ANALYSING LONGITUDINAL DATA ON STUDENTS' DECIMAL UNDERSTANDING USING RELATIVE RISK AND ODDS RATIOS

Vicki Steinle and Kaye Stacey

University of Melbourne

The purpose of this paper is to demonstrate the use of the statistics of relative risk and odds ratios in mathematics education. These statistics are widely used in other fields (especially medical research) and offer a useful but currently under-utilised alternative for education. The demonstration uses data from a longitudinal study of students' understanding of decimal notation. We investigate the statistical significance of results related to the persistence of misconceptions and the hierarchy between misconceptions. Relative risk and odds ratio techniques provide confidence intervals, which give a measure of effect size missing from simple hypothesis testing, and enable differences between phenomena to be assessed and reported with impact.

This paper demonstrates some possibilities for analysing educational data, which draw upon methods that are widely used in reporting results of medical, environmental and epidemiological research. We believe that these measures provide very useful techniques for testing for statistical significance and reporting confidence intervals, which will enhance mathematics education research. Capraro (2004) draws attention to important recent policy changes within the American Psychological Association, evident in their publication manual, that stress the importance of researchers supplementing statistical significance testing with measures of effect size and confidence intervals, with many journals making them mandatory.

For those worried about deep vein thrombosis (DVT) after long flights, BUPA's website (Newcombe, 2003) cites Farrol Kahn as saying that "Several studies have shown [wearing flight socks] to be of benefit and it reduces the risk by up to 90 per cent." However, we can be cheered that "The researchers discovered the risk of developing DVT after a long-haul flight seemed to be low - at about 1 per cent of all long-haul passengers." Some of us can be further comforted by the observation of Runners' World (Reynolds) that "Being athletic accounts for ten times more victims than any other risk factor."

These reports in the popular press, along with reports in research literature, are mostly describing the results in terms of *relative* rather than *absolute* risk. So, for example, instead of commenting that DVT developed in only about 0.10% (10% of 1%) of passengers wearing flight socks, the website reports that the risk is reduced by up to 90%. This puts what might be seen as a tiny reduction in risk (just 90% of 1%) into perspective and shows its importance.

In this paper, we will show how these ideas of relative risk can be applied to educational data and discuss the benefits and issues arising. We illustrate the methods and challenges by some reanalyses of longitudinal data on students' understanding of

2005. In Chick, H. L. & Vincent, J. L. (Eds.). *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education*, Vol. 4, pp. 217-224. Melbourne: PME.

decimals. This was a cohort study, which tracked the developing understanding of over 3000 students in Years 4 – 10 at 12 schools for up to 4 years, testing them with the same test at intervals of approximately 6 months. Details of the sampling, the test and its method of analysis and many results have been described elsewhere; for example, Steinle and Stacey (2003), and Steinle (2004). For the purpose of this paper, it is sufficient to know that students are classified into 4 *coarse codes*, (A, L, S and U) on the basis of their answers to one section of this Decimal Comparison Test. In general terms, students in coarse code A are generally able to compare decimals; students in coarse code L generally treat longer decimals as larger numbers (for a variety of reasons); students in coarse code S generally treat shorter decimals as larger numbers (again for a variety of reasons); and coarse code U contains all remaining students. Answers to the other items in the test refine these coarse codes into 12 *fine codes*, which represent *expertise* (A1, which is a subset of A), various particular misconceptions, or students who cannot be classified. The longitudinal study traced student's understanding in terms of the coarse and fine codes and used this to examine questions such as which misconceptions are prevalent at different ages, whether some misconceptions are better to have than others, how often students appear to lose expertise, and whether students tend to move between misconceptions in predictable ways.

