Insight into Subtraction from Large-Scale Assessment Data

Patricia Morley
Monash University
<patricia.morley@monash.edu>

The data from national tests such as the National Assessment Program Literacy and Numeracy (NAPLAN) and its precursor Victorian Achievement Improvement Monitor (AIM) are an important resource. The 2006 Year 3 AIM assessment included two subtraction items that are similar in content, and which were presented without text or images. The detailed, novel analysis of the children's responses presented here provides insight into children's fluency and understanding of these items.

Large-scale assessments have received much criticism in the press and from within the education research community in part because aggregated assessment scores are widely claimed to be a means of measuring a school's quality. This interpretation disregards factors outside the school's influence that contribute to achievement scores (Ferrari, May 3, 2011; Popham, 1999).

The Victorian Achievement Improvement Monitor (AIM) assessment was a precursor to the National Assessment Program Literacy and Numeracy (NAPLAN). The Victorian Curriculum and Assessment Authority (VCAA) stated that the purpose of the AIM assessment was to provide an "indication of how well the literacy and numeracy skills of students were developing" (Victorian Curriculum Assessment Authority, 2006c). All Victorian children in years 3, 5, 7 and 9 undertook the assessment. The details of children's responses were provided to the schools, so that they could be analysed by the school and hence provide feedback that could be used to inform decisions about the school's teaching program (Victorian Curriculum Assessment Authority, 2006a).

The data from the national tests have been acquired with considerable effort, and remain an underutilized resource. As Leder (2012) asks, "The NAPLAN national reports contain much valuable and potentially usable data. But how much of these are actually understood and used constructively?" (p. 17).

The analysis investigates the responses of Victorian Year 3 children in Year 3 to two subtraction items on the same assessment. This paper investigates the patterns that exist in the children's responses to these items, and the extent to which the format of the items, that is whether the item is multiple-choice or write-in, affects children's responses.

Literature Review

Skemp (1976) identified two types of mathematical understanding: relational understanding, which involves understanding the underlying ideas, and instrumental understanding, which involves understanding what to do to solve a mathematical problem. Popham (2009), discussing assessment in a classroom context, noted that there is an inherent limitation when trying to make inferences about what knowledge another person possesses: "we are dealing with the unseen" (p. 126), even in a classroom context, and it is not possible to determine whether instrumental or relational understanding is being assessed. The ability for this analysis to draw inferences about children's understanding from the results of large-scale assessments is inherently limited because the only source of information is the written result of the assessments. Nonetheless, there is information present in the incorrect choices that were made. Guidelines for writing multiple-choice

In V. Steinle, L. Ball & C. Bardini (Eds.), Mathematics education: Yesterday, today and tomorrow (Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australasia). Melbourne, VIC: MERGA.

items generally recommend that distractors provide plausible options to an unambiguously correct answer to avoid the correct answer being made obvious, and to a lesser extent, that distractors correspond to common errors (Haladyna, Downing, & Rodriguez, 2002). Children are approximately 8 years old at the Year 3 level, and at this age are "struggling to learn the complex arithmetic skills associated with addition and subtraction" (Romberg, Collis, & Grouws, 1987, p. 115), and their capability to solve such items changes rapidly at this stage. Common errors for children of this age on subtraction items include subtracting the smaller digit from the larger digit, regardless of which number the digit is associated with, and regrouping errors (Young & O'Shea, 1981; Fuson, 1990).

The distractors of the multiple-choice items are important sources of information in this analysis. Rather than using the assessment data to provide a measurement of children's progression, this analysis seeks to gain an insight into underlying factors that support their progress, utilising the information provided by the selection of distractors to gain an insight into their mathematical understanding.

While the two items analysed in this paper are similar in that they are subtraction which some form of decomposition, they differ in three respects: in magnitude, format and presentation. The first item, 423−106, is a t hree-digit multiple-choice item presented horizontally, while the second, 71 − 26, is a two-digit write-in item presented vertically.

