 girls and 54 boys), aged 4–5 years who were soon to attend school. Each kindergarten teacher was asked to provide a list of 5–8 children whom they thought had "higher levels of numeracy", as well as 5–8 with "lower levels of numeracy". Only these children were tested, 104 in all. I stress that there was no measure of achievement—these were teachers' perceptions—but for convenience will call them "achievement groups" in this paper. The mean ages of higher and lower-achieving children were 4 years 7 months and 4 years 5 months respectively. Children from the random lists of 52 "higher achievers" and 52 "lower achievers" were allocated Test 1 or Test 2 in turn.

Clinical tests/interviews, taking 30-40 minutes each, suited the age of the children and enabled notes to be taken about responses, including actions and comments. No deep probing for reasons for responses was undertaken. The children were asked to shake their

heads of they did not know an answer, rather than to guess—although some guessing is bound to confuse the results of any testing.

Correct questions were scored 1, and incorrect answers scored 0. Frequencies, standard deviations and levels of significance (by *t*-test) were calculated for mean scores on Tests 1 and 2, each pair of questions, all original questions, and all modified questions. *T*-score analysis was used to compare the means for each original compared with modified question, all original compared with all modified questions, the 2 achievement groups, and the "higher" and "lower" achievement groups against the original and the modified questions. (Detailed results are presented in Mousley, in press). *T*-scores facilitate comparison of scores from matched pairs and can be used with relatively small samples. Examination of notes about children's comments and actions supported statistical analysis.

Findings

As shown in Table 1, the modified questions (those using objects) were answered correctly more frequently than original questions (using illustrations). The difference was noted particularly with money questions and counting questions, a point I return to below.

Table 1.
Comparison of Results for 20 Original and 20 Modified Questions

Original		Modified		*t* [52]
M	*SD*	*M*	*SD*	
11.38	2.54	15.62	1.41	0.00*

Note. *$p < 0.01$

Table 2 shows how results of the achievement groups differed. For higher achievers there was a difference in results between original and modified questions (with means of 7.04 compared with 6.9); but the lower achievers performed better with the modified questions, with means of 2.6 with illustrations and 6.1 with objects.

Table 2.
Comparison of Results for Achievement Groups

	Original		Modified		*t* [26]
	M	*SD*	*M*	*SD*	
Higher achievers	9.02	2.96	9.67	3.06	0.42
Lower achievers	2.37	1.47	5.94	1.41	0.00*

Note. *$p < 0.01$

This statistically significant difference (at $p = .05$) suggests that lower-achieving young children may have lower levels of graphicacy and that this may affect their scores in numeracy tests. I report below some observations of how the higher and lower achievers acted differently with the available resources.

Discussion

Test scores are descriptive and not explanatory, so analysis of children's actions and comments is useful. They pointed to the need for careful consideration of the need for quality illustrations or other images. For instance, it was noted that the illustration-based money question proved difficult for lower achievers. Adults and children rely largely on

the size of coins (rather than the numbers or animals depicted on them) to pronounce their value/name. The illustrations of coins in Question 3 (Q3) of *I Can Do Maths—Level A*, and hence in the relevant original question, were not drawn to the correct size. Indeed the illustration of the 20c was the size of a real 10c coin, and the 50c was the size of a real 20c. Perhaps the intention of the question was to get the children to read the numbers drawn on the coins, but the key skill demanded by the question seemed to be coin recognition.

Further, having real money (as in the modified question) may evoke a better sense of value. The original Q7 showed an illustration of 5 coins, one being a (small) 50-cent piece, along with a drawing of a pencil with a 50c tag tied to it with string; children with this original question being asked to point to the coin needed to buy the pencil. The modified question made available a real pencil, with a 50c price sticker typical of those used in shops, and real coins. The majority of children handling the real materials picked up the pencil and read the 50c sticker then chose the 50c relatively quickly (61% correct), while interpreting the illustrations of the pencil and coins proved more difficult (32% correct).

