Mathematics Teacher Education and Development Vol. 17.1, 62-75

Discrepancies between standardised testing and teacher judgements in an Australian primary school context

Colin Carmichael
University of Southern Queensland

Received: 7 August, 2014 / Revised: 4 September, 2015 / Accepted: 29 September, 2015

This study compares the judgments that teachers make on their students' mathematics achievement with results taken from Australia's National Assessment Program: Literacy and Numeracy (NAPLAN). Using a sample of 2144 students, drawn from the Longitudinal Study of Australian Children (LSAC), the study develops two regression models: one with teacher ratings of achievement as the outcome variable, and the other with NAPLAN numeracy results as the outcome. A number of individual and environmental factors are then regressed onto these outcome variables, and the magnitudes of their effects are compared. The results indicate a consistency between teachers' judgements and NAPLAN test results, except for students with special needs, where a significant discrepancy exists. Implications of these results are discussed.

Keywords: Assessment · NAPLAN · primary school mathematics · teacher bias

Introduction

Assessment is an essential component of education as it informs important decisions about children's learning. Timely and accurate feedback, for example, is known to have one of the greatest effects on learning achievement (Hattie, 2009). Further, information on children's achievement is used to direct scarce resources in the school setting, identifying students who, for example, possess learning disabilities or instead are gifted and talented. It is imperative that such information is accurate, as incorrect under-assessments of children's achievements can contribute to a "Pygmalion effect" (Jussim & Harber, 2005) resulting in long-term deleterious effects. In the school setting, achievement data can come from: teachers' judgements of their student's achievements; and, from standardised tests such as Australia's National Assessment Program: Literacy and Numeracy (NAPLAN). This study seeks to identify discrepancies between primary school teachers' judgements of their students' achievements and their students' results in NAPLAN numeracy tests, because such discrepancies signal potential problems with one or both methods of assessment.

Given their access to a range of authentic assessment strategies, their deep knowledge of the child, and the naturalistic setting in which classroom assessment occurs, teachers should be well placed to make valid judgements on the achievement of their students. External evidence for the validity of these judgements can be obtained through a comparison with students' performances in a standardised test (Messick, 1995). In their meta-analysis of 16 studies comparing teachers' judgements with students' performances in standardised tests, Hoge and Coladarci (1989) reported a median correlation coefficient of 0.66 concluding that there were high levels of validity associated with these judgements. They noted, however, that despite this moderate association there were some studies reporting considerably less agreement between teacher judgements and student performances. Whereas some of these low associations might be explained by measurement error in tests, there is concern that some result from poor classroom assessment practices (Black & Wiliam, 1998). In this regard, Bates and Nettelbeck (2001) examined teachers' abilities to assess the reading age of children. They reported a moderate association between primary school teachers' predictions of their students'

reading accuracy and students' actual performance but found that, in terms of absolute accuracy, up to 75% of teachers misjudged reading age by more than 6 months. Poor assessment practices include teacher bias, with Kenealy, Frude, and Shaw (1991) reporting that teachers' judgements of their students' achievements were moderately correlated with their judgements of these students' attractiveness, raising the possibility that teacher judgements are influenced by non-cognitive characteristics of the child. In a more recent study, Hay and Macdonald (2008) reported that instead of using pre-specified performance criteria for the assessment of their students, the physical education teachers in their study tended to rely on intuition and were guided by affective characteristics of the students. In response to concerns regarding teacher bias, a number of U.S. based studies have utilised large longitudinal data-sets to investigate factors that might indicate biased teacher judgements (see for example, Hinnant, O'Brien, & Ghazarian, 2009; Martínez, Stecher, & Borko, 2009; Ready & Wright, 2011).

Given the doubts, raised in the last paragraph concerning poor teacher assessment practices, including bias in teacher judgements, it is of no surprise that considerable research in Australia has focussed on improving assessment practices in the mathematics classroom (see for example the review by Lowrie, Greenlees, & Logan, 2012). Fewer studies, however, have sought to gather external evidence for the validity of these assessment practices. Bobis (2009), for example, conducted a small study of three primary schools to evaluate, in part, teachers' judgements of their students' mathematical development using the Learning Framework in Number (LFIN) (Wright, 1994). No studies noted, however, have examined assessment practices and in particular teacher judgements on a large scale in Australia.

As envisaged by Lowrie et al. (2012), the current study utilises affordances associated with the availability of NAPLAN data to investigate Australian teachers' judgements at the macro-level. Foremost amongst these affordances is the alignment between NAPLAN test items and the Australian National Curriculum: Mathematics (Australian Curriculum Assessment and Reporting Authority [ACARA], 2013), the latter guiding teaching in most Australian primary school classrooms. The notion of teacher bias discussed in the U.S. based studies is not emphasised in this study, as some authors question the validity of the NAPLAN numeracy test and its items (Greenlees, 2010; Perso, 2009), instead the focus is on *systematic* discrepancies between teacher judgements and those obtained from the NAPLAN numeracy measures, hereafter termed "teacher/test discrepancies" .

