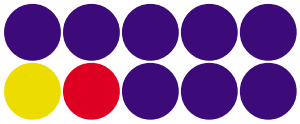
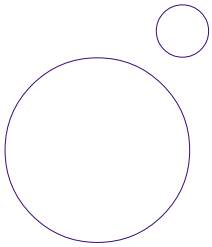


Count Me In Too

1997 Report



A report prepared on behalf of
the NSW Department of Education & Training
by
Dr Janette Bobis
University of Western Sydney, Macarthur
November 1997

Count Me In Too

1997 Report

CONTENTS

Executive Summary

An investigation of the 1997 Count Me In Too Project (NSW Department of School Education) was conducted to examine the degree of agreement between teachers when judging the arithmetical ability of young children on the Schedule for Early Number Assessment (SENA), a performance-based assessment instrument. The SENA was an integral component of the Count Me In Too Project with children's performances being assessed at the start and conclusion of the project. This section presents a summary of the report's findings. The results are organised according to the two data gathering strategies employed in the study - quantitative (teacher ratings of students' performances on the SENA) and qualitative (interviews with teachers).

Quantitative data findings showed that:

1. Generally, there was a high degree of inter-rater reliability between teachers when rating children's performances on the SENA.
2. There was some degree of inter-rater variability on the Forward Number Word Sequence, Backward Number Word Sequence and Numeral Identification aspects of the SENA, but it was not significant.
3. Most of the variability could be accounted for by a small group of raters, and in particular, one rater.
4. Except for three cases, teachers' ratings were highly correlated to the expert's ratings.
5. There was a high degree of correlation between individual teacher's ratings and the mean rating of the whole group.

Qualitative data findings showed that:

1. Some insights into why raters differed on their judgements could be gained by examining teachers' verbal explanation of how they rated each child.
2. There were three possible explanations for the variability in ratings. These were characterised by three strategies that different raters adopted to assist them make decisions regarding children's performance levels. The factors driving these strategies related to the perceived student confidence level, *teacher uncertainty* and teacher *initial impressions* of students' ability levels.
3. Teachers were competent in interpreting strategy use from behavioural indicators that were easily observed, but did not detect less overt clues to the strategy use of children, such as subtle eye movements.

Recommendations

Courses of action that may alleviate the problem of rater severity when attributed to any of the three strategies identified in this study could include training teachers to:

1. Actively search for and take account of less overt behavioural indicators, such as eye movements, to help them make decisions regarding students' thinking strategies and performance levels.
2. Be aware of the impact factors such as student confidence, teacher uncertainty and initial impressions have on some teachers' judgements.

Report of the Count Me In Too Project

This report presents the findings of an investigation into the Early Number Project (Count Me In Too) conducted by the NSW Department of School Education in Terms 1 and 2 of the 1997 school year. The aim of the investigation was to determine the degree of agreement between teachers when judging the arithmetical ability of young children on a performance-based assessment instrument that was an integral component of Count Me In Too. To set the context for the study and so that the implications of the findings may be fully comprehended, background information relating to the Count Me In Too project has also been included.

Background to the Study

Origins and Aims of Count Me In Too

In 1996 the NSW Department of School Education trialed an early number project (Count Me In) in 13 schools throughout NSW. The aim of the project was to develop the knowledge of K-2 teachers in early number with the ultimate aim of improving young children's mathematical abilities.

The project employed a work-based model of professional development, with mathematics consultants working in classrooms alongside teachers. Exactly how consultants became involved varied from school to school, but basically their role was to assist teachers with the implementation of the learning framework espoused by the CMI project. Generally, this was achieved by consultants helping teachers assess the mathematical development of children in their class, and by helping them plan and implement developmentally appropriate learning and teaching experiences.

The evaluation of Count Me In (CMI) indicated through observational and self-report anecdotal evidence that the project had overwhelming success (Bobis, 1996). The report found that teachers generally increased their knowledge and understanding of mathematical content, of children's thinking strategies and of how children learn mathematics. It also found that 90% of the students participating in the project had progressed in their numerical development as indicated by a performance-based assessment instrument called the Schedule for Early Number Assessment or SENA.