Objects also facilitated a range of physical strategies such as taking away or movement into groups. For example, Question 30 from the ACER test was "Jill has 36 pencils in her pencil case. She puts 17 pencils on the table. Write how many pencils are in the pencil case now". This was answered correctly by a surprising number of the children using real pencils and a pencil case (32%), but by only 11% who of the children faced with an illustration of an open pencil case. Note, though, that this was still a surprisingly good result given the magnitude of the numbers in the more abstract, illustrated problem.

Achievement groups *and performance*

Overall, the majority of high achievers coped relatively well with the drawn representations as well as the manipulatives. It was the lower achievers who seemed to benefit most from being able to see objects and handle them.

Many of the lower achievers seemed to need to manipulate the objects. For instance, when choosing the smallest star (Q1), they acted by picking up the smallest one and placed it on top of the next star, stacking them or lining them up in order or size, etc. This happened after answering the question in 40% cases. On the other hand, nearly all of the higher achievers merely looked at the stars and pointed to the smallest. Similarly, when asked to choose "the shape that makes the side of the cube", the lower achievers generally picked up the square and placed against the side of the cube to test their choice, and were less able to complete the task correctly when presented with drawings; while many higher achievers confidently pointed to the square in both original and modified questions. Nevertheless, the lower-achievers' initial choice of the correct shapes indicated a good grasp of the required response and hence of the mathematical knowledge involved. The act of checking made the lower achievers seem slower and less confident, but it is important to note that their first choice was often correct when the range of manipulatives was offered. A surprising number of both groups (63%) knew which was "the cube".

Being able to move the objects also overcame some common counting problems experienced in early childhood. For example, when counting butterflies (Q8), the higher achievers just touched butterflies or buttons and counted them accurately, so the ability to move the buttons in the modified question did not make the problem much easier. 33% of the lower achievers either did not seem to be able to say the necessary number words in order or could not "touch and count" any objects with 1:1 correspondence, while the others had this skill. (Again a good proportion, considering the first year school curriculum.) When faced with illustrations of 11 butterflies not drawn in rows, 82% of the lower

achievers who were able to count made the mistake of counting some butterflies twice or of missing some. However, when presented with 11 buttons that they could physically moved or picked up, 82% of the lower achievers who could count were able to answer the question correctly. Some moved each button from one area to another while saying a counting word, and the others counted while picking them up, only .8% just pointed at or touched the buttons. These lower achievers being able to count 11 moveable objects but making mistakes when counting 11 illustrations suggests that systematic movement of objects is an important stage in learning to count immovable objects—an hypothesis that needs further investigation. The important point is that for the lower achieving children who could count, it seems that the presentation of immovable illustrations made the question quite difficult.

Q14 in *I Can Do Maths* asks children to identify a cone from an illustration of 4 three-dimension shapes, and Q15 uses the same illustration for "Put a cross on the cylinder". 61% from the higher achievers' group and 46% from the lower group identified the cone correctly from the picture set. When faced with 3D wooden shapes, all children in both achievement groups picked up the cone, many making comments or actions about ice-cream cones. It seemed clear that using real objects helped children to identify the shape that had a familiar name. Only 3 children (6%), 1 higher and 2 lower achievers, identified the "cylinder" correctly from the illustrations, and 3 others (all high achievers) identified the wooden cylinder.

It is important to note that provision of objects may change the nature of what is being tested. For example, in the ACER's Q4 and Q5, children are asked to judge the length of illustrations of sets of snakes and indicate the longest snake and the shortest snake. While the snakes were drawn relatively parallel in the *I Can Do Maths* booklet (and hence in the original research question too), they could not be "lined up" to make a comparison. With snakes of the same length as the originals made from thick wool and presented in the same orientation, all of the higher achieving children answered correctly, with 77% and 61% just pointing to the longest and then the shortest snakes respectively, while 23% (longest) and 38% (shortest) physically lined up the "noses" of the model snakes. Of the lower achievers faced with the illustration, 34% picked the longest snake correctly and 28% picked the shortest. However, all but 12% of the lower achievers who were faced with model snakes lined up "noses" to compare their lengths, and then gave the correct answers. Thus if the question is only about length, testing understanding of "longest" and "shortest", the lower achievers using the models performed as well as the higher achievers; but if it is also about understanding the more complex conservation of length then being able to move the model snakes destroys that component of the test. Again, it is important to note that all of the children understood "longest" (100%) and most comprehended "shortest" (91%), and the majority of both groups (63%) recognized the need to have a common starting point when comparing lengths—and these are both key measurement concepts. Many teachers, when faced with the incorrect results for this question in to original *I Can Do Maths* test, may assume that children still needed to be taught to compare and describe lengths, and may not realise that the children could already do this with manoeuvrable objects.