Theoretical background

Teacher/test discrepancies will emerge when teachers' judgements of their students' achievements differ to these students' performances in tests in some systematic way. Südkamp, Kaiser, and Möller (2012) argued that these judgements are influenced by characteristics of the students, the teachers, the test itself, and the type of judgement required of the teacher. Obviously students' knowledge of mathematics, for example, will influence their performance in a mathematics test, but other factors such as their sex may also impact on this performance causing systematic discrepancies with teacher judgements. Similarly, a more experienced teacher might make a different judgement on a student's achievement than a less experienced colleague. The test itself may contribute to teacher/test discrepancies with the possibility that some children will perform differently on tests containing short-response items than those containing long-response items. Finally a normative judgement on a child's achievement will differ from a criterion based judgement.

Teacher judgements and student test performances do not occur in isolation and broader features related to the child's environment should be considered. Bronfenbrenner (1977) has argued that this environment consists of five nested systems of interaction ranging from the proximal "microsystem" – a system including the child's immediate environment – through to the more distal "chronosystem", which includes the broad historical circumstances of the child's life course. In line with this approach,

it is argued that influences of parents and the broader community may impact differentially on both children's performances and teachers' judgements thus generating teacher/test discrepancies. Carmichael, MacDonald, and McFarland-Piazza (2014) for example, reported that whereas parents' reports of their involvement with homework did not impact on children's mathematics achievement, teachers' perceptions of the parent's involvement did. It is possible that teachers' judgements of achievement are influenced by their perceptions of parental involvement to a greater extent than is evidenced by test performance.

Review of the literature

In this review the literature is explored to identify factors that are associated with teacher/test discrepancies. As described above, the review focuses on student, teacher, and environmental factors. It also discusses characteristics of the test that students take and the judgements that teachers make.

Student characteristics

Other than the true ability of the child, the literature suggests that teacher/test discrepancies are associated with the child's sex, ethnicity, special-needs status and behaviour.

Sex. Several studies have reported that teacher/test discrepancies are associated with the sex of the child (Hinnant et al., 2009; Martínez et al., 2009; Ready & Wright, 2011). Martínez et al. (2009), for example, found that boys outperformed girls in primary school mathematics (Years 1, 3, and 5) irrespective of whether performance was based on teacher judgements or standardized test results, but that the influence of sex was greatest in standardised test results. In other words, teachers perceived a much smaller effect than that detected by the tests. Hinnant et al. (2009) reported similar results but noted that teacher/test discrepancies were also associated with measures of the child's social competence, suggesting that teachers' judgements of children's mathematics achievement are influenced by factors related to the development of the child in general.

Ethnicity. Studies have also suggested that the ethnic background of a child is associated with teacher/test discrepancies (Bergin, 1999; Hinnant et al., 2009; Martínez et al., 2009; Rubie-Davies, Hattie, & Hamilton, 2006). Martínez et al. (2009), for example, reported that teacher/test discrepancies in mathematics were only associated with ethnicity later in primary school when children were in Year 5. Children from minority groups reported on average lower marks in standardised tests than those received from teacher judgements suggesting that teachers over-estimated their mathematics ability. Hinnant et al. (2009) also reported that teacher/test discrepancies emerged in Year 5, but that teachers underestimated the mathematical achievement of students from ethnic minorities. This difference in direction may reflect differences in the instruments used by the researchers and points to the importance of judgement and test characteristics when analysing teacher/test discrepancies (Südkamp et al, 2013).

Special needs. A number of studies have examined the accuracy of teachers as they assess children who require special needs (Bennett, Gottesman, Rock, & Cerullo, 1993; Helwig & Tindal, 2003; Martínez et al., 2009; Ritter, 1989; Sideridis, Antoniou, & Padeliadu, 2008) because this usually involves the allocation of additional scarce educational resources. As far as the accuracy of these assessments, Helwig and Tindal (2003) reported that teacher recommendations were little better than random. Of these studies, only Martínez et al. (2009) have reported on the occurrence of teacher/test discrepancies, noting that teachers tended to under-estimate, with respect to test scores, the achievement of children who were known to have a disability, though they failed to clearly define what was meant by disability.

Perceived behaviour of the child. Bennett et al. (1993) reported that teachers' perceptions of students' behaviours were strongly predictive of their academic judgements. Further, that teachers consistently perceived boys to be worse behaved in early primary than girls, resulting in lower expectations for boys. Van Houtte and Demanet (2013), however, reported no association between students' behaviour and teachers judgements of their achievement, though the students in their study were much older and the behavioural perceptions were made by the students rather than the teacher. Teachers' judgements may also be influenced by their perceived motivation of the students. Kaiser, Retelsdorf, Südkamp, and Möller (2013), using a simulated classroom situation with pre-service teachers, were able to demonstrate that the perceived motivation of children can influence teacher judgements of assessment, even when there was no association between the motivation and achievement of the children.