In 1997 the NSW Department of School Education decided to extend the project to include 53 DSE funded schools and 40 consultants. Count Me In Too (CMIT) started in Term 1 of the school year and continued into Term 2 with many schools not completing the follow-up assessment of children until early Term 3 (August 1997). At the time this investigation was conducted, most teachers had not yet started the post-project assessment of their children.

Description of the SENA

The SENA was developed over a period of approximately five years and has been used extensively by teachers and researchers to assess the early arithmetical development of young children (Wright, 1996). It was used by all teachers to monitor the arithmetical development of their children throughout CMI and throughout the 1997 implementation of the program, Count Me In Too (CMIT). The SENA involves the presentation of a number of 'tasks' or problems to a child in an individual interview situation. Examples of tasks include: asking the child to say the number words from one to twenty,

or given two covered collections of counters and asking the child how many in all (see Appendix A for a copy of the SENA). It is the role of the interviewer (the classroom teacher) to elicit a child's most sophisticated strategy and then determine where each response might be categorised within a Learning Framework of predetermined stages or levels of development (Wright, 1994). Having teachers assess and monitor the development of the children in this manner is an integral component of CMIT. From initial and subsequent assessments, teachers make decisions regarding learning experiences necessary for individual children and groups of children to help them advance through the stages and levels of the Learning Framework.

Description of the Learning Framework

Analysis of student performance on the SENA includes determination of a level or stage on each of five aspects concerned with the arithmetical development of young children. These five aspects relate to a child's:

(a) level of sophistication of counting and other strategies to solve relatively simple addition and subtraction problems (Early Arithmetical Stages or EAS);

(b) facility with forward number word sequences (FNWS);

(c) facility with backward number word sequences (BNWS);

(d) ability to identify numerals (NID); and

(e) understanding of tens and ones (Base 10).

The predetermined stages and levels, along with statements or criteria describing behavioural indicators, were devised by Wright (1994) and presented to teachers as a Learning Framework in Early Number (see Appendix B for a summary of the Framework and Appendix C for the criteria). Teachers were given a 1 day training session at the start of CMI and CMIT to assist them with the assessment and interpretation of their students' performances on the SENA. SENA interviews with children were video-taped and later analysed by the classroom teachers. This was done initially with the assistance of district mathematics consultants and eventually by the teachers themselves.

Rationale for the Investigation

Performance-based assessment is by no means a new procedure for evaluating the academic merits of children in educational institutions across Australia (Clauser, Subhiyah, Nungester, Ripkey, Clyman & McKinley, 1995; Joffe, 1990). However, the current emphasis in curriculum documents (for example, Department of School Education, 1994) on the use of a variety of assessment procedures for evaluating the achievement of students has seen performance-based

assessment procedures develop a more prominent status. Such large scale assessment makes sending 'expert' raters to every school impractical - besides which, such a plan would not promote the benefits of performance-based assessment techniques to teachers (Mendelovits, 1997). Using teachers, who have limited access to training, to rate the performances of students raises questions in regard to the reliability of their judgements.

In a study designed to investigate the accuracy and reliability of teacher ratings of Years 3, 7 and 10 students' speaking performances, Mendelovits (1997) found that the "raters are inconsistent in the degree of severity they apply in assessing different performances, and vary widely from each other in their assessments" (p.17). This finding applied particularly when teachers were rating students from a Year level other than the one they teach. Furthermore, it was found that inter-rater reliability did not improve even after exposure to a video-tape designed to provide teachers with training in the procedure. However, it was suggested that when the mean standards of achievement from a state-wide or sub-group of teachers is considered, "the ratings seem about as accurate as those that would have been obtained had teams of expert markers been sent" across the state (p. 18).

Unlike the study described by Mendelovits, where teachers were required to rate the performances of students from Year 3 to Year 10, CMIT teachers were only dealing with students in the K-2 range. Thus, it is plausible to predict that the variability of teacher ratings in the CMIT project would be less pronounced than that found my Mendelovits. However, such findings emphasise the necessity of determining the reliability of teacher ratings in the CMIT project, particularly given the pivotal role such ratings play in determining instructional decisions for individual and groups of children to help them advance through the stages and levels of the Learning Framework. In addition, if the performance of students from different classes, schools or regions are to be compared for any reason in the future, it is imperative that the inter-rater reliability be evaluated. While the scope of the present investigation does not allow any speculation as to the educational benefit of collating and comparing achievement levels of children assessed by different teachers on an assessment instrument of this nature, an investigation into the degree to which teachers' judgements concur regarding student performances may be necessary simply to provide educational authorities an indication of the SENA's reliability.

