African Journal of Research in Mathematics, Science and Technology Education, 2013
Vol. 17, Nos. 1–2, 162–172, http://dx.doi.org/10.1080/10288457.2013.829598

The Instability of KwaZulu-Natal Grade 6 Learners' Mathematics Multiple Choice Test Responses

Iben Maj Christiansen* **and Yougan Aungamuthu**

University of KwaZulu-Natal, South Africa
**Corresponding author, email: christianseni@ukzn.ac.za*

Learning or performance gains are often measured in improvements of test scores over time. However, learners may also answer questions incorrectly which they appeared to master previously, making claims about overall learning gains problematic. Using data from 1,211 grade 6 learners in the Umgungundlovu district in KwaZulu-Natal, we interrogated the consistency of the improvement of their test scores after a year in grade 6. The learners had completed the same 40 question multiple choice test at the onset and the end of grade 6, with no interventions only normal schooling in-between. The content of the test was grade 5 and 6 mathematical content. We know from previous research that the learners were not randomly guessing, yet we found that on average, the learners only got 6 out of 40 questions correct on both test 1 and test 2, did worse on 5 questions and better on 6. Further, almost half the learners both improved on 5 or more questions *and* declined on 5 or more questions. On 35 questions, more than half the learners changed their answers between the two tests. In other words, their performance is very 'unstable'. This indicates that learners are not consistent in their thinking, possibly reflecting guessing, albeit not random. Furthermore, the findings point to a methodological issue of making claims about average performance gains, as these may disguise very unstable learner performance. Such claims should therefore be interrogated before they are applied in making statements about links between learning and teaching or other background factors.

Keywords: learning gain; test methodology; multiple choice; learner performance; middle school; mathematics performance; south africa

Background

Large-scale studies often explore the correlation between learners' performance and factors in the school or larger context, such as teacher education, teacher knowledge, socio-economic status, parents' educational level, and so forth (Carnoy, Chisholm, & Chilisa, 2012; Kanyongo, Schreiber, & Brown, 2007; Mullis, Erberber, & Preuschoff, 2008; Spaull, 2011; van der Berg et al., 2011). In order to hone in on the specific effects of the teaching, or of a particular teaching approach, studies may try to determine the learning gain of learners by comparing their performance on tests before or after the teaching. This was the case in the large-scale grade 6 study done recently in South Africa and Botswana (Aungamuthu, Bertram, Christiansen, & Mthiyane, 2010; Carnoy & Arends, 2012; Carnoy et al., 2012). For instance, Carnoy and Arends (2012) make claims about aspects of teaching related to (assumed) learning gains across two countries.

It is thus important that we understand more about the nature of any apparent gain in performance, such as in which areas the learners' performance has improved and if the overall performance gain disguises reduced performance in some areas. That is what this paper explores for the KwaZulu-Natal grade 6 mathematics study. Specifically, we asked 'How did learners' answers on the multiple

African Journal of Research in Mathematics, Science and Technology Education is co-published by NISC (Pty) Ltd and Routledge, Taylor & Francis Group

choice test change from the test to the re-test, and what does that say about the performance gains of the learners?'

Before discussing relevant literature, we describe the methodology of the study.

Methodology

The study we report on here used previously collected data from a large-scale survey. The data collection for the study took place in 2009. There was no intervention; the study was surveying regular practice in grade 6 teaching and learning. It was linked to a larger study comparing the performance of learners in South Africa and Botswana (Carnoy & Arends, 2012; Carnoy et al., 2012).

Sample

The sampling of grade 6 classes in KwaZulu-Natal was done in three steps. First, for convenience, only schools in the Umgungundlovu district were included. Second, the wealth of the schools was taken into account. In South Africa, schools are divided into quintiles according to the wealth of the school. For the purpose of this study, a list of schools with grade 6 teaching was obtained from the Department of Education, and schools were divided into two groups. The one consisted of schools from the two wealthiest quintiles, the other of schools from the three poorest quintiles. Third, in order to reflect the distribution of schools in the province, 10 schools were randomly chosen from the first group, 30 from the second. When schools were not willing to take part in the study, they were replaced by the next school within the group. In the end, we ended up with 39 schools, and data were collected from all grade 6 classes in these schools. In some cases, this was one class, in other cases two or three. Only some learners were present for both tests, so this reduced the sample for comparisons between the two test performances. In addition, one class was excluded from the study because the very unrealistic test performance gain (from a mean of 25.9% to a mean of 58.7%, or more than three times the improvement of any other class in the sample) indicated possible dishonesty.