In summary, while there was no significant difference for high achievers between the original questions involving illustrations and the modified questions involving objects ($t[26] = .42$, $p>0.05$) there was a significant difference in the performance of the lower achievers ($t[26] = .00$, $p<0.05$). The important finding of this research is that when objects rather than illustrations were used with the same test questions, children identified as having lower levels of numeracy achieved higher test scores. From observation and

analysis of specific results, it seemed clear that the difference was often not in mathematical knowledge but in the way children coped with two-dimensional illustrations or used the objects provided.

Conclusion

Of course, the ACER used much larger numbers of young children when calculating norms for the *I Can do Maths* test kits, and this paper in no way challenges the accuracy and usefulness of the kits, particularly given the difference in teacher time that is required for the ACER tests compared with one-to-one interviews with objects. They also did extensive testing of each question. That is, I am not attempting to address, here, the efficacy of clinical interviews compared with broader-based testing or to challenge the usefulness of any specific *I Can Do Maths* questions. The main advantage of the ACER's *I Can Do Maths* test is that it can be administered in a pencil and paper format with groups of children in their first year of schooling. This is convenient in terms of teachers' time. The test is widely used, and it is easier to produce illustrations in the test papers than it would be to provide and manage the use of manipulatives with groups of children. However, the introduction to the *I Can Do Maths* Teachers' Guide (Doig & de Lemos, 2000) suggests that its purpose is to inform teachers and parents about children's development as well as to inform teaching and learning programs, and the results of this research suggest that further probing by teachers of children's individual knowledge and skill would be necessary for these aims to be met.

To draw empirical conclusions, it would be necessary to trial the use of illustrations and manipulatives with many more children of school-entry age, using a range of tests including the illustrated computer-based programs currently in use as well as print materials with higher-quality and more accurate drawings and/or photographs. However, the findings of this research do suggest that illustrations in test questions may affect the results, with the need to decode illustrations making some questions harder. Interpreting illustrations requires graphicacy, so the test results from printed or computer-based standardised tests (as well as research and interview instruments) are not only evaluations of children's mathematical knowledge and skills, so children's mathematical abilities may be under estimated because they are struggling with "graphicacy" rather than the mathematical knowledge and skills required by the questions themselves.

The differences in results were most significant for children who had been identified as having "lower levels of numeracy", but such children may be scored lower than they deserve because of the test format. If tests like the ACER one are used to determine readiness for school, graphicacy needs to be recognised as a factor. A recommendation arising from this research is that teachers do not just accept results from assessment instruments that use illustrations, and another is that illustrations should be as realistic as possible (such as illustrations of coins being their correct size). The results here would also have implications for school-entry advisers as well as researchers, teachers, and others who construct or use pre-school and school entry-level assessment instruments. As Lids (2003) noted, "Standardized tests can be invaluable aids for determination of risk and program eligibility, but the use of these tests should not be a substitute for good judgment and thinking" (p. 153).

References

Anning, A. (2003). Pathways to the graphicacy club: The crossroad of home and pre-school. *Journal of Early Childhood Literacy, 3*(1), 5–35.