Design and Methodology

Materials

Evaluation Tape

For the purposes of this investigation, a video tape was made of children performing tasks from the SENA. Existing video-taped SENA interviews were made available by the Department of School Education and segments were selected for inclusion in the study on the basis of their clarity and audibility of children's responses. The resultant tape contained excerpts from SENAs conducted with 5 different children of varying abilities. Two children were recorded performing Numeral Identification tasks, Forward and Backward Number Word Sequence tasks, two were recorded performing tasks that would allow their Early Arithmetical Strategies to be categorised and one child was recorded performing Base Ten Strategy tasks. Children used in the evaluation tape were not known to the teachers involved in the investigation.

Participants and Procedure

The invitation to participate in this study was extended to over 40 K-2 teachers already involved in CMIT. Sixteen K-2 teachers from 7 different primary schools in the Sydney Metropolitan area and one 'expert' mathematics consultant volunteered to participate.

The teachers and consultant viewed the prepared video-tape showing children performing tasks from the SENA in an individual interview situation. The teachers and consultant were asked to rate the children's performances on each task by allocating them to the stages or levels of development indicated on the Learning Framework. These ratings were recorded on a standard response sheet that was similar to the recording sheet teachers used when rating their own children on the SENA (see Appendix D). In addition, participants were asked to explain their reasons for rating each child's performance. To this end, the video-tape was paused frequently to allow verbal comments to be audio-taped. These tapes were transcribed and were used to provide qualitative information regarding variability in teacher ratings of student performances. It was anticipated that the consultant's qualitative responses would allow 'key features' to be identified providing an indication as to the level of sophistication of explanations provided by teachers.

Interviews were of approximately 1 hour in duration. Each video segment could be viewed a number of times until teachers made their final decisions. They could 'go back' and change their ratings to any segment at any time and could refer to the Learning Framework documentation to assist them make their decisions. Teacher ratings were collated and analysed to provide quantitative data in respect to the inter-rater reliability of teachers.

Results and Discussion

Characteristics used to assess the reliability of teacher ratings in the investigation were:

(a) the variability of raters' judgements on the different aspects of the Learning Framework;

(b) the variability of raters' judgements across aspects for different children;

(c) the correlation of individual teacher's ratings with that of the expert's ratings; and

(d) the correlation of individual teacher's ratings with the mean rating of the whole group.

Quantitative Data

Table 1 presents the raw score ratings awarded by teachers and the expert to each child on the various aspects of the Learning Framework evident in the SENA video excerpts. It can be seen that for a few aspects there was some degree of variability in the judgements of certain students' performances, namely child 1 on the Forward Number Word Sequence (FNWS), Backward Number Word Sequence (BNWS) and Numeral Identification (NID) aspects and child 2 on the FNWS and BNWS aspects.

Table 1 *Individual teacher and expert raw score ratings for each child on various aspects of the Learning Framework.*

School/ Teacher Code*	Forward Number Word Sequence (Levels 0-5)		Backward Number Word Sequence (Levels 0-5)		Numercial Identification Levels (0-4)		Early Arithmetical Strategies		Base 10 Strategies (Levels 1-3) (Stages 0-4)
	Child1	Child 2	Child1	Child 2	Child1	Child 2	Child 3	Child 4	Child 5
Expert	4	3	3	1	2	1	2	3	1
A1	4	3	3	3	2	1	2	3	2
A2	4	3	3	2	2	1	1	3	1
A3	2	2	2	1	1	1	2	3	1
A4	3	3	3	3	2	2	3	4	1
A5	4	3	3	2	3	1	3	3	1
B6	4	4	3	1	1	1	3	3	1
B7	5	3	3	0	3	1	2	3	2
C8	3	1	1	1	3	0	1	4	1
C9	4	3	3	0	1	1	1	3	2
C10	1	4	1	0	1	1	2	3	1
D11	4	2	3	1	1	1	2	3	1
E12	4	3	3	1	1	1	1	3	1
F13	4	1	1	0	2	1	2	3	1
G14	3	0	1	0	1	1	2	3	1
G15	4	3	3	2	1	1	2	3	0
G16	3	3	3	0	1	0	2	4	1