Despite evidence that numbers of increasing magnitude become increasingly difficult from a cognitive perspective (see Nuerk, Moeller, Klein, Willmes, and Fischer, 2011), it does not necessarily follow that the three-digit item is conceptually more difficult than the two-digit item. For English speakers, among others, two digit numbers are problematic because of irregular linguistic representations of these numbers (Fuson et al., 1997; Ma, 1999). Fuson et al. (1997) recommended that leaving little delay between the introduction of two-digit numbers and the introduction of three-digit numbers, as the linguistic regularity and alignment with numerical representation of three-digit numbers allows children to construct a coherent model of the base 10-system more readily than if only two-digit numbers are available to them.

The third difference, between the multiple-choice format and write-in format, affords the possibility of guessing the correct answer by choosing at random, and also to correct mistakes that do not appear in the choice of distractors (Martinez, 1991). This paper takes the opportunity to compare the results of these two different subtraction items on a large population.

Methodology and Approach

The data used in this investigation was provided by the Victorian Curriculum Assessment Authority (VCAA) from the Year 3 2006 AIM assessment and is based on the anonymised responses for the 53 174 Victorian children who undertook the assessment and answered at least one item correctly. The assessment was undertaken by children simultaneously across the state within a time limit of 35 minutes for 32 items (Victorian Curriculum and Assessment Authority, 2006b). The 2006 Year 3 AIM assessment included two subtraction items of similar content, both presented without text or images. This provided the opportunity to compare the children's responses to the two items, and thereby gain insight into how children respond to the difference between the formats.

The analysis presented in this paper uses exploratory techniques to find and investigate patterns in the data and confirmatory techniques to test hypotheses that did not arise from the same data (see Behrens, 1997; Tukey 1977, 1980).

Analysis

Responses to multiple-choice item

The first item, item 21, is a multiple-choice item that presented the expression 423−106 horizontally, together with four options: 317, 327, 329 and 529. The correct response, 317, was chosen by 69% of children.

The most popular distractor for the multiple-choice subtraction item, 327, was chosen by 15% of children. It is an error that can be explained if regrouping between the units and tens was incompletely performed; that is 400 + 20 + 3 - 100 - 6 was regrouped as 400 + 10 +13 - 6 to perform the subtraction of the units digit, but then the regrouping was forgotten and the result was calculated as 400 + 20 + 13 - 6. The third option, chosen by 8%, reflects the common error of subtracting the smaller digit from the larger digit, regardless of whether the digit is in the minuend or the subtrahend. The fourth option, chosen by 6%, is the result of treating the subtraction problem as an addition problem.

Responses to write-in item

The second item, item 25, is a write-in item that presented the expression 71 − 26 vertically, with space underneath to record the result. The correct answer of 45 was given by 46% of children. There were more than 100 distinct responses to item 25, which asked for the answer to 71 − 26. The next most common answer was 55, given by 29% of children. This response corresponds to both the regrouping error found in the most common distractor chosen in the multiple choice subtraction item, and the subtraction of the smaller digit from the larger digit regardless of position, corresponding to the third option of the multiple choice item. The next most frequent response was 97, at 3%, which is what would be obtained if the children added instead of subtracted, and 50, also at 3%, which is what would be obtained as an estimate using the tens column only (avoiding the subtracting the larger digit from the smaller).

Of the children who chose the option corresponding to the regrouping error in the multiple choice item, 45% gave the answer corresponding to the same error in the write-in item. Of the children who chose the option corresponding to the subtracting the smaller digit in the multiple choice item, 45% gave the answer corresponding to the same error in the write-in item. The number of children who did not provide a response was 6% for each of the two items, an indicator that the items were initially assessed by the children to be of equal difficulty (see Martinez, 1991).

The analysis of the combined responses to the two items analysis addresses three separate issues: firstly, what the responses to the two items together can tell us about children’s understanding of the content; secondly, what responses to the other items on the assessment can tell us about the extent to which write-in items reflect content knowledge compared with multiple-choice items; and thirdly, what the responses to other items on the assessment can tell us about which factors are important to children’s ability to answer the subtraction items correctly.

Theoretical difference between multiple-choice and write-in items

For the assessments being discussed in this analysis, a write-in item, which asks the respondent to fill in a blank box, differs from a multiple-choice item in a fundamental way. The instruction associated with the write-in item, whether explicit or implied, might be phrased as ‘Write your response to the item in the blank space provided’, whereas the instruction associated with the multiple-choice item might be phrased as 'Given that one, and only one, of the multiple responses shown here is correct, select the response that you think is most likely to be correct'.