Teacher characteristics

Experience and qualifications. Teachers' experience and/or qualifications may influence the accuracy of their judgements and thus lead to teacher/test discrepancies. Mashburn and Henry (2004), for example, found that Kindergarten teachers with higher educational qualifications were more likely to accurately assess the school readiness of children than their less qualified peers. Ready and Wright (2011) reported that increased teacher education was associated with over-estimation of literacy achievement for children from minority groups. Martínez et al. (2009), however, reported that elementary qualifications in mathematics were not associated with teacher/test discrepancies but that teacher's age was, in that judgements of achievement from older teachers were more likely to agree with test outcomes than judgements from younger teachers. Ready and Wright (2010), also found that younger teachers tended to over-estimate, with respect to test scores, the literacy achievement of their students. Sideridis et al. (2008), however, in their study of Greek teachers did not find any association between teacher age/experience and the accuracy in which the teachers diagnosed learning disabilities, indicating that cultural or systemic variations may occur.

Teacher sex. Sideridis et al. (2008) reported that male teachers were more likely to incorrectly identify a student as having a learning disability than their female peers, suggesting that female teachers were more accurate in making these assessments. Similarly, Ritter (1989) found that male teachers were less accurate in identifying behavioural problems than females. Given that perceived behavioural problems were associated with teacher/test discrepancies, a teacher's sex may influence these discrepancies through differentiation in the labelling of poorly behaved children.

Environmental characteristics

Parental involvement. There is no research, to date, that investigates the influence that parental involvement might have on teacher/test discrepancies.

Income level of home. U.S. based studies suggest that teachers over-estimate, relative to test performance, the achievement of children from low socio-economic backgrounds in both mathematics (Martinez et al., 2009) and literacy (Ready & Wright, 2011), though these findings were not supported by Hinnant et al. (2009).

Test and judgement characteristics

Eckert, Dunn, Codding, Begeny, and Kleinmann (2006) reported that teachers were more accurate in reading as opposed to mathematics and results reported in Hinnant et al. (2009) suggest that teacher/test discrepancies are influenced by the subject matter in question. Begeny, Eckert, Montarello, and Storie (2008) noted that teachers were able to assess mastery quite well, but not partial mastery. Consequently a proficient/non-proficient judgement might be quite accurate than one

requiring a grading. This finding may have influenced reports that teachers are more accurate with higher achieving students (Demaray & Elliot, 1998), though Feinberg and Shapiro (2009) have disputed this, instead arguing that there is a greater variability in the lower distribution.

Research questions

In view of the earlier discussion, the study aims to answer the following research questions.

1. How closely correlated are Australian teachers' ratings of their students' mathematics achievement with these students' performance in the NAPLAN numeracy test?
2. To what extent are teacher/test discrepancies associated with student and teacher characteristics, and environmental factors?
3. How do test and judgement characteristics influence these teacher/test discrepancies?

Methodology

The current study examined data obtained from the Kindergarten cohort of children in the fourth wave of LSAC, which utilizes a cross-sequential design to follow two cohorts of approximately 5000 Australian children (see Sanson et al., 2002). Data collection for this wave occurred between March 2010 and February 2011, when the children were aged between 10 and 11 years. Data came from interviews with the primary parent (95% female), and questionnaires sent to the child's teacher. Children's results in NAPLAN numeracy tests were available, provided their parents had given approval. The NAPLAN numeracy tests were conducted when the children were in Year 5 at school, however due to the age range of the children and state differences in school commencement ages, the children in this cohort completed their NAPLAN tests in May 2009, 2010, and 2011.

Participants

A total of 4169 children from the original 4983 initially recruited, were still participating in the LSAC study during the fourth wave. The current study focuses on those students for whom Year 5 NAPLAN numeracy results and teacher ratings of achievement were both available. Whereas NAPLAN numeracy data were available for 3915 of these students, only 2805 of these did their test in 2010. Moreover, of these students matching teacher data were only available for 2144 children, the final sample in this study. The teacher response rate for this wave was 80%, though a small proportion of parents (3.4%) failed to give permission for researchers to contact their children's teachers.

Summary statistics for key demographic variables in the original and reduced samples are shown in Table 1 and suggest that in most respects the study sample is as representative as the original LSAC sample. A detailed analysis of possible bias caused by the sub-sampling is reported in the results.