* Each letter indicates a different school

This observation is supported by the variance and range scores shown in Table 2. Ideally, the range score should be close to 0, indicating a high degree of agreement between the raters. Similarly, and more importantly, variances closer to 0, would also indicate a greater degree of inter-rater agreement.

Results in Table 2 show clearly that disagreement between raters was more pronounced for some aspects and not others, (namely FNWS and BNWS) and only for some children (for example, compare NID of child 1 with that of child 2). In the case of children 1 and 2 for FNWS and BNWS ratings, a range of 4 is quite high. However, these must be considered in the light of the corresponding variances. For example, a range of 4 for ratings associated with the FNWS ability of child 1 can be accounted for largely by two teachers, who awarded a rating at opposite ends of the scale. Conversely, FNWS ratings for child 2 also have a range of 4, but have a much higher variance. This is caused by the fact that many more teachers awarded ratings at opposing ends of the scale. Hence, there was far more variability in teacher ratings for child 2 on this aspect of the framework.

Table 2 *Mean ratings, standard deviations, range of ratings and variances for the group on each aspect for individual children.*

	Child 1			Child 2			Child 3	Child 4	Child 5
	FNWS (0-5)	BNWS (0-5)	NID (0-4)	FNWS (0-5)	BNWS (0-5)	NID (0-4)	EAS (0-4)	EAS (0-4)	BASE 10 (1-3)
Mean	3.53	2.471	1.65	2.59	1.06	0.94	1.94	3.17	1.12
SD	0.94	0.87	0.78	1.06	1.03	0.42	0.65	0.39	0.48
Variance	0.89	0.76	0.61	1.13	1.05	0.18	0.43	0.15	0.23
Range	4.00	2.00	2.00	4.00	3.00	2.00	2.00	1.00	2.00

FNWS Forward Number Word Sequence
BNWS Backward Number Word Sequence
NID Numeral Identification
EAS Early Arithmetical Strategies
BASE 10 Base 10 Strategies

For the purposes of the present investigation it was important to discern if the variability calculated for any set of ratings was significant. To this end, an analysis of variance (ANOVA) was employed. Results indicated that there were no significant differences in the mean ratings for the different aspects of FNWS ($df=16$, $F = 0.78$, $p > 0.001$), BNWS ($df=16$, $F = 0.89$, $p > 0.001$), NID ($df=16$, $F = 0.59$, $p > 0.001$), and EAS ($df=16$, $F = 0.30$, $p > 0.001$), or for the overall ratings of individual students who were rated on more than one aspect, $df=16$, $F = 0.91$, $p > 0.001$ and $F = 0.83$, $p > 0.001$ (child 1 and 2 respectively). This means that while there was some degree of inter-rater variability on the various aspects of the SENA, it was not significant.

Having confirmed that the disagreement between raters was not significant, it was still of interest to locate the source(s) of variability to determine if it was a result of ambiguities in the criteria used to rate students' performances on the SENA, a result of teacher error or whether such differences in opinions will need to be tolerated given the fact that we are unlikely to ever gain complete unanimity of ratings even among a

group of expert raters. To this end, individual teacher's ratings were correlated with that of the expert's ratings and with the mean rating of the whole group (see Table 3).

Table 3 *Correlations between teacher ratings and the expert, and with the whole group*

School/Teacher Code	Correlations with Expert's Ratings	Correlations with the mean rating of the whole group
Expert	1.00	.98**
A1	.76*	.77*
A2	.89**	.87**
A3	.76*	.85**
A4	.62	.69*
A5	.89**	.85**
B6	.89**	.89**
B7	.89**	.85**
C8	.58	.66*
C9	.87**	.86**
C10	.45	.51
D11	.92**	.92**
E12	.93**	.93*
F13	.72*	.77*
G14	.55	.66*
G15	.88**	.88**
G16	.88**	.93**

* $p < .05$, ** $p < .01$.