Australian Council for Educational Research (2000). *I can do maths: Level A*. Camberwell, Vic: ACER.
Ball, D. L. (1990). Breaking with experience in learning to teach mathematics: The role of a preservice methods course. *For the Learning of Mathematics 10*(2), 10–16.
Clarke, D., Cheeseman, J., Gervasoni, A., Gronn, D., Horne, M., McDonough, A., Montgomery, P., Roche, A. Sullivan, P., Clarke, B., & Rowley, G. (2002). *Early Numeracy Research Project (ENRP): Final report.* Available online: http://www.sofweb.vic.edu.au/eys/num/enrp.htm. Retrieved Sept. 9, 2002.
Cox, M. (1991). *The child's point of view*. Hemel Hempstead: Harvester Wheatsheaf.
Curriculum Evaluation and Management Centre (CEM) (2006). *Performance indicators in primary schools: ePIPS Electronic On-entry baseline assessment*. Durham, UK: University of Durham. Retrieved Feb. 12, 2009, from http://www.cemcentre.org/RenderPage.asp?LinkID=22218005.
De Lemos, M. (2000). *Curriculum and organisation in the early years of schooling*. Camberwell, Vic: ACER.
Department for Education and Employment (1998). *The implementation of the National Numeracy Strategy: The final report of the numeracy taskforce*. Sudbury, Suffolk: DfEE.
Doig, B., & de Lemos, M. (2000). *I Can Do Maths Teachers' Guide.* Camberwell, Vic: ACER.
Doig, B., McCrae, B., & Rowe, K. (2003). A good start to numeracy: effective numeracy strategies from research and practice in early childhood. Camberwell, Vic: ACER.
Gervasoni, A. (2000). Using growth point profiles to identify Year 1 students who are at risk of not learning school mathematics successfully. In J. Bana & A. Chapman (Eds.), *Mathematics education beyond 2000: Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 275–283). Perth: MERGA.
Ginsburg, H. P., Kaplan, R. G., Cannon, J., Cordero, M. I., Eisenband, J. G., Galanter, M., et al. (2006). Helping early childhood educators to teach mathematics. In M. Zaslow & I. Martinez-Beck (Eds.), *Critical issues in early childhood professional development* (pp. 171–202). Baltimore, MD: Paul H. Brookes.
Kress, G. (1997). *Before writing: Rethinking the paths to literacy*. London: Routledge.
Lidz, C. S. (2003). *Early childhood assessment*. Hoboken, NJ: John Wiley Associates.
McDonough, A., Clarke, B. A., & Clarke, D. M. (2002). Understanding assessing and developing young children's mathematical thinking: The power of the one-to-one interview for preservice teachers in providing insights into appropriate pedagogical practices. *International Journal of Education Research, 37*, 107–112.
Ministry of Education (1996). Revised statement of desirable objectives and practices (DOPs) for chartered early childhood services in New Zealand. *New Zealand Gazette*, October 3, 1996.
Mousley, J. (in press). *Evaluation and assessment of numeracy in early childhood*. Research in Practice series. Watson: Early Childhood Australia.
Pearn, C. (1998, December). *Mathematics intervention: A school-based project informed by mathematics education research.* Paper presented to the 1998 annual conference of the Australian Association for Research in Education, Adelaide.
Perry, B. (2000). *Early childhood numeracy*. Canberra: DETYA, Commonwealth of Australia.
Perry, B. Harley, E., & Dockett, S. (2006). Powerful ideas, learning stories and early childhood mathematics. *Proceedings of the 30th Annual Conference of the International Group for the Psychology of Mathematics Education* (Volume 1, pp. 270–277). Prague: PME.
Sarama, J., & DiBiase, A-M. (2004). The professional development challenge in preschool mathematics. In D. H. Clements & J. Sarama (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 415–46). Mahwah, NJ: Lawrence Erlbaum Associates.
Stenmark, J. K. (1991). *Mathematics assessment: Myths, models, good questions, and practical suggestions.* Reston, VA: National Council of Teachers of Mathematics.
Twaddel, P. (2000). *The learning place: early childhood assessment: education*. Sydney: Learning Place.
Tymms, P. (1999). *Baseline assessment and monitoring in primary schools: achievements, attitudes and value-added indicators*. London: David Fulton.
Wildy, H., Louden, W., & Bailey, C. (2001, July). *High stakes testing in a low stakes environment: PIPS baseline assessment in Australia*. Paper presented to the 3rd International Inter-disciplinary Conference on Evidence-based Policies and Indicator Systems, Durham, UK.
Wright, R. J., Martin, J., & Stafford, A. K. (2006). *Early numeracy: Assessment for teaching and intervention*. London: Paul Chapman Publishing.