Table 1
Key variables for original and study samples

Variable	Original sample	Study sample
Sex of child (% male)	51.1	50.7
Indigenous status (% yes)	2.8	2.7
Non English speaking background (%)	8.2	7.5
Teacher recorded special needs (%)	13.0	12.3
Parent recorded disability	5.7	5.1
Socio-economic position (Mean/SD)	(0.01/0.77)	(0.03/0.75)
Sex of teacher (% male)	28.5	28.4
Teacher post-grad qualification (%)	7.0	7.1
Number of participants	4169	2144

Outcome measures

NAPLAN numeracy score. The numeracy score is derived from children's performance in the 2010 NAPLAN Year 5 numeracy test, which contained 40 short response items that sampled mathematical content from Australia's National Curriculum. Actual test items are no longer published, however, the format of the test is shown through a published practice test (ACARA, 2012a). Using a Rasch measurement model, children's results are converted to a single score between 0 and 1000 that is comparable across years. NAPLAN numeracy results for the study sample (M = 508, SD= 75) were slightly higher than those reported in the population (M = 489, SD = 70) (ACARA, 2010).

Teacher judgments. Teachers were asked to rate their students' mathematical performance across ten items adapted from the Academic Rating Scale (ARS) (National Center for Educational Statistics, nd). When making their judgments, teachers were asked to compare the child with others of the same age. Items sampled two of the three content strands from the Australian National Curriculum: No items addressed content from the Statistics and Probability strand. Each item provided the teacher with a clear example of the type of mathematics involved, for example "Uses strategies to multiply and divide (e.g. calculates 5 lengths of 3.25 metres; or divides by 4 to determine 25% of 32)". Teachers then rated the child on a five-point scale, ranging from 1 (*Not yet*), through to 5 (*Proficient*). The scale also included a "Not applicable" category, which in this study was treated as a missing value. An exploratory factor analysis for this sample reported one clear factor explaining 78% of the variance and yielding a good reliability (a = 0.97).

Controlling for time. Teacher assessment data were collected from April 2010 through to February 2011. Almost all (98.3%), however, were collected in 2010 and all but three of these occurred at the same time or after the NAPLAN test in May 2010. The length of time (in months) separating the teacher assessment and the NAPLAN test was therefore considered in the analysis, as research suggests that children's normal growth from one school year to the next is on average equivalent to an effect size of 0.4 (Hattie, 2009).

Predictor variables/measures

Student characteristics. Children's ethnicity was assessed in two ways, whether they came from non-English speaking backgrounds (NESB) and whether or not they were Indigenous. In regards to special needs, two items were also considered. One asked the parents "Does the family member have a condition or disability that has lasted, or is likely to last, for 6 months or more?" Parents were also

asked to note the nature of the disability, with only 18 of the 106 responses indicating it was a learning disability. The second item asked the teachers "Does this child receive any specialised services provided within the school because of a diagnosed disability or additional need?" Teachers were also asked to identify the main reason why children received these services. Children receiving the service for gifted and talented (n = 34) were not included in this factor. Of the remaining 262 children, more than half (52%) had learning problems in maths or reading, and approximately one fifth (21%) emotional or behavioural problems.

Children's behaviour was assessed using the Conflict Subscale from the Student Teacher Relationship Scale (STRS) (Pianta, 2001). This subscale consisted of seven items, such as "this child and I always seem to be struggling with each other (i.e. having a hard time getting along)" that teachers answered on a 5-point Likert scale that ranged from 1 (*Definitely does not apply*) to 5 (*Definitely applies*). Reliability estimates (a = 0.9) for this sample were good.

Teacher characteristics. In relation to their experience, teachers were asked "How many years teaching experience do you have as a teacher at this grade level?" Results for this sample ranged from 0 through to 38 years (M = 6 years). Teachers were also asked for their highest educational qualification. In this sample 7.1% of respondents indicated they had Masters or Doctoral degrees.

Environmental characteristics. Parental involvement was assessed through teachers' responses to the question "In your opinion, how involved are this child's parents in his/her learning and education?" with response categories "very involved", "somewhat involved", "not involved", and "not known". The majority of parents (54.8%) were perceived to be very involved in their child's education. Teachers were also asked about parents' involvement in the school. In this sample most (77.3%) parents had visited the classroom, but few (22.6%) had volunteered to help out on an excursion. In regards to the family's socio-economic status, the socio-economic position (SEP) (Blakemore, Strazdins, & Gibbings, 2009) was an index constructed from data in this study that is based on family income and the educational background of parents.