The difference between the two implied instructions leads to a different interpretation of what can be inferred from a correct response. If a write-in item response is correct, there is some confidence that the respondent was able to both interpret the item correctly and was able to perform the necessary calculation. If a multiple-choice item is answered correctly, it is less likely that the respondent was able to both interpret the question and provide the correct response, because in this case there is a possibility that the correct response was selected without a similar level of knowledge by eliminating implausible distractors, or even was selected at random. It is also possible for the respondent to have made an error that was not one of the choices offered, allowing an opportunity to self-correct that the write-in format does not provide.

When comparing two similar items where one item is a write-in item and the other is multiple-choice, there are four possible outcomes: firstly, both items are answered correctly; secondly, the multiple-choice item is answered correctly and the write-in item is answered incorrectly; thirdly, the write-in item is answered correctly and the multiple-choice item is answered incorrectly; or finally that both items are answered incorrectly. What can be inferred from each of these four outcomes is dependent on t he likely probability of whether the response to the multiple-choice option is a reliable indicator of the respondent's knowledge or not.

In the first case, a correct response to both the write-in item and the multiple-choice items is most likely to result when the respondent is fluent with the material being assessed. In this case, the probability that the multiple-choice item response was chosen at random is likely to be low.

In the second case, where the write-in item is answered correctly but the multiple-choice item is answered incorrectly, the implications are less clear. It would be surprising if this case accounted for a large proportion of respondents, but these items offer some indication of the prevalence of mistakes among children who understand the material well enough to have answered the write-in item correctly. However, the presence of a sizeable proportion of respondents in this category may also indicate that the items, although similar, may be different in some significant way.

In the third case, where the write-in response was incorrect but the multiple-choice response was correct, it is more likely that the multiple-choice response was the result of a guess, or perhaps indicates that the responses provided an opportunity for self-correction that was not available with the write-in item. However, there is not sufficient information to conclude that this response represents a guess, as it remains a possibility that the respondent has made a mistake with the write-in response. It would be interesting to examine the type of mistake made on the write-in item, and whether it corresponded to one of the multiple-choice options.

In the fourth, and final, case, where both items were answered incorrectly, it is likely that the material is not well understood by the respondents. In this case, the comparison of the item responses will give an indication of whether the same type of mistake was made on both items.

This framework for comparing students' responses to the subtraction items is used for comparing the actual responses of the two subtraction items. The facility of an item is an indicator of the relative ease with which children correctly solve mathematical problems and is measured by the proportion of children who correctly answered the problem (Hart, 1980). The assumption that answering the write-in item correctly is a better test of content knowledge can be tested by comparing the difference in the facility of each item of the assessment between the second and third cases, where one item was answered correctly and not the other. If the write-in item assesses content knowledge in the same way as the multiple-choice item, it follows that the content knowledge of those who answer the multiple-choice item correctly would be similar, and that this would be reflected to some extent in other items on the assessment, particularly those items with overlapping content knowledge. To investigate this hypothesis, the facility on all items of the assessment where the population is students who answered one of the subtraction items correctly, but not the other, are compared.

Effects of different formats

The principal difference between the two subtraction items is that one is in a multiple-choice format and the other asks children to write their response in to a box. The multiple-choice item involves three-digit numbers, including a place-holding zero in the number to be subtracted (subtrahend), while the write-in item involves two-digit numbers with no place-holding zero. Table 1 shows the actual percentages of children who responded according to the different cases.

Table 1
Subtraction Item Outcomes (n=53 174)

Item Outcomes	Multiple-choice item correct 423−106 =	Multiple-choice item incorrect
Write-in item correct 71 − 26 =	First case: 38% Likely to be fluent with item content	Second case: 8% Less likely outcome indicates rate of mistakes or that the three-digit subtraction may have differed sufficiently from the two-digit subtraction to have an impact on the results
Write-in item incorrect	Third case: 30% Possible guess with multiple-choice item or mistake with write-in item Not likely to be fluent with item content	Fourth case: 24% Not likely to be fluent with item content

In the second case, the correct response to the write-in item indicates that these children are able to solve a two-digit subtraction with regrouping, but the incorrect response to the multiple-choice three-digit subtraction item with regrouping indicates either that these children made a computational error not based on a conceptual difficulty, or that place-value concepts were not yet sufficiently in place to be able to deal with three-digit numbers.