AN EXAMPLE USING RELATIVE RISK AND ODDS RATIOS

The main ideas in this paper will be illustrated by considering the question of whether it is *better* for a student to be in code L or S, i.e. from which of these groups are students more likely to become experts (i.e., move to code A1) on their next test? Table 1 summarises the data. Looking over the whole sample[1], there were 847 occasions where a student completed a test coded as S and then completed another test. On this subsequent test, 230 of the S students became experts and 617 did not, giving a 27% (230/847) chance of an S student becoming an expert and a 73% chance of an S student *not* becoming an expert. Similarly, from Table 1, there were 1257 occasions where a student completed an L test and was followed to their next test. The L students had 20% (251/1257) chance of becoming an expert. It seems that it is better to be an S student[2].

Conditions	Outcome_1 (A1 on next test)	Outcome_2 (not A1 on next test)	Total
Condition_1 (S)	$n_{11} = 230$	$n_{12} = 617$	$n_1 = 847$
Condition_2 (L)	$n_{21} = 251$	$n_{22} = 1006$	$n_2 = 1257$

Table 1: Numbers of A1 and non-A1 tests following S and L tests

[1] More careful analysis, as in Steinle (2004), would define the samples to reduce the effect of confounding variables such as age. The purpose here is to illustrate the procedures; the results here broadly match the refined analysis.

[2] This result is consistent with responses to individual items reported by large-scale studies around the world since the 1980s. See, for example, Foxman et al. (1985).

There are several ways in which this result can be tested statistically. A chi-squared test rejects the null hypothesis that the proportions of L and S students becoming expert are the same ($\chi^2 = 14.82$, d.f.=1, p=0.0001). However, the chi-squared test simply indicates the degree of evidence for association and does not give other information such as a confidence interval.

Analysing absolute differences in proportions becoming experts

A second method is to test whether the proportions of students going to A1 (moving to expertise) from S and L are the same. Assuming the counts for S and L are independent binomial samples, the difference of the proportions from Table 1 is distributed approximately normally with mean (0.27 – 0.20) and standard error 0.019 (Agresti, 1996). Hence a 95% confidence interval for the true difference in probabilities is $0.07 \pm 1.96 \times 0.019$, i.e. the interval (0.03, 0.11). This confidence interval provides more information than the chi-squared test. The interpretation of this confidence interval is in terms of *absolute differences* in the chance of moving to A1 from S and L. With 95% confidence, the percentage of S students becoming expert is between 3 and 11 more than for L students. In other words, if we have 100 L and 100 S students, and if 20 L students become experts on the next test, we can be confident that between 23 and 31 S students will become experts.

Analysing relative risk of becoming an expert

Another approach to testing whether two proportions are the same is to consider the *relative*, rather than the *absolute* difference in the proportions as above. This is especially useful when the proportions are small, as the absolute differences will also be small, although their ratios may be large. Because of its origins in epidemiological studies, the proportions of interest are classically labelled *risk*, but in our circumstance (where becoming an expert is a benefit rather than a harm) *chance* seems a more appropriate label. To answer the question of whether it is better to be S or L, the *relative risk (chance)* of becoming an expert on the next test is calculated as the ratio of the chances of S to A1 and L to A1. Figure 1, where the steps involved are demonstrated and given to two decimal places, shows that the relative chance of becoming an expert (from S and L in that order) is 0.27/0.20 = 1.36. This number indicates that an S student is 36% more likely to become an expert on the next test than is an L student.

Is this a significant difference? As indicted in Figure 1, the natural logarithm of this relative chance (i.e. relative risk) is normally distributed (Agresti, 1996; Bulmer, 2005), and the 95% confidence interval for the relative chance of becoming an expert is (1.16, 1.59). As 1.00 is not inside this interval, we are 95% confident that an S student is more likely to become an expert than an L student. In fact it is reasonable to say that an S student has at least a 16% greater chance of becoming an expert and possibly up to 59% more chance, compared with an L student. The best estimate is 36% more chance since the relative chance is 1.36. This is an intuitive way of presenting the results, with some impact.