Judgment/test characteristics. Teachers' judgements of mathematics achievement from the ARS items discussed earlier were relative to other children that age. No criteria were available for the teachers to make absolute ratings. In regards to test characteristics, however, teachers' assessments of children's literacy were also considered. This was gauged using nine items adapted from the ARS, such as "conveys ideas clearly when speaking (e.g. presents an oral report from an outline that is logically organised, supports ideas with specific details, and presents a simple argument)" that were answered using the same scale as described for mathematics. For this sample the items loaded onto one factor explaining 77% of the variance and reported good reliability (a = 0.96). In addition to this, children's NAPLAN reading, writing, grammar and spelling scores were used to construct a composite literacy test score. An exploratory factor analysis indicated that these scores loaded onto one factor explaining 73% of the variance and reporting a reliability of 0.87. Sample items are available from ACARA (2012b).

Method of analysis

In order to determine the extent to which teacher/test discrepancies were associated with student, teacher and environmental characteristics, two regression models were tested. In the first, NAPLAN numeracy scores were regressed onto each of the factors listed earlier. The second model regressed ARS teacher responses onto each of the identified factors, but after firstly controlling for the influence of the time separating the dates of the two measurements. Structural equation modelling procedures were used so that the influence of measurement error could be minimised. Unfortunately this only occurred for the teacher judgment measures, because error data associated with the NAPLAN measures were not released by the relevant authorities. All analyses was undertaken using M-Plus (Muthén & Muthén, 2011). Due to the complex sampling design used in LSAC, stratification and

clustering effects were modelled using the type=COMPLEX option in M-Plus (see Asparouhov & Muthen, 2006, for details). Weights were provided to account for the probability of unequal inclusion in the sample (Daraganova & Sipthorp, 2011) and were modelled using procedures described in Muthén and Muthén (2011).

The effect of test/judgement characteristics was investigated by undertaking a similar analysis to the above, but instead using the NAPLAN composite literacy scores and ARS literacy scores. The analysis also addresses the issue of bias that may have occurred through the sub-sampling process used in the study.

Results

Teachers' judgments of these students' mathematics achievement were moderately correlated with the students' NAPLAN numeracy scores (r = 0.61) and this result is very similar to the median of 0.66 reported by Hoge and Coladarci (1989).

Both outcome measures were initially regressed onto time, which significantly (at the 5% level) predicted teacher judgements of their students' mathematics achievement (β = 0.03). Each of the identified factors associated with the student, teacher and environmental characteristics were then included with time as predictors of teacher's judgements (ARS model) and without time for test scores (NAPLAN model). Significant standardised effects are shown in Table 2, which also shows the standard errors associated with these effects, the difference in effects, the standard error associated with this difference, and its associated t-value. Factors reporting significantly different standardised effects depending on the outcome measure (ARS or NAPLAN), were flagged as being associated with teacher/test discrepancies. Fit statistics for all models were within acceptable bounds (Byrne, 2001), in that *RMSEA* < 0.08 and *CFI* > 0.95.

As is seen from the table, a number of identified factors predicted mathematics achievement as measured by teacher judgements and NAPLAN numeracy tests. Children identifying as Indigenous, for example, achieved significantly lower than their peers, irrespective of the assessment type. Interestingly, teacher sex impacted positively on both measures of mathematics. Given earlier research suggesting male teachers were less accurate this could explain the positive effect on ARS, but the significant positive effect on NAPLAN is surprising. Only one of the factors tested, however, was associated with a significant teacher/test discrepancy. Teachers tended to underestimate, compared with test results, the mathematical ability of children who they knew received specialised services. The effect of being identified as receiving specialised services on the ARS outcome ($\beta = -1.00$) was significantly lower than its effect on the NAPLAN numeracy score ($\beta = -0.76$). Interestingly, parent reported disability status was associated with NAPLAN numeracy performance but not with teacher judgements, though the discrepancy between them was not significant.

In order to assess the influence of test/judgment characteristics on teacher/test discrepancies, the above analyses were repeated using ARS literacy and the NAPLAN composite literacy score as outcome variables. The latter was strongly associated with the ARS measure (r = 0.79). Again, time was a significant predictor of teacher ratings (β = 0.04). The results for literacy, not reported here, were very similar to those for mathematics in that the key variables such as Indigeneity predicted literacy achievement irrespective of assessment method. The only teacher/test discrepancy occurred for SEP. Its effect on the ARS model was 0.27, whereas its effect on the NAPLAN literacy score was 0.37, the difference in these effects was statistically significant ($t = 3.3$). Fit statistics for the literacy models were all within acceptable bounds, in that *RMSEA* < 0.10 and *CFI* > 0.90.