Data collection

Learners, principals and teachers completed questionnaires on background information, one lesson in each class was video recorded, a small sample of learners' books were collected at the end of the year, research assistants made observation notes about the school, and learners completed the same test at the beginning and again at the end of the year. This was with the purpose of correlating any apparent learning gain to the quality of the teaching, learners' home situation, etc. It is on the test results we focus in this paper.

The test had 40 multiple choice questions with four answer options each. One thousand two hundred and eleven learners wrote both tests. The gender distribution is show in Table 1.

The questions were identical from test 1 to test 2, but the order was different. All the questions were in the lower ranges of Bloom's taxonomy, with a focus on 'understanding' and 'application' to routine tasks. As the test was to be used in an international comparison, it had been benchmarked around the grade 5–6 curriculum.

Data analysis

In a previous combined-methods study of the data (Aungamuthu et al., 2010), we had interrogated the extent to which the learners' answers could be said to be random from looking at the distribution of answers on correct as well as incorrect answer options. We engage this later in this paper. We now

Table 1: Gender distribution in the sample

Females	Males	No response	Total
624	584	3	1,211
51.5%	48.2%	0.2%	100%

Table 2: An example of relative distribution of learners on their chosen response across the two tests

		Answer on the second test					
Question 5		a	b	c	d	Not answered	Total
Answer on the first test							
	a	*1.5*	1.9	2.2	4.0	0.1	9.7
	b	1.7	*2.9*	2.4	3.8	0.4	11.2
	c	2.8	5.0	*7.8*	8.9	0.3	24.8
	d	4.8	4.6	11.5	*29.3*	0.3	50.5
	Not answered	0.7	0.7	0.7	1.6	*0.2*	3.8
	Total	11.5	15.0	24.6	47.6	1.3	100.0

wanted to know not just how they answered but how consistent their answers were across the two tests. For this purpose, we compared each learner's performance on each question of the second test to their performance on the first test in SPSS and Excel. We did this in two ways. (In the remainder of the paper, we will refer to the situation where a learner answered differently on the two tests as 'the learner changed her answer' or 'movement between answers', though we are aware that the learner did not go back and correct a previous answer, but in most cases answered the question anew in the second test).

First, we assigned +1 to each question which the learner got correct on the second test but had incorrect on the first test, -1 to each question which the learner got incorrect on the second test but had correct on the first test, and a 0 in all other cases. Second, we added these scores for each learner to get an indication of the overall performance gain/reduction between the two tests. Thus, a gain of 5 would indicate that the difference between how many questions the learner had improved on and how many questions the learner had performed worse on was 5, reflecting an assumed performance gain of 5 out of 40 or 12.5%—but potentially disguising a combination of 'learning' and 'unlearning'.

It is possible that learners learn without it being reflected in necessarily choosing the correct answer. For instance, a learner may move from being misled by a drawing accompanying a question to being aware of having to work with the written information but then do so incorrectly. For that reason, we set out to explore not only if learners had changed from incorrect to correct answers (or vice versa) between the two tests, but also to what extent they had changed between incorrect answer options. We did this by summarising for each question how many learners had chosen answer option a in test 1 but answer b in test 2, and so forth for all the options. This resulted in tables such as the following, with the percentage of learners indicated for each cell (numbers rounded for clarity of presentation only):

In this case, the correct answer was option c, but the preferred answer was option d. By adding up along the diagonal (numbers in italics), we can see which percentage of learners did not change their answer (in this case, 41.6%), and thus also which percentage of learners *did* change their answer (in this case, 58.4%). Of these, only 7.8% of the learners stayed with the correct answer across the two tests, while 17.0% of the learners moved away from the correct answer, and 16.8% moved from an incorrect to the correct answer.