Relative Risk (RR)		**Odds Ratio (OR)**	
Risk of Out_1 given Con_1 *Chance of A1 given S*	$p_{1,1} = \frac{n_{11}}{n_1} = \frac{230}{847}$ $= 0.27$	Odds for Out_1 given Con_1 *Odds of A1 given S*	$o_1 = \frac{n_{11}}{n_{12}} = \frac{230}{617}$ $= 0.37$
Risk of Out_1 given Con_2 *Chance of A1 given L*	$p_{1,2} = \frac{n_{21}}{n_2} = \frac{251}{1257}$ $= 0.20$	Odds for Out_1 given Con_2 *Odds of A1 given L*	$o_2 = \frac{n_{21}}{n_{22}} = \frac{251}{1006}$ $= 0.25$
Relative Risk of Out_1 (Con_1, Con_2) *Relative Chance of A1 (S, L)*	$RR_1 = \frac{p_{1,1}}{p_{1,2}} = \frac{0.27}{0.20}$ $= 1.36$	Odds Ratio for Out_1 (Con_1, Con_2) *Odds Ratio of A1 (S, L)*	$OR_1 = \frac{o_1}{o_2} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$ $= 1.49$
$Ln(RR_1)$ is normally distributed around $Ln(p_{1,1}/p_{1,2})$ with $SE = \sqrt{\frac{1-p_{1,1}}{n_{11}} + \frac{1-p_{1,2}}{n_{21}}}$	Ln 1.36 = 0.31 $SE = 0.08$	$Ln(OR_1)$ is normally distributed around $Ln(o_1/o_2)$ with $SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$	Ln 1.49 = 0.40 $SE = 0.10$
95% confidence interval for $Ln(RR_1)$ is $Ln(p_{1,1}/p_{1,2}) \pm 1.96 \times SE$	$0.31 \pm 1.96 \times 0.08$ $= (0.15, 0.46)$	95% confidence interval for $Ln(OR_1)$ is $Ln(o_1/o_2) \pm 1.96 \times SE$	$0.40 \pm 1.96 \times 0.10$ $= (0.20, 0.61)$
Conclusion based on whether 1 is included in the 95% confidence interval for RR_1 (which is the antilog of above).	95% CI for RR_1 is $(e^{0.15}, e^{0.46}) = (1.16, 1.59)$ which does not include 1.00	Conclusion based on whether 1 is included in the 95% confidence interval for OR_1 (which is the antilog of above).	95% CI for OR_1 is $(e^{0.20}, e^{0.61}) = (1.22, 1.83)$ which does not include 1.00

Figure 1: Calculations of Relative Risk and Odds Ratio from Table 1

Analysing the odds ratio for becoming an expert

The right-hand side of Figure 1 also provides an explanation of a related measure of association called the *odds ratio*. The odds of an S student becoming an expert on the next test are 230:617 = 0.37 and the odds for an L student are 251:1006 = 0.25. The *odds ratio* is therefore 0.37/0.25 = 1.49. The calculation of the 95% confidence interval for the odds ratio is (1.22, 1.83) (Agresti, 1996; Bulmer, 2005). As 1.00 is not inside this interval, we can be 95% confident that there is a true difference between the odds for an S and an L student becoming an expert on the next test.

The odds ratio is harder to interpret than the relative risk considered above, but it is widely used because it can be applied to a wider range of research designs than relative risk, and has strong mathematical properties giving it a role in other statistical testing. Moreover, when the risks of the event under both conditions are low (e.g. less than 10%), the odds ratio is a good approximation to the relative risk and can be

interpreted as such. SPSS performs odds ratio calculations under the Crosstabs menu, as the *Mantel-Haenszel common odds ratio estimate.*

Features of Relative Risk and Odds Ratio Analyses

In using both relative risks and odds ratios, it is important to think carefully about what is a good comparison to demonstrate an effect. If we had carried out the odds ratio analysis for L compared to S (i.e. swapping conditions in Table 1) then the odds ratio would have been the reciprocal i.e. 1/ 1.49 (i.e. 0.67). Similarly, the relative risk of moving to expertise of L compared to S is the reciprocal of 0.20/0.27, i.e. 74%. When the relative risk is less than one, it is common to use *relative risk reduction* to present the results. Instead of saying that an L student has only 74% of the chance (risk) of becoming an expert that an S student has, it is common to talk about a 26% reduction in the chance of becoming an expert, as was done in the DVT example in the introduction. This again is an intuitive way of presenting the results with impact.