Table 2:
Standardised effects on Teacher and NAPLAN estimates of mathematics achievement

	ARS model[1]		NAPLAN model		Difference		
Predictor	β_1	se(β_1)	β_2	se(β_2)	$\beta_2 - \beta_1$	se(dif)	*t*
Student characteristics							
Behaviour	-0.24*	0.03	-0.18*	0.03	0.06	0.04	1.50
Disability (parent report)	-0.04	0.09	-0.27*	0.11	-0.23	0.14	1.62
Disability (teacher report)	-1.00*	0.08	-0.76*	0.07	0.24	0.11	2.18*
Indigenous	-0.66*	0.15	-0.84*	0.15	-0.18	0.21	0.85
Teacher characteristics							
Sex (male)	0.22*	0.05	0.14*	0.05	-0.08	0.07	1.13
Post-grad	0.27*	0.08	0.23*	0.09	-0.04	0.12	0.33
Environmental characteristics							
SEP	0.35*	0.03	0.34*	0.02	-0.01	0.04	0.25
Parent very involved	0.50*	0.05	0.41*	0.05	-0.09	0.07	1.29
Parent attend excursion	0.20*	0.06	0.18*	0.06	-0.02	0.08	0.25

1. After controlling for time. * Significant at the 5% level

Analysis of sample bias

In order to assess the effects of possible bias from using a subset of the Wave 4 LSAC data, a number of key variables were compared across the study sample (n = 2144) and the excluded sample (n = 2025). In particular chi-square tests of association were used to analyse relationships between inclusion and exclusion in the study and key categorical variables and a comparison of means was undertaken for continuous variables.

Of the categorical variables tested, including student sex, teacher sex, teacher qualifications, disability status, Indigeneity, Non-English speaking background status, parental involvement and school sector, only the latter two were significantly associated with sample. Children included in the study were more likely to attend non-government schools (χ^2 = 4.8) and to have parents who were very involved (χ^2 = 6.4) than those excluded. Possibly more parents with low perceived involvement in their children's education failed to provide permission for their children's teacher to be contacted, thus contributing to this small bias. In relation to continuous variables, teachers in the study sample were more likely to have taught longer than those in the excluded sample (ΔM = 6 months, t = 2.2), though data for only one half the excluded sample were available. Children in the study sample were also more likely to come from wealthier homes (ΔM = 0.08, t = 3.4) than those in the excluded sample.

Discussion

The study results indicate that Australian primary teachers' assessments of their children's mathematics achievement tend to agree with equivalent assessments obtained from the NAPLAN numeracy test, in that there was a moderate correlation between the two measures. The results suggest, however, that this agreement is much stronger for measures of children's literacy than for mathematics. The weaker correlation between teacher judgements and test scores in mathematics may simply be a result of the teacher scale not containing any statistics and probability items. It may also be that primary school teachers are not as accurate in assessing mathematics as they in assessing literacy, if agreement with NAPLAN results is taken as the benchmark.

Few of the student and teacher characteristics identified in the literature were associated with teacher/test discrepancies in mathematics. Both models detected a bias in favour of boys, but this was not statistically significant in either. Similarly, the NESB status of children did not impact on their mathematics achievement. Indigeneity and poor behaviour were both negatively associated with mathematics achievement irrespective of the type of measure used. Teachers' qualifications were positively related to both outcomes as was teacher sex, but again not differentially. Surprisingly, teachers' experience did not predict either outcome measure.

The major finding of the study is that children with noted specialised needs performed worse on teacher assessments of mathematics than in standardised tests. In terms of effects, the noted difference of 0.24 suggests that teachers perceived these students to be, on average, six months in development behind (in terms of Hattie's average effect of 0.4) their level as indicated by NAPLAN test scores. A similar result was detected in the U.S. study undertaken by Martinez et al. (2009), though their effect difference was 0.1. It is unlikely that the noted deficiencies of the ARS mathematics instrument would contribute to this discrepancy, in that children with learning difficulties are unlikely to perform better (or worse) on items sampling statistics and probability. Instead the finding appears to confirm reports that teachers have difficulty judging the performance of children with lower mathematics achievement (Demaray & Elliot, 1998; Begeny et al. 2008) and/or that they struggle with judging partial mastery (Begeny et al., 2008). Surprisingly, however, no such teacher/test discrepancy was observed for literacy, suggesting again that these teachers were more adept at making judgements on literacy rather than mathematics achievement.

In regards to environmental factors, the socio-economic position (SEP) of the child's family impacted significantly on all measures, but was only associated with teacher/test discrepancies for literacy. This apparent discrepancy is likely a manifestation of the class-level comparisons that teachers were making, and the finding that SEP influences teacher judgements at the class level (Ready & Wright, 2011). A child from a high SEP background is likely to gain a higher NAPLAN literacy score than one from a low SEP background. This child, however, is likely to be in a classroom of children with similar backgrounds. A teacher's judgement of the relative achievement of this child should note a smaller effect compared to others in this high SEP classroom. That SEP was not associated with a teacher/test discrepancy in mathematics, however, is surprising, again suggesting teachers' judgements of children's mathematics achievement may be less valid than their judgements of literacy achievement.