Correlations between the teachers' and the expert's overall ratings range from 0.45 to 0.93, with 12 teachers' ratings being significantly correlated to that of the expert's ratings (teachers A4, C8, C10 and G14 were not significant). However, when each teacher's ratings were correlated with the mean rating of the whole group, it was found that only one teacher's ratings were not correlated significantly (teacher C10). While this further supports the finding that generally, there was a high degree of inter-rater agreement, it also indicates that most of the variability could be accounted for by one rater.

Interview Data

While a large proportion of the variability between raters (shown to be insignificant by an ANOVA) could be traced to a small group of raters (namely C10), it does not provide reasons for the disagreement that existed. The interview data, however, does provide some insights into why raters differed on their judgements and is an important aspect of this study.

The interview data provides evidence of each participant's rationalisation for rating students' performances at particular levels or stages of arithmetical development. The more extensive knowledge base and experience in analysing SENA interviews, enabled the expert to provide a more elaborate rationale than most other raters who had only used the SENA as part of their CMIT involvement. As intended from the outset of the investigation, this rich text provided a useful 'benchmark' to gauge the extent to which other raters were able to identify behavioural clues as to the strategies children used to complete SENA tasks.

Four teachers were selected for closer analysis - A4, C8, C10 and G14. However, due to a technical fault, the taped interview with A4 was inaudible and had to be excluded from the analysis. Transcripts of interviews with each of the other three raters were scrutinised for clues to why their ratings did not correlate to that of the expert, or to the group's mean rating, to the same extent as other teachers' ratings. In the case of each rater, three instances were examined when they rated more severely relative to the mean rating of the group and/or to that of the expert - FNWS ratings for children 1 and 2, and the BNWS rating for child 1.

Three possible explanations for the variability in ratings emerged from the interview data. These explanations take account of the 'strategies' that different raters adopted to assist them make decisions regarding children's performance levels. The factors driving these strategies seemed to relate to the *perceived student confidence* level, *teacher uncertainty* and teacher *initial impressions* of students' ability levels. Each of these strategies and their related factors are discussed here, drawing upon excerpts from interviews as exemplars.

First, there was a tendency to rate more severely, or conservatively, when a child paused before responding to a task - whether the child's response was judged to be correct or not. For example, C8 commented that child 1

> paused at 66, a long pause before she went on. But she still knows her double digit numbers. She paused to change into nine-ty. She needed to work that out. That indicated that she's still not sure of her "ty's", but she is still quite good.....She's really into Level 3 for her forward number word sequence, nearly Level 4. Her confidence puts her backwards.

The fact that a child 'paused' before providing a response that was characteristic of a particular developmental level on the Learning Framework caused some raters to award a lower rating simply because the child appeared not to be "confident" with their response. This rating strategy was also identified in the transcripts of other raters noted for their severity, namely C10.

She was quite competent as far as 10. She could go up to 32 and with prompting she could go up to 73. But she was quite panic stricken about those higher numbers...She was definitely Level 1 for her forward number word sequence and moving into Level 2. She had no confidence at all.

This type of strategy usually resulted in a rater allocating a child to a level that was one or two levels lower than that judged by the expert rater (except in the case of C10 rating the FNWS performance of child 1, where the she rated three levels lower). However, uncertainty as to where a child might 'belong' on the Framework, caused teachers to rate two or even three levels lower than the expert or the mean rating of the group. This was characterised in the interview data by a rater's inability to elaborate upon a child's performance or to provide a rationale for a decision. For example, when C8 and G14 were unsure of how to rate a child they tended to rate more severely. In the case of child 1's BNWS performance, C8 stated, "Oh! I'm not sure. She doesn't give much away here... Level 1 maybe". Similarly, G14 was unsure of where to place child 2 on her FNWS tasks, rating her at Level 0, while the expert rated her performance at Level 3. "I think she's in the initial stages - Level 0 ... no Level 1. No I'll go back to Level 0 just to be sure." In both cases, the raters were unable to rationalise their decisions, possibly due to their inability to identify any behavioural indicators of strategies the children were employing while contemplating or conducting tasks.