It is the frequency of learners changing their answers which we refer to as 'the instability of learners' responses'. One could argue that we should not include the change from incorrect to correct answers in this measure, as indeed education generally strives to facilitate learning which would be expected to result in a change towards correct answers. However, as we can see from the example above, many learners also moved away from the correct answer, and thus we cannot assume that a change to the correct answer reflects learning. Excluding it could give us 'false positives'. We wanted to engage the extent to which learners changed their answers as a supplementary measure of performance/learning, and have therefore included the change to correct answers. In data sets where fewer learners change their answers, it may make more sense to exclude the change to correct answers.

Notions of Validity

The exploration engaged in this paper rests on a commitment to validly measuring learning/performance. It is premised on the assumption that learners have a level of mathematics competency/knowledge which it is possible to approximate through measurements such as tests. While it is recognised that a multiple choice test of the type used in this study only engages a limited range of mathematical competency, it is none the less a premise of most assessment studies that it is possible to gain some measure of competency through testing.

The instrument and data collection must then be able to measure the construct it is intended to measure (construct validity). In post-positivist research, it is also normally assumed that a test must be reliable, that is, produce the same or similar results when repeated. One way to get a measure of this is to look at the correlation between test 1 and test 2 scores. However, this does not, as we show in the result section of this paper, necessarily mean that the test measures a performance which is reflective of an underlying competency.

As a result, we need a notion of validity which includes being able to capture the underlying patterns in learner responses and attempt to explain the behaviour which led to these responses. This moves towards a more interpretivist or constructivist (Guba & Lincoln, 1994) notion of validity to include trustworthiness and credibility.

Excluding Random Guessing

The exploration of the data we share in this paper is based on the first recognition that the distribution of learner answers was similar to what would have been obtained if the learners had guessed randomly. Modelling the number of learners who would get 0, 1, 2, …, 40 correct answers and comparing them to the distribution of the answers in our study shows that performance is not far from equal to that of learners selecting answers randomly (see Figure 1).

When we research learner performance on multiple choice tests, guessing must always be considered a possible factor. Other factors generally considered are distractors which relate to learner misconceptions, language issues, and item difficulty level.

A Rasch analysis of the learners' responses showed that the test overall was too difficult for the majority of learners. This was anticipated, given previous studies of South African learners' Mathematics performance, and it would have been better had the difficulty level been adjusted after a pilot. This was not done, in part because this study was linked to a larger study which involved international comparisons.

We had previously considered guessing as an explanation for learners' responses. If learners were selecting answers randomly, it would seem reasonable with such a large sample to expect a fairly evenly distribution on considered answers—for instance, 25% on each of the options, or a small

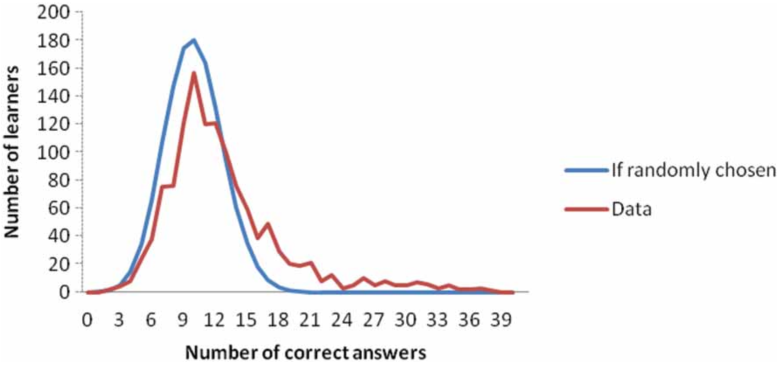

Figure 1: Number of learners as a function of number of correct answers in our data versus in a model assuming completely random choice of answers

percentage on one answer because it has been excluded by many learners, and then a roughly equal distribution across the three remaining options. It was, however, evident that the learners were *not* randomly choosing answers, as the majority would often show a preference for a particular incorrect answer. Instead, the distractors and the language of the test had to be considered.