If the odds ratio test gives a significant result for S compared to L, will the test also be significant for L compared to S? The answer is yes: the only disadvantage is that the point estimates less than 100% are harder to describe in words, as indicated above. The formulas in Figure 1 show that the confidence interval would have been obtained from the reciprocals (1/1.83, 1/1.22) = (0.55, 0.82). So, if one of these confidence intervals includes 1.00 (so that the null hypothesis is accepted), then the other will automatically. The same situation applies for relative risk: if S compared to L is significant, then L compared to S will be significant. The choice of whether to discuss condition_1 to condition_2 or vice versa is therefore a choice between interpreting ratios greater or less than one.

Another important question is: if a test of the relative risk (or odds ratio) shows a significant difference in the chances that an event E happens, would these tests show significant differences in the chances that the event not-E happens? In our example, both the relative risk and odds ratios show S students have more chance of becoming expert than L students. Is it also the case that there is a significant difference in the chance of S students, compared with L students, *not* becoming an expert on their next test? Note that in this case, *risk* is a good term because not becoming an expert is a perceived harm. For the odds ratio, this result is true – a significant result for event E implies a significant result for event non-E. This is an advantage of the odds ratio analysis. This situation, however, does not automatically follow for relative risk analysis. For example, the relative difference between risks of 1% and 2% for event E is much larger than the relative difference for risks of 99% and 98% for event non-E.

ESTIMATES OF RISK IN THE LONGITUDINAL DATA

In this section, we apply the techniques described above to questions of hierarchy (which misconceptions are *better* to have), and persistence (are some misconceptions likely to trap students more than others), and consider some of these questions by comparing students in primary school (Years 4 – 6) with secondary school students

(Years 7 – 10). As noted above, Steinle (2004) presents an analysis where confounding variables related to the sampling are treated carefully. The results presented here are in agreement with those from more careful analyses and therefore summarise some of the major results of the refined data analysis.

Hierarchy: which misconceptions are best to have?

The preceding analysis demonstrated that a student in code S is more likely to become an expert (A1) by the next test than a student in code L. The 95% confidence intervals of both relative risks (RR) and odds ratios (OR) determined in Figure 1 are provided graphically in Figure 2 (see the lowest two rows). Confidence intervals which are larger than 1.00 indicate a significant difference with the condition first listed having the larger result. So, it is clear being in L is *worse* than being in S, but which is the *best* code to have: S or U or A?

The intermediate rows of Figure 2 show the confidence intervals for both measures (RR and OR) for a comparison of code U with code S. The RR indicates that a U student is between 1.2 and 1.6 times more likely than an S student to be an expert on the next test. (Typically, students who answer the test inconsistently and hence are not classified by the test, belong to the group U). The top two rows in Figure 2 show that in turn, students in code A are more likely than those in code U to be experts on the next test. This is to be expected, since the numerically largest group in code A is in fact the experts (A1). Note that the confidence interval for OR in row 2 is off the graph to the right (it is between 7.8 and 10.9).

Together these results show that the hierarchy of these four codes is (highest to lowest) A, U, S, then L. It is best to be an expert or near expert (i.e. in A), then it is best to be undecided (U), then to have a shorter-is-larger (S) misconception and worst to have a longer-is-larger (L) misconception.