In reality, the decision-making strategy characterised by 'rater uncertainty' may never be utilised, since teachers in normal classroom settings are free to seek opinions from other teachers. However, training teachers to take account of less overt behavioural indicators, such as eye movements, may assist them in their decision-making process. The ability to recognise subtle gestures provided the expert with more evidence to verify continuously judgements and to rationalise student ratings. For instance, it was noted that child 1 "did well on '13', was quite quick on '27' and for '69' she was going into a counting phase, you could see slight movement with her eyes. Her eyes were up, indicating that she was really working on that problem".

A third strategy, which may be related to the 'lacking in confidence' strategy described above, was evident when raters ignored a child's performance and rated according to their first impression of a child. For example, C8 rated the FNWS skills of child 2 at Level 1 because "she's just

rote counting, she obviously doesn't really know the names of the numbers". The fact that the child got the number names correct was ignored by the teacher. Analysis of subsequent tasks was not entered into to verify her initial impression. On the other hand, the expert rated her performance at Level 3 only after examining all the tasks requiring FNWS skills.

Her FNWS, through very good teaching I'd say, is quite strong within the range 1 to 30 and beyond 30... Now we have to watch what she does with the number afterward task before we can be confident about her level.

Almost all teachers were competent in interpreting the behavioural indicators that were easily detected, such as subvocalised counting accompanied by mouth or finger movements. However, the majority of teachers did not choose to review segments of the video. They were satisfied with their decisions based solely on the more overt behaviours. While this procedure resulted in the majority of teachers making similar ratings, it was not a reliable one when judging the performance of children who openly displayed few 'clues' as to the type of thinking strategies they were employing. Only the expert and rater A5 initiated more thorough searches to detect additional clues to verify their judgements. This would often involve repeated reviews of the same video segments until the rater was satisfied that every clue to a child's strategy use had been detected.

Summary and Conclusion

This report has presented findings of an investigation intended to determine the degree of agreement between teachers when judging the arithmetical ability of young children on the SENA, a performance-based assessment instrument that was an integral component of the 1997 Count Me In Too Project. Background information relating to the origins of CMIT, its aims and key elements such as the Learning Framework in number (Wright, 1994) were included to provide a context for the study and so that the implications of the findings might be appreciated more fully.

Characteristics used to assess the reliability of teacher ratings in this investigation were: (a) the variability of raters' judgements on the different aspects of the Learning Framework; (b) the variability of raters' judgements across aspects for different children; (c) the correlation of individual teacher's ratings with that of the expert's ratings; and (d) the correlation of individual teacher's ratings with the mean rating of the whole group. Results showed that while there was some degree of inter-rater variability on the Forward Number Word Sequence, Backward Number Word Sequence and Numeral Identification aspects of the SENA, it was not significant. In addition, it was found that teachers' ratings were similar to those that were given by the expert and that there was a high degree of agreement between individual teacher's ratings and that of the whole group.

The source(s) of variability, while not significant, was investigated further so as to determine, if possible, the reasons for disagreement. It was found that most of the variability could be accounted for by a small group of raters, and in particular, C10.

The interview data provided some insights into why raters differed on their judgements. Possible explanations for the variability in ratings were suggested. These were characterised by three 'strategies' that different raters adopted to assist them make decisions regarding children's performance levels. Each of these strategies were characterised by particular 'factors' that seemed to initiate and fuel the use of the strategy.

First, there was the student confidence factor. Teachers tended to rate more severely when a child was seen to be lacking in confidence. Secondly, the rater uncertainty factor. If a teacher was uncertain as to where a child might 'belong' on the Framework, they would rate more severely. This was characterised by a rater's inability to elaborate about a child's performance or to provide a rationale for a rating. Thirdly, the initial impression of a student influenced the final judgements of some raters. This strategy was evident when raters ignored a child's performance and rated according to their first impression of a child regardless of the child's performance on subsequent tasks.

While teachers were competent in interpreting behavioural indicators that were easily detected, such as finger counting, the majority did not engage in detailed searches intent on detecting less overt clues to the strategy use of children. It is possible that training teachers to take account of less overt behavioural indicators, such as eye movements, may assist them in their decision-making process. It is also possible that making teachers aware of the impact certain factors have on some teachers' judgements may alleviate the problem of rater severity to some degree.