Literature on Distractors, Language, and Validity of Multiple Choice Tests

The international literature on common learner misconceptions is based on the idea that learners generally do not guess completely randomly; the fact that their incorrect answers follow certain patterns is an indication that they are based on having tried to make sense of the content and the question (Ben-Zvi & Garfield, 2004; Christiansen & Jess, 1999; Ma, 1999; Molina, Castro, & Castro, 2010; Walcott, Mohr, & Kastberg, 2009).

Learners also appear to draw on a number of strategies that are more language informed, such as choosing a word from the question which also appears to be one of the answer options provided; ignoring the difficult words in the question and answering the 'new' question; or choosing the most familiar option (Dempster, 2007; Dempster & Reddy, 2007). The effect of language also showed up in a study of 755 grade 8 learners in the USA, tested on language and mathematics, which found that shortening the stem of the answer was a very effective modification, while adding graphics to the question was not (Kettler et al., 2011).

In a recent paper on the KZN grade 6 mathematics learners' choice of answers, we looked at both misconceptions and language strategies as possible explanations for the learners' responses (Christiansen & Aungamuthu, 2012). We found that learners with a home language other than English made more interpretation mistakes as well as displayed more misconceptions. That analysis provides a background to the current paper in the sense that it excludes, in all reasonable ways, that learners were simply choosing answers randomly.

Many studies of validity of pre-post-test studies address issues in intervention studies, where issues of biased sampling, the Hawthorne effect, lack of a control group, and taking into account 'natural' maturation are amongst those considered (see for instance the review by Marsden & Torgerson, 2012). While the Hawthorne effect has to be considered in our case, it does not explain the patterns in the changes of answers. The other issues are not relevant, as this was not an intervention study, and as we are not looking at the overall gain in performance but at the patterns of changes in responses.

What Marsden and Torgerson propose as another validity issue is the 'regression to the mean' phenomenon, which manifests in a strong negative correlation between the score on test 1 and the performance gain. They discuss this as the result of high scoring learners having less chance of a large performance gain, and opposite for learners who scored low on the first test. This it is a validity concern in studies which analyse data with extreme scores on the first test—such as very low scores overall. They explain:

> Scores towards the ends of a distribution will, on average, be more likely to have a higher error term than those nearer the mean. When students are re-tested the results towards the ends of the distribution will tend to move closer to the mean (to their 'true' value) than the results in the middle range. (p. 585)

This should of course only be the case if the very high and very low scores were not true reflections of learners' competencies, in other words if some guessing or luck was involved, not with very skewed distributions like the one in our study. Yet it is worth considering, and we will return to it in the results section.

All the studies we have found engaging performance gains or assumed learning gains deal with the overall scores of correct answers as the change over time—Carnoy and Arends (2012) and Graham and Provost (2012) are two recent examples. Sicoly (2002) looked at stability of scores across 2 years, and found high degrees of instability of results, but this was not on the same test, and reasons for the instability were not investigated empirically. We have found no other studies which have engaged changes in learners' incorrect answers on two versions of the same test. Of the 16 intervention studies reviewed by Marsden and Torgerson (2012), none considered regression to the mean, and this was not engaged in Carnoy and Arends' (2012) work either despite the low overall improvement in performance. Thus there is a need to not only interrogate the extent to which regression to the mean occurs but also to explain the underlying patterns of learner behaviour. Maybe looking into how learners change their answers could be a step in that direction.

Results

In the following, we present our findings from the two stages of analysis, as per the 'data analysis' section. First, we look at the changes of learners answers from correct to incorrect and from incorrect to correct. Secondly, we look at the changes between all answer options.

Learners' gains and declines in performance

Comparing the learners' performance on the two tests, we found an overall average performance gain of just under 1 answers (mean = 0.998), from a sample mean of 29.6% to a sample mean of 32.1%. A paired samples t-test revealed that learners scored 2.5 percentage points higher on average in the second test than on the first test ($t(1210) = 8.203$, p-value <0.01). This is compatible to the 3 percentage point performance gain found for the grade 6 learners in the North West Province (Carnoy et al., 2012).

We found a correlation of –0.30 between the scores on the first test and the performance gain, indicating little likelihood of the presence of 'regression to the mean'. However, even if there was a 'movement towards the mean' this would not reflect or explain the large movement between answers which we discuss later in this section.