In the third case, it seems at first glance to be a reasonable expectation that the children who answered the three-digit multiple-choice item correctly were also likely to have answered the two-digit write-in item correctly, except for those children who either made a mistake that was not represented by the distractors and revised their response, or guessed.

However, this would not account for the likelihood that some children made simple errors even though they were familiar with the concept and a method of finding a solution.

In the fourth case, in which the children answered both of these items incorrectly, 52% (6 556) children answered 55 to the write-in item. This result may be at arrived at by either of two ways: one being the same regrouping error as the most commonly chosen multiple-choice item distractor, and the other being a common error arising from taking the smaller digit from the larger digit regardless of which number is being subtracted (Young and O'Shea, 1981), which does not have a corresponding distractor in the multiple-choice item. It seems to be a reasonable inference that the majority of children who answered 55 to the write-in item would be likely to make a similar error in the multiple-choice item, if there is a corresponding distractor. Since only 54% of the children in this case who answered 55 (3 546 children) chose the multiple-choice distractor corresponding to the regrouping error, the inference is that approximately half of the children who wrote 55 as their answer to the write-in item incorrectly subtracted the smaller digit from the larger digit. The children who chose the third multiple-choice distractor (a result obtained from a combination of subtraction and addition) were likely to be having more difficulty with three-digit numbers than two-digit numbers, a situation that is reflected in the third case above to some extent.

Of the 69% of children who answered the multiple-choice item correctly, one third answered the write-in item incorrectly, 55% of whom wrote 55 which was the most common error corresponding to a regrouping error, indicating that this type of error for these children is not conceptual, but a common procedural error.

The proportion of children who answered the write-in item correctly but who did not answer the multiple-choice item correctly was 8%. If the items were so similar as to be considered identical, this proportion would reflect the rate of mistakes made by the children, would be surprisingly high error rate. However, the write-in item, 71−26, involves two-digit numbers, while the multiple-choice item, 423 − 106, involves three-digit numbers, and so this 8% includes both the rate of mistakes and the difference in difficulty between the items. This suggests that the difference between the items is modest and that most children who are fluent with two-digit numbers are also fluent with three-digit numbers.

Reliability of write-in responses compared to multiple-choice responses

Write-in responses are generally assumed to be a more accurate measure of content knowledge than multiple-choice, if only because write-in responses do not allow the same scope for guessing the correct response items (Martinez, 1991). If this assumption is true, it follows that children who answered the write-in subtraction item correctly but not the multiple-choice item were more likely to have greater content knowledge than those who answered the multiple-choice item correctly but not the write-in item correctly, also assuming that the two subtraction items are sufficiently similar to be considered to be measuring the same construct.

To investigate the assumption that write-in items reflect content knowledge more reliably than multiple-choice items, the responses of the children who had answered one subtraction item correctly to other items on the assessment were compared. The facility for each item was calculated separately for each of the groups of children who had answered one subtraction item correctly. The difference between the facility of the group that answered the multiple-choice item correctly and the facility of the group that answered the write-in item correctly was calculated for each item, providing the correlation between the format and the facilities for all of the items on the test. For most items, the facility for the group that answered the write-in item correctly was higher than the group that answered the multiple-choice item correctly, with the difference in values ranging from -2 to 6%.

A Welch 2-sample t-test, appropriate for two samples with differing sample sizes and means, showed that although there was a statistically significant difference between the groups in favour of the write-in item, the difference was small, with the 95 percent confidence interval of the difference in mean of the total assessment score between -0.61 and -0.33 [$t = -4.46$, $df = 37\,430$, $p\text{-value} < 0.001$].

Discussion

Over a third (38%) of the children answered both subtraction items correctly, while a similar proportion (38%) of children answered either one of the subtraction items correctly, but not both, confirming the earlier findings that children are learning and consolidating their understanding in this area (Romberg et al. 1987). Of the group indicating less fluency with the material than those who had answered both items correctly, there was little difference in responses to other items on the assessment. Only 8% of children answered the write-in item correctly and the multiple-choice item incorrectly, suggesting that there was a moderately small difference in difficulty between the two-digit subtraction and the three-digit subtraction, supporting the recommendation of Fuson et al. (1997) that being presented with three-digit numbers supports children's understanding of base-10 representation.