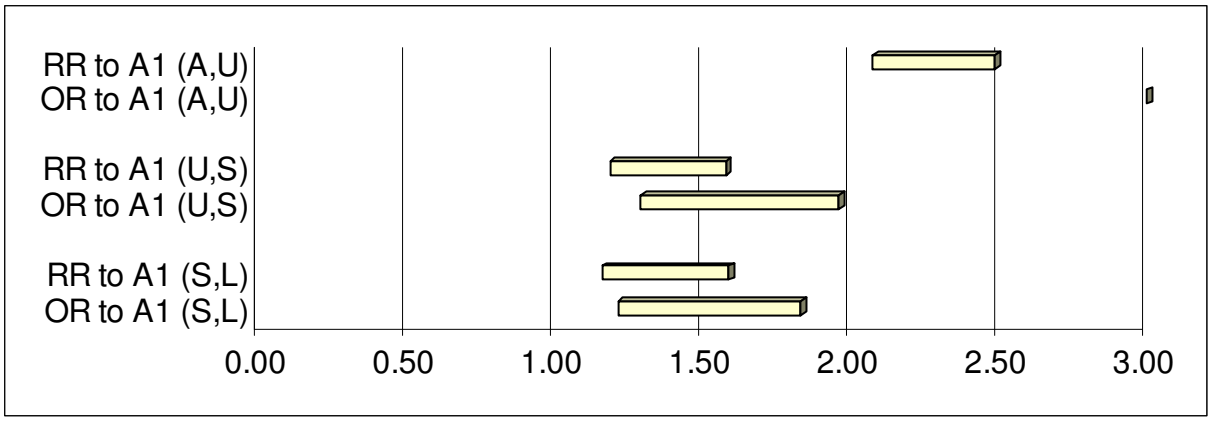

Figure 2: Confidence intervals for RR and OR analyses of the movement to expertise from various codes. (Note. Row 2 is off the scale to the right, so not shown)

Persistence: do some misconceptions keep students longer than others?

Steinle (2004) examined various measures of persistence – how often students retest in the same code on their next test. The basic finding was that 89% of A1 students retest as A1 at the next test (i.e., *persist* in A1), compared to 44% of L students persisting in L, 38% of S students persisting in S and 29% of U students persisting in U. Note that persisting in A1 is desirable, while persisting in other codes is not. Closer analysis showed interesting variations between older and younger students, some of which are summarised graphically in Figure 3.

The top two rows of Figure 3 show that, as both confidence intervals include 1.00, there is *not* a significant difference in the persistence in A1 by students in Secondary school compared with students in primary school. The next two rows are to the left of 1.00 indicating that there is a significant difference and it is the younger L students who have higher levels of persistence than the older L students. Rows 5 and 6 indicate that the opposite result holds true for the S students. In particular, row 5 indicates that older S students are approximately 1.5 times more likely to persist in S than the younger S students. The last two rows indicate that there is no significant difference between older and younger U students in their persistence in U.