Of the 1,211 learners who wrote both tests, 542 showed an overall improvement of 2 or more, 339 learners' performance changed by no more than 1 in either direction, and 330 learners reduced their performance by 2 or more. This means that while 44.8% of the learners improved, 27.3% of the learners did worse on the second test. The remaining 28.0% did not change their performance by more than 1. And 8.6% of the learners reduced their overall performance by 5 or more, on a test with 40 questions. (See Table 3 for an overview of these data.)

The correlation between learner scores on each test is $r = 0.712$ (p-value < 0.01). Thus, a learner's mark on test 1 is significantly associated with their test 2 mark, which normally is taken to indicate a reasonable test-retest reliability. Together with the summarised changes in learner performance, it is tempting to assume that learners' responses reflect real learning and real forgetting. However, our scoring shown in Table 3 disguises situations where learners improved their performance on some questions but also decreased their performance on other questions. About 18.8% of the learners got less than 3 questions correct on both tests, and 2.0% got no questions correct on both tests. On average, the learners only got 6 questions correct on both test 1 and test 2, did worse on 5 questions and better on 6. We therefore looked at how many learners had gone from correct to incorrect on several questions. The results are shown in Table 4. The categories (less than 4, 5–7, and 8 or more) are arbitrary in the sense that other cut-off points could have been used to illustrate the point.

To test if answer changing (from correct on test 1 to incorrect on test 2) between tests had a significant effect on test 2 performance, the sample was divided into three groups based on Table 4 above. In other words, students who scored incorrect on less than 5 questions in test 2 but had got those questions correct on test 1 were labelled group 1. Similarly, those students who got from 5 to 7 questions

Table 3: Summarised changes in learner performance

Change in score btw test 1 and test 2	-5 or worse	[-4:-2]	[-1; + 1]	[+2: + 4]	+ 5 or more
Number of learners	104	226	339	323	219
Percentage of learners	8.6%	18.7%	28.0%	26.7%	18.1%

Table 4: Number and percentage of learners decreasing their performance on a range of questions

	Scored incorrect on less than 5 questions in test 2, which were answered correctly in test 1	Scored incorrect on 5–7 questions in test 2, which were answered correctly in test 1	Scored incorrect on 8 or more questions in test 2, which were answered correctly in test 1
Number of learners	471	493	247
Percentage of learners	38.9%	40.7%	20.4%

wrong on test 2 which they had previously got right on test 1 were labelled group 2. Group 3 consisted of students who changed 8 or more questions from a correct answer in test 1 to an incorrect answer on test 2. Thus, group 1 consists of students with the least number of changes in correct answers from test 1 to test 2.

A one way ANOVA revealed that there was a significant effect of number of answer changes (from correct on test 1 to incorrect on test 2) on test 2 performance: the Levene statistic of 55.943 was significant with $p < 0.05$. This meant that the homogeneity of variance assumption for ANOVA was violated requiring the use of the Welch *F* statistic. The Welch *F* (2, 721.963) = 80.018 and $p < 0.05$. Thus there was a significant difference in test 2 performance among the three groups.

The Games-Howell *post hoc* procedure established that the mean performance on test 2 for group 1 was greater than the mean performance of group 2 and group 3; group 3 performed worst, on average, compared to the other two groups. (Refer to Appendix 1 for the Games-Howell statistics.)

Compare this to how many learners had gone from incorrect to correct on several questions (Table 5):

Table 5: Number and percentage of learners improving their performance on a range of questions

	Scored correct on less than 5 questions which were answered incorrectly in test 1	Scored correct on 5–7 questions which were answered incorrectly in test 1	Scored correct on 8 or more questions which were answered incorrectly in test 1
Number of learners	282	565	364
Percentage	23.2%	46.7%	30.1%

To test if answer changing (from incorrect on test 1 to correct on test 2) between tests had a significant effect on test 2 performance, the sample was divided into three groups based on Table 5 above. In other words, students who scored correct on less than 5 questions in test 2 but had got those questions wrong on test 1 were labelled group 1. Similarly, those students who got from 5 to 7 questions correct on test 2 which they had previously got wrong on test 1 were labelled group 2. Group 3 consisted of students who got 8 or more questions correct in test 2 which they had answered incorrectly on test 1. Thus, group 3 consists of students with the most number of changes from incorrect answers on test 1 to correct answers on test 2.