Taken together the results suggest that primary school teachers' would benefit from more professional development related to the mathematical assessment of children and in particular those with lower abilities or demonstrating only partial mastery. A review of current state education web-sites, reveal a plethora of documents providing good, but general, advice on assessment. Some sites also point teachers to excellent developmental frameworks that can and are used to assess children's mathematics achievement (e.g. LFIN). Whereas web-based resources are useful, active professional development in the use of these frameworks is far more beneficial (see for example, Bobis, 2009).

Methodological considerations and limitations

The study has sought to address some of the weaknesses in studies undertaken in the U.S. (Hinnant et al., 2009; Martínez et al., 2009; Ready & Wright, 2011). In the first instance the measurement error associated with the ARS was modelled in the study. Whereas this error is often ignored because of large sample sizes, it is known to be considerably larger in the tails of achievement distributions (Wu, 2010) and of consequence when examining students with, for example, specialised needs. It was not possible, however, to model the measurement error associated with NAPLAN results, though the tests upon which they are based contain many more items than the ARS and are thus likely to report considerably lower levels of measurement error (Wu, 2010).

The study has also addressed the issue of bias caused when sub-samples of large designed samples are used. As reported, parents in the study who were more involved in their child's education were also more likely to give permission for the use of their child's achievement data than parents who were not as involved. This bias may have contributed to the result that parental involvement was not associated with teacher/test discrepancies. In addition to this, more experienced teachers were likely to be in the study sample and thus correctly returned their surveys. Such teachers, however, are more likely to make judgements that agree with standardised tests (Martinez, et al. 2009; Ready & Wright, 2011). This bias may have contributed to the non-significant association between teacher experience and teacher/test discrepancies.

In the study teacher/test discrepancies have been used to highlight potential bias in teachers' judgements of their students' mathematical achievements. These discrepancies, however, could equally point to problems (validity threats) with the NAPLAN tests. For this reason, the study also considered teacher/test discrepancies in literacy. Different results from these parallel analyses strengthen the argument that these teachers had difficulty judging mathematics achievement, conclusions reported elsewhere in the literature (Eckert et al., 2006; Hinnant et al., 2009).

Secondary data sets such as LSAC are an important research resource permitting analysis at the macro-level. With this use, however, is the limitation that questions cannot be modified to suit the particular research question. In this regard, the ARS mathematics scale did not contain any items sampling the statistics and probability strand of the national curriculum and this deficiency could have contributed to the study's findings.

Conclusion

In terms of teacher/test discrepancies, primary school teachers in Australia appear to score higher on the mathematics assessment report card than their international colleagues. The study identified only one factor associated with such a discrepancy, students reported by their teachers as requiring specialised services were judged to achieve lower in mathematics than their NAPLAN numeracy scores would suggest. This discrepancy is disturbing as it suggests that labelling could occur when children are referred to specialised support services. That it did not occur in literacy, however, indicates that rather than labelling children, Australian primary school teachers may be less accurate in making judgements on their students' mathematical achievements than on their literacy achievements, especially for those students who have gained partial mastery.

Several years ago, the Australian Association of Mathematics Teachers (AAMT) reported that professional development in assessment is regularly identified as a priority by teachers and schools (AAMT, 2008). Perhaps it is now time to act on this information.

Acknowledgements

The author would like to thank all involved in the LSAC study. *Growing Up in Australia* was initiated and funded as part of the Australian Government Stronger Families and Communities Strategy by the Australian Government Department of Housing, Families, Community Services and Indigenous Affairs (FaHCSIA). The study is being undertaken in partnership with the Australian Institute of Family Studies, with advice being provided by a consortium of leading researchers at research institutions and universities throughout Australia. The data collection is undertaken for the Institute by the Australian Bureau of Statistics. All views expressed in this paper are the authors', and do not represent the views of FaHCSIA or the Australian Institute of Family Studies.

References

Asparouhov, T., & Muthén, B. (2006). *Multilevel modeling of complex survey data*. Paper presented at the Joint Statistical Meeting. ASA Section on Survey, Seattle. http://www.amstat.org/sections/SRMS/Proceedings/y2006/Files/JSM2006-000803.pdf

Australian Association of Mathematics Teachers. (2008). *Position paper on the practice of assessing mathematics learning*. Adelaide: AAMT.

Australian Curriculum Assessment and Reporting Authority. (2010). *NAPLAN Achievement in reading, writing, language conventions and numeracy: National report of 2010*. Canberra: ACARA. Available from: http://www.nap.edu.au/results-and-reports/national-reports.html

Australian Curriculum Assessment and Reporting Authority. (2012a). *NAPLAN Numeracy Example test: Year 5*. Canberra: ACARA. Retrieved from: http://www.nap.edu.au/verve/_resources/Example_Test_Numeracy_Y5.pdf

Australian Curriculum Assessment and Reporting Authority. (2012b). *NAPLAN Reading Example test: Year 5*. Canberra: ACARA. Retrieved from: http://www.nap.edu.au/verve/_resources/Example_Test_Reading_Y5.pdf

Australian Curriculum Assessment and Reporting Authority. (2013). *Australian Curriculum: Mathematics (Version 5.1)*. Retrieved from http://www.australiancurriculum.edu.au/mathematics.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177-187.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgements and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*(1), 45-55.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behaviour perceptions and gender on teacher's judgements of students' academic skill. *Journal of Educational Psychology, 85*(2), 347-356.

Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist, 34*(2), 87-98.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Blakemore, T. J., Strazdins, L., & Gibbings, J. (2009). Measuring family socio-economic position. *Australian Social Policy, 8*, 121-169.

Bobis, J. (2009). *The Learning Framework in Number and its impact on teacher knowledge and pedagogy*. Sydney: NSW DET Retrieved from http://www.curriculumsupport.education.nsw.gov.au/primary/mathematics/assets/pdf/cmitreport08.pdf.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513-531.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS*. Mahwah, NJ: Lawrence Erlbaum Associates.

Carmichael, C. S., MacDonald, A., & McFarland-Piazza, L. (2014). Predictors of numeracy performance in national testing programs: Insights from the Longitudinal Study of Australian Children. *British Educational Research Journal, 40*(4), 637-659.

Daraganova, G., & Sipthorp, M. (2011). LSAC Technical paper no.9: Wave 4 weights. Melbourne: Australian Institute of Family Studies.

Demaray, M. K., & Elliot, S. N. (1998). Teachers' judgements of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*(1), 8-24.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*(3), 247-265.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgements of students' reading with differing achievement levels. *The Journal of Educational Research, 102*(6), 453-462.

Greenlees, J. (2010). The terminology of mathematics assessment. In L. Sparrow, B. Kissane, & C. Hurst (Eds.), *Shaping the future of mathematics education: Proceedings of the 33rd Annual Conference of the Mathematics Education Research Group of Australasia.* (pp. 218-224). Adelaide: MERGA, Inc.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, UK: Routledge.

Hay, P. J., & Macdonald, D. (2008). (Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject. *Assessment in Education: Principles, Policy & Practice, 15*(2), 153-168.

Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children, 69*(2), 211-225.

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*(3), 662-670.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgements of academic achievement: A review of the literature. *Review of Educational Research, 59*(3), 297-313.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131-155.

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgements. *Learning and Instruction, 28*, 73-84.

Kenealy, P., Frude, N., & Shaw, W. (1991). Teacher expectations as predictors of academic success. *The Journal of Social Psychology, 131*(2), 305-306.

Lowrie, T., Greenlees, J., & Logan, T. (2012). Assessment beyond all: The changing nature of assessment. In B. Perry, T. Lowrie, T. Logan, A. MacDonald, & J. Greenlees (Eds.), *Research in Mathematics Education in Australasia 2008-2011*. Rotterdam: Sense Publications.

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgements, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*(1), 78-102. 0.21

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice, 23*(4), 16-30.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist, 50*(9), 741-749.

Muthén, L. K., & Muthén, B. (2011). *Mplus User's Guide* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

National Center for Educational Statistics. (nd). Early Childhood Longitudinal Study - Kindergarten (ECLS-K). Washington, DC: Department of Education.

Perso, T. (2009). Cracking the NAPLAN code: Numeracy in action. *Australian Mathematics Teacher, 65*(4), 11-16.

Pianta, R. C. (2001). Student Teacher Relationship Scale (STRS). Lutz, FA: Psychological Assessment Resources.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335-360.

Ritter, D. R. (1989). Teachers' perceptions of problem behaviour in general and special education. *Exceptional Children, 55*(6), 559-564.

Rubie-Davies, C., Hattie, J. A. C., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Pscyhology, 76*(3), 429-444.

Sanson, A., Nicholson, J., Ungerer, J., Zubrick, S., Wilson, K., Ainley, J. & Wake, M. (2002). Introducing the Longitudinal Study of Australian Children *(LSAC Discussion Paper No. 1)*. Melbourne: Australian Institute of Family Studies.

Sideridis, G. D., Antoniou, F., & Padeliadu, S. (2008). Teacher biases in the identification of learning disabilities: An application of the logistic multilevel model. *Learning Disability Quarterly, 31*(4), 199-209.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743-762.

Van Houtte, M., & Demanet, J. (2013). Curriculum tracking and teacher evaluations of individual students: Selection, adjustment or labeling? *Social Psychology of Education, 16*(3), 329-352.

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

Wu, M. J. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*(4), 15-27.

Author details

Colin Carmichael, Open Access College, University of Southern Queensland, Baker St, Toowoomba, Queensland, 4350.
Colin.Carmichael@usq.edu.au