References

Bobis, J. (1996). *Report of the Evaluation of the Count Me In Project*. Report submitted to the NSW Department of School Education.

Clauser, B., Subhiyah, R., Nungester, R., Ripkey, D., Clyman, S., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgements of experts, *Journal of Educational Measurement*, 32 (4), 397-415.

Department of School Education (1994). *Assessment in Mathematics K-6*. Sydney: Author.

Joffe, L. (1990). Evaluating Assessment: examining alternatives. In S. Willis, (Ed.), *Being numerate: What counts?* Hawthorn: ACER.

Mendelovits, J. (1997). *Evaluating the reliability of teachers as raters in a performance assessment program*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, March 24-28.

Wright, B. (1994). A study of the numerical development of 5-year-olds and 6 year olds. *Educational Studies in Mathematics,* 26, 25-44.

Wright, B. (1996). Problem-Centred Mathematics in the First Year of School. In J. Mulligan and M. Mitchelmore (Eds.), *Children's Number Learning* (pp. 35-52). Adelaide: MERGA/AAMT.

Appendix A

Overview of the Learning Framework in Number

Part A - Primary Dimension Early Intervention Arithmetical Strategies & Base Ten Strategies	Part B - Secondary Dimension Number Words and Numerals	Part C - Procedural Dimension Partitioning, Patterns and other Aspects
Early Arithmetical Strategies 0 - Emergent Counting Cannot count visible items. 1 - Perceptual Counting Can count perceived items but not those in concealed collections. 2 - Figurative Counting Can count concealed items but counts from one rather than counting on. Has figurative notion of numbers and thus does not need to count perceived items but counts from one to construct a number in additive situations. 3 - Counting on Can use advanced count-by-one strategies. Counts-on rather than counting from "one", to solve addition or missing addend tasks. 4 - Facile Number Sequence Can use a range of non-count-by-one strategies. **Base Ten Arithmetical Strategies** 1 - Ten as ones (Initial Concept of Ten) 2 - Ten as unit (Intermediate Concept of Ten) 3 - Tens & Ones (Facile Concept of Ten)	**Forward Number Word Sequences (FNWS)** 0 - Emergent FNWS 1 - Initial FNWS up to "ten" 2 - Intermediate FNWS up to "ten" 3 - Facile with FNWSs up to "ten" 4 - Facile with FNWSs up to "thirty" 5 - Facile with FNWSs up to "one hundred" **Backward Number Word Sequences (BNWS)** 0 - Emergent BNWS 1 - Initial BNWS up to "ten" 2 - Intermediate BNWS up to "ten" 3 - Facile with BNWSs up to "ten" 4 - Facile with BNWSs up to "thirty" 5 - Facile with BNWSs up to "one hundred" **Numeral Identification** 0 - Emergent Numeral Identification 1 - Numerals to "10" 2 - Numerals to "20" 3 - Numerals to "100" 4 - Numerals to "1000"	Combining and Partitioning Spatial Patterns and Subitising Temporal Sequences Finger Patterns Quinary based Strategies **Early Multiplication and Division** 1 - Counts using perceptual materials 2 - Uses counting, additive or subtractive strategies without perceptual material 3 - Uses known or derived addition or multiplication facts. **Early Fraction Knowledge** 1 - Subdivides a unit into equal parts

Appendix C

Criteria

Model for Development of Early Arithmetical Strategies

Stage 0: Emergent Counting. Cannot count visible items. The child either does not know the number words or cannot coordinate the number words with items.

Stage 1: Perceptual Counting. These children are limited to counting items they can perceive (i.e. see, hear or feel)

Stage 2: Figurative Counting. Children can count concealed items but may include unnecessary activity. For example, given a collection of 5 items and a collection of 3 items (both screened) the child will count from 1 in an attempt to determine the total number of items.

Stage 3: Counting-on. The child typically counts-on rather than counting from 1 when solving tasks involving hidden items. These children count-on to solve additive and missing addend tasks and may use counting-down-from strategies (eg 17-3 as 16, 15, 14 - answer 14) but not counting-down-to strategies (eg 17-14 as 16, 15, 14 - answer 3).