A one way ANOVA revealed that there was a significant effect of number of answer changes (from incorrect on test 1 to correct on test 2) on test 2 performances: the Levene statistic of 23.648 was significant with $p < 0.05$. This meant that the homogeneity of variance assumption for ANOVA was violated requiring the use of the Welch *F* statistic. The Welch *F* (2, 612.062) = 44.942 and $p < 0.05$. Thus there was a significant difference in test 2 performance among the three groups.

The Games-Howell *post hoc* procedure established that the mean performance on test 2 for group 3 was greater than the mean performance of group 2 and group 1; group 1 performed worst, on average, than the other two groups. Refer to Appendix 2 for the Games-Howell statistics.

There were 42 learners (3.5%) who *both* improved on 8 or more questions and declined on 8 or more other questions, and there were 552 learners (45.6%) who *both* improved on 5 or more questions and declined on 5 or more other questions. In other words, almost half of the learners changed their answers substantially. There were only 68 learners (5.6%) who improved on 8 or more questions without also declining their performance on more than 2 questions.

Overall, what appears to be learning gains according to test performance are mostly offset by declines in performance on other questions of the test. As a result, the correct answers on test 2 were often not given by the same learners who had provided correct answers to test 1. This indicated to us that it was worth looking deeper into the changes of answers to include changes *between* incorrect answer options.

Instability of answers

As described under 'Data Analysis,' we created tables for each question showing the learners' change of answers from test 1 to test 2 (See Table 2). It is important to note here that we were no longer interested in the changes between incorrect and correct, but in how frequently learners changed their answers.

When looking across questions, we found that on 35 of the 40 questions, at least half the learners changed their answer from test 1 to test 2, varying between 36% and 71% of learners changing their answers to a question. This is what we refer to as 'instability of answers'. There was a negative correlation ($r = -0.591$, p-value < 0.01) between how many learners got the correct answer to a question on both tests, and how many learners changed their answer to the same question, indicating that they are less likely to change answers on the more established content areas. This sounds credible.

Looking instead across learners, they changed between 3 and 40 answers on the test. On average, the learners changed 24 of their answers (median = 24, mean = 23.6). There was a negative correlation between the learners' score on their first test and the number of answers they changed ($r = -0.627$, p-value < 0.01), and a negative correlation between the number of questions a learner got correct on both tests and the number of answers they changed ($r = -0.77$, p-value < 0.01). Thus, as we would have expected, better performing learners were less likely to change a lot of answers. Yet, there were but small positive correlations between numbers of answers changed and improving on more than 4 questions ($r = 0.29$, p-value < 0.01) or improving on more than 7 questions ($r = 0.241$, p-value < 0.01).

If changes between answer options for a question showed an overall move away from one or more options or an overall move towards one or more options, such a pattern may credibly reflect a change in performance—whether this then reflects a desired or an undesired learning. We therefore looked at the overall change in learners choosing each of the answer options for each question, and calculated what the overall change in percentage of learners choosing each option was. For instance, 0.16% fewer of the learners preferred the correct answer to question 5, while 3.72% more preferred option (b).

For 27 of the questions, the correct answer was more popular in the second test, by more than one percentage point. For 7 of these, there was however an incorrect option which got a higher change in frequency. For 5 of the questions, the correct answer was less popular in the second test, by more than one percentage point. It would require further exploration to explain to what extent these changes reflect learning or not.

Discussion

On face value, there is a small average performance gain of 2.5%. However, this overall performance gain disguises that while 44.8% of the learners performed better on the second test, more than a quarter of the learners performed worse. It also does not reflect that 45.6% of the learners did better on 5 or more questions but also worse on 5 or more questions. On average, learners changed their answers on 60% of the questions mostly without getting a much better result. Thus, the answer to the first part of our question is pragmatically that learners changed their answers very frequently, from correct to incorrect, from incorrect to correct, and between incorrect options. These results indicate that the overall performance gain disguises reduced performance for many learners, and thus illustrates how problematic it is to make claims about performance gains (and even more so about learning gains) and possible causes of learning without looking deeper into the flurry which may hide behind a small change in test score.