Whether the horizontal or vertical presentation was problematic for children is not apparent from the data. The Early Numeracy Research Project established key growth points in the development of children's numeracy, noting that only 10% of children are proficient in strategies for use in addition and subtraction items, and that the teaching of column-based written algorithms in the early years is therefore inappropriate. As Perso (2011) pointed out, the NAPLAN assessment writers take some care to present the items that are atypical in some way, so that children are encouraged to use their numeracy skills rather than rote-learn item types. It is possible that the higher facility of the multiple-choice item was influenced by the horizontal presentation of the item, forcing a consideration of how to solve the problem rather than immediately performing an algorithm.

Conclusion

This paper analysed Victorian children's responses to two subtraction items on the 2006 Year 3 AIM assessment for Numeracy. The three-digit multiple-choice item presented horizontally was answered correctly by 69% of children, while the two-digit write-in item presented vertically was answered correctly by 46% of children. The most common error for both items was a regrouping error. The analysis suggests that children at this stage have progressed to being as fluent with three-digit numbers as they are with two digit numbers.

Acknowledgements

AIM data are used and reproduced with permission of the VCAA, and were provided on the understanding that the data would be approached in a holistic fashion. Analysis and findings using that data are not connected with or endorsed by the VCAA.

References

Behrens, J. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2(2)*, 131-160.

Clarke, D., Cheeseman, J., Clarke, B., Gervasoni, A., Gronn, D., Horne, M., . . . Sullivan, P. (2001). Understanding, assessing and developing young children's mathematical thinking: Research as a powerful tool for professional growth.

Ferrari, J. (May 3, 2011). NAPLAN risks 'repeating US mistakes'. *Australian, The* (Australia).

Fuson, K. (1990). Issues in place-value and multidigit addition and subtraction learning and teaching. *Journal for Research in Mathematics Education*, *21*, 273–280.

Fuson, K., Wearne, D., Hiebert, J., Murray, H., Human, P., Olivier, A., ... Fennema, E. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *28(2),*130–162.

Leder, G. (2012). Mathematics for all? the case for and against national testing. *12th International Congress on Mathematical Education.*

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28(2)*, 131–145.

Nuerk, H.-C., Moeller, K., Klein, E., Willmes, K., & Fischer, M. H. (2011). Extending the mental number line. *Zeitschrift für Psychologie*, *219(1*), 3–22.

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

Popham, W. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56(1),* 8–16.

Popham, W. (2009). *Unlearned lessons: Six stumbling blocks to our schools' success.* Cambridge, MA, USA: Harvard Education Press.

Reys, R., Reys, B., McIntosh, A., Emanuelsson, G., Johansson, B., & Yang, D. C. (1999). Assessing number sense of students in Australia, Sweden, Taiwan, and the United States. *School Science and Mathematics, 99(2)*, 61-70.

Romberg, T., Collis, K., & Grouws, D. (1987). *Learning to add and subtract.* Reston, VA: NCTM.

Skemp, R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77,* 20-26.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Pub. Co.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, *34(1)*, 23-25.

Victorian Curriculum and Assessment Authority. (2006a). *Achievement Improvement Monitor (AIM) Year 3 statewide assessment mathematics test.* East Melbourne, Victoria, Australia: Author. Retrieved from http://www.vcaa.vic.edu.au/Documents/aim/AIM2006-Y3Maths.pdf

Victorian Curriculum and Assessment Authority. (2006b). *2006 principals' guide.* East Melbourne, Victoria, Australia: Author.

Victorian Curriculum and Assessment Authority. (2006c, March). *Achievement Improvement Monitor (AIM) Years Prep to 10.* East Melbourne, Australia: Author. Retrieved 18/9/2012, from http://web.archive.org/web/20100323084658/http://www.vcaa.vic.edu.au/vcaa/prep10/aim/aimindex2.html

Young, R., & O'Shea, T. (1981). Errors in children's subtraction. *Cognitive Science, 5(2),* 153–177