Stage 4: Facile Number Sequence. The child can use a range of strategies that involve procedures other than counting by ones but may include counting by ones. For instance, the child might solve an additive problem using strategies such as compensation, adding to ten, or commutativity.

Model for the Construction of Forward Number Word Sequences (FNWSs)

Level 0: Emergent FNWS. Cannot produce the FNWS from 1 to 10.

Level 1: Initial FNWS up to 10. The child can produce a number word sequence from 1 to around ten. The child cannot produce the number word just after a given number. Dropping back to 1 does not occur at this stage.

Level 2: Intermediate FNWS up to 10. The child can produce the number word just after a given number but drops back to 1 when doing so.

Level 3: Facile with FNWSs in the range 1 to 10. Produces the number word just after a given number in the range 1 to 10 without dropping back, but typically drops back for numbers after 10.

Level 4: Facile with FNWSs up to 30. The child produces the number word immediately following given numbers in the range 1 to 30 without dropping back.

Level 5: Facile with FNWSs up to 100. Produces the number word immediately following given numbers in the range 1 to 100 without dropping back.

Construction of Backward Number Ward Sequences (BNWSs)

Level 0: Emergence of BNWS. Cannot produce the BNWS from 10 to 1.

Level 1: Initial BNWS from 10 to 1. Can produce the BNWS from 10 to 1 but cannot produce BNWSs from number words less than 10. The child cannot produce the number word immediately before a given number, and the "dropping back to 1" strategy is not available to the child.

Level 2: Intermediate BNWS from 10 to 1. The child can produce the number word immediately before a given number up to 10, but typically drops back to 1 when so doing.

Level 3: Facile with BNWSs up to 10. Produces the number word immediately before a given number in the range 1 to 10, without dropping back, but typically drops back to 1 for numbers after 10.

Level: Facile with BNWS up to 30. Produces the number word immediately before given numbers in the range 1 to 30 without dropping back.

Level: Facile with BNWSs up to 100. Produces the number word immediately before given numbers in the range 1 to 100 without dropping back.

Model for the Development of Numeral Identification

Level 0: Emergent Numeral Identification. Cannot identify some or all of the numerals in the range 1-10.

Level 1: Numerals to 10. Can identify numerals in the range 1-10.

Level 2: Numerals to 20. Can identify numerals in the range 1-20.

Level 3: Numerals to 100. Can identify one and two digit numerals.

Level 4: Numerals to 1000. Can identify one, two and three digit numerals.

Model for the Development of Base-Ten Arithmetical Strategies

Level 1: Initial Concept of Ten. The child does not see ten as a unit of any kind. The child's focus is on the individual items that make up the ten. A necessary condition for attaining Level 1 is attainment of at least Stage 3 in the Stages of Early Arithmetical Learning.

Level 2: Intermediate Concept of Ten. Ten is seen as a unit composed of ten ones. The child is dependent on re-presentations* of units of ten such as hidden ten-strips or open hands of ten fingers. The child can perform addition and subtraction tasks involving tens where these are presented with materials such as covered units of tens and ones. The child cannot solve addition and subtraction tasks involving tens and ones when presented as written number sentences.

Level 3: Facile Concept of Ten. The child can solve addition and subtraction tasks involving tens and ones without using materials or re-presentations of materials. The child can solve written number sentences involving tens and ones by adding or subtracting units of ten and ones.

* A re-presentation can be thought of as a mental replay of a prior experience (ie in reflection) that is distinct from and separated in time from the experience itself.

Appendix D

COUNT ME IN TOO EVALUATION PROJECT
Assessment Summary

Teacher Code: _____ **School Code:** _____ **Date:** _____

Student	Age at Initial Interview	Early Arithmetical Strategies (Stages 0-4)	FNWS (Levels 0-5)	BNWS (Levels 0-5)	Numerical Identification (Levels 0-4)	Base 10 (Levels 1-5)
1. Female	N/a					
2. Female	N/a					
3. Female	N/a					
4. Male	N/a					
5. Male	N/a					