The first possible explanation for the large number of learners changing their answers is that they are guessing, but our previous analysis of learner responses (Christiansen & Aungamuthu, 2012) had in substantial ways challenged the hypothesis that learners were choosing answers at random. Furthermore, if many learners had guessed answers randomly, we would have expected a strong regression to means, which we did not find. Thus these learners appear to be making an effort to answer the questions, but they are not consistent in their use of alternative strategies and/or misconceptions.

We cannot explain why this is the case, but we speculate that it is because many learners are so uncertain about what the correct strategy is, that the choice of how to approach the question in itself becomes somewhat arbitrary. In that sense, it is likely to reflect the incomplete and/or incoherent coverage of the curriculum documented by other studies (Reeves, 2005; Reeves & McAuliffe, 2012) and generally low attainment of the curriculum outcomes. For instance, learners may look at this question

25 x _____ = 50 x 2

(a) 2
(b) 4
(c) 50
(d) 100

and exclude the second option (the correct one), because they have an operative perception of the equal sign (cf. Brodie & Shalem, 2011; Molina et al., 2010). Preferred answers were ‘100’ (reading from right to left) or ‘2’ (reading from left to right), in both cases ignoring the rest of the equation. Of these, ‘2’ was selected by 33.7% of the learners in the first test and 36.3% of the learners in the second test, while ‘100’ was selected by 27.1% respectively 27.7% of the learners. However, though the overall distribution on options remained almost the same, 61% of the learners changed their answer on this question. Thus, they may be aware of more than one way of reading the question but being uncertain which one to choose. This then becomes somewhat arbitrary. This, of course, is only one possible explanation, albeit a credible one.

A concern is that the test was well matched to the intended learning outcomes, but was not well matched to the achieved learning outcomes; it was too difficult for most of the learners. (Until we can test learners electronically, with difficulty levels of questions based on responses to previous questions, this is going to remain a problem.) This would explain why the learners performed poorly overall, and why they were often using language strategies or were guided by misconceptions. It does not, however, explain why their choice of answers varied so much.

To us, the results are highly relevant. First, they show much higher degrees of arbitrary answers than analysing one test indicates, while still excluding completely random guessing. Second, they challenge the validity of large-scale studies in our context which treat differences in scores as indicative of a real gain in learning (or unlearning, for that matter). Rather, statistical analyses have to allow for the ‘instability’ of responses, reflected in the high frequency of changed answers between tests. When the aim is to validly measure learners’ performance, considering the number of changed answers may well be a useful addition to summaries of the number of correct answers learners choose. Finally, our results indicate that learners might not always be as consistent in their answers as the literature on misconceptions may lead one to believe.

References

Aungamuthu, Y., Bertram, C., Christiansen, I. M., & Mthiyane, N. (2010). *Grade 6 Mathematics classrooms in KwaZulu-Natal*. UKZN, South Africa: School of Education and Development, Faculty of Education.

Ben-Zvi, D., & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht: Kluwer Academic.

Brodie, K., & Shalem, Y. (2011). Acountability conversations: Mathematics teachers' learning through challenge and solidarity. *Journal of Mathematics Teacher Education*, *14*, 419–439.

Carnoy, M., & Arends, F. (2012). Explaining mathematics achievement gains in Botswana and South Africa. *Prospects*, *42*, 453–468.

Carnoy, M., Chisholm, L., & Chilisa, B. (Eds.). (2012). *The low achievement trap*. Cape Town, South Africa: HSRC Press.

Christiansen, I. M., & Aungamuthu, Y. (2012). Language issues, misconceptions and confusion in abundance: A qualitative analysis of grade 6 learners' responses on a mathematics test. *Education as Change*, *16*(1), 51–67.

Christiansen, I. M., & Jess, K. (1999). Learning from children's work, Part one: Young children's addition. *Pythagoras*, (50), 13–19.

Dempster, E. (2007). Textual strategies for answering multiple choice questions among South African learners : What can we learn from TIMSS 2003? *African Journal of Research in Mathematics, Science and Technology Education*, *11*(1), 47–60.

Dempster, E., & Reddy, V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, *91*, 907–925.

Graham, S. E., & Provost, L. E. (2012). Mathematics achievement gaps between suburban students and their rural and urban peers increase over time. *Carsey Institute Brief no 52*. Durham, NH: University of New Hampshire.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: SAGE Publications.

Kanyongo, G. Y., Schreiber, J. B., & Brown, L. I. (2007). Factors affecting mathematics achievement among 6th graders in three sub-Saharan African countries: The use of hierarchical linear models (HLM). *African Journal of Research in Mathematics, Science and Technology Education*, *11*(1), 37–46.

Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (2011). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. *Applied Measurement in Education*, *24*(3), 210–234.

Ma, L. (1999). *Knowing and teaching elementary Mathematics: Teachers' understanding of Fundamental Mathematics in China and the United States*. Mahwah, NJ: Lawrence Erlbaum.

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, *38*(5), 583–616.

Molina, M., Castro, E., & Castro, E. (2010). Elementary students' understanding of the equal sign in number sentences. *Electronic Journal of Research in Educational Psychology*, *7*(17), 341–368.

Mullis, I. V. S., Erberber, E., & Preuschoff, C. (2008). Chapter 13: The TIMSS 2007 international benchmarks of student achievement in mathematics and science. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 Technical Report* (pp. 339–347). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Reeves, C. (2005). *The effect of opportunity-to-learn and classroom pedagogy on mathematics achievement in schools serving low socio-economic status communities in the Cape Peninsula* (Unpublished PhD thesis). University of Cape Town, Cape Town.

Reeves, C., & McAuliffe, S. (2012). Is curricular coherence slowing down the pace of school mathematics in South Africa? A methodology for assessing coherence in the implemented curriculum and some implications for teacher education. *Journal of Education*, (53), 9–36.

Sicoly, F. (2002). Stability of school-level scores from large-scale student assessments. *Applied Measurement in Education*, *15*(2), 173–185.

Spaull, N. (2011). A preliminary analysis of SACMEQ III. *Stellenbosch Economic Working Papers*, (11/11). Retrieved from www.ekon.sun.ac.za/wpapers/2011/wp112011/wp-11-2011.pdf (accessed 5 March 2012).

van der Berg, S., Burger, C., Burger, R., de Vos, M., du Rand, G., Gustafsson, M., . . . von Fintel, D. (2011). *Low quality education as a poverty trap*. Stellenbosch, South Africa: Stellenbosch University.

Walcott, C., Mohr, D., & Kastberg, S. E. (2009). Making sense of shape: An analysis of children's written responses. *The Journal of Mathematical Behavior*, *28*(1), 30–40.

Appendix 1

Dependent Variable: SCORE on test 2
Games-Howell

					95% Confidence Interval	
(I) Group	(J) Group	Mean Difference (I–J)	Std. Error	Sig.	Lower Bound	Upper Bound
1.0	2.0	7.8888*	.9422	.000	5.676	10.101
	3.0	12.3977*	.9802	.000	10.096	14.700
2.0	1.0	–7.8888*	.9422	.000	–10.101	–5.676
	3.0	4.5089*	.7609	.000	2.721	6.297
3.0	1.0	–12.3977*	.9802	.000	–14.700	–10.096
	2.0	–4.5089*	.7609	.000	–6.297	–2.721

* The mean difference is significant at the 0.05 level.

Appendix 2

Dependent Variable: SCORE on test 2
Games-Howell

					95% Confidence Interval	
(I) group	(J) group	Mean Difference (I–J)	Std. Error	Sig.	Lower Bound	Upper Bound
1.0	2.0	–1.4712	1.2166	.448	–4.333	1.391
	3.0	–8.7068*	1.2815	.000	–11.720	–5.694
2.0	1.0	1.4712	1.2166	.448	–1.391	4.333
	3.0	–7.2356*	.8243	.000	–9.171	–5.300
3.0	1.0	8.7068*	1.2815	.000	5.694	11.720
	2.0	7.2356*	.8243	.000	5.300	9.171

* The mean difference is significant at the 0.05 level.