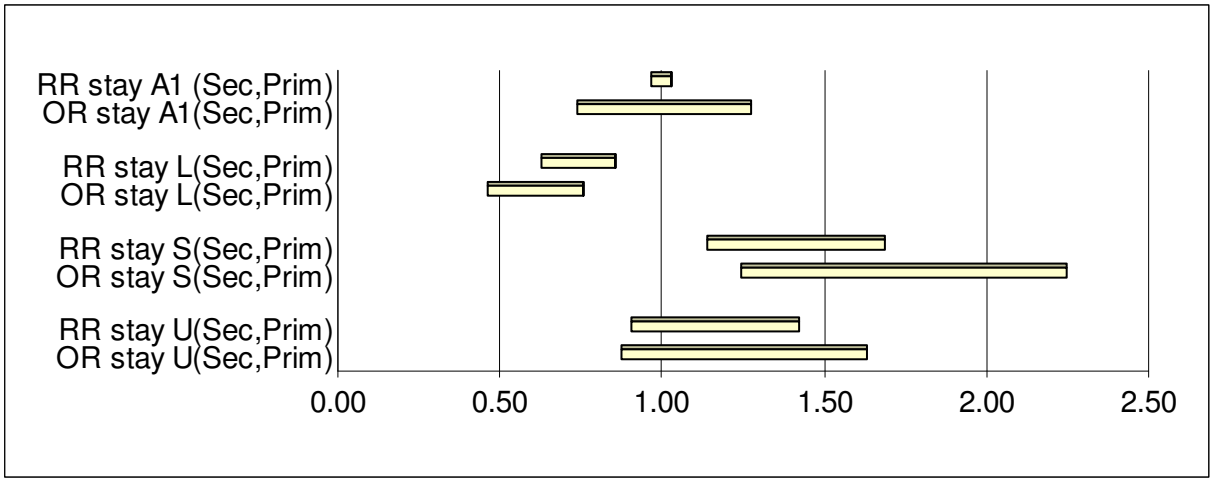

Figure 3: Confidence intervals for RR and OR analyses of the persistence in various codes between students in Secondary and Primary schools

CONCLUSION

The main aim of this paper has been to explore the application of techniques of relative risk and odds ratio analysis to our educational data. Reporting relative risk (or reduced relative risk) is very common in the popular press as well as in the scientific literature in other fields, and so it seems worthwhile investigating it for our context. There are several advantages, which relate to the ease of interpreting the change in risk and the way in which it provides an alternative presentation of results in possibly a more memorable form. Contrast these two statements: *An S student has an extra 30% to 40% chance of becoming an expert, compared with an L student*, with, *The rate that S students become experts (27%) is 7% more than the rate of L*

students becoming experts (20%). The difficulty of describing the last absolute, rather than relative, result highlights the inadequacy of ordinary language in distinguishing absolute and relative change, especially when it is a change in rate or percentage that is being discussed. Analysing relative risk approach has advantages here, along with providing confidence intervals.

We expect that some members of the mathematics education community will be uncomfortable when we draw upon the medical context for research methods, even to analyse results. It is inherent in applying these concepts, to take an undesirable outcome (such as a disease or even death) as an implied metaphor for mathematical error or misunderstanding. When using any metaphor, different aspects of the metaphor will be carried across to the target situation by different people. Our position is that we can focus on the positives, as mathematical error as something to be overcome by joint effort of student and teacher. Other people may feel some discomfort in the use of techniques from medical research because of concerns about the way in which medical research has been simplistically held up as the "gold standard" for educational research in debates on funding principles in the USA (NCTM Research Advisory Committee, 2003). We contend that choice of methodology or data analysis techniques should not be judged by political or social associations, but by scientific reasoning. On the other hand, terminology needs to be chosen with sensitivity to the social needs in the area of application. Analysis of odds ratio and relative risk seems to have much to offer, although the language with which they are expressed needs modification.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. John Wiley: New York.

Bulmer, M. (2005). *A Portable Introduction to Data Analysis*. 3rd edition. The University of Queensland Printery.

Capraro, R. M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. *Journal for Research in Mathematics Education*, 35(1), 57 – 62.

Foxman, D., Ruddock, G., Joffe, L., Mason, K., Mitchell, P & Sexton, B. (1985). *A Review of Monitoring in Mathematics 1978 to 1982. (Vol. 1).* London: Dept of Ed.& Science.

Newcombe, R. (2003). DVT plane risk may be lower than thought. BUPA Health news. http://www.bupa.co.uk/health_information/html/health_news/ Accessed 11 Jan 2005.

NCTM Research Advisory Committee. (2003). Educational research in the No Child Left Behind environment. *Journal for Research in Mathematics Education*, 34(3), 185-190.

Reynolds, M. (n.d.) Air Travel: Runners at Risk? Accessed 11 Jan 2005, from http://www.runnersworld.com/article/0,5033,s6-188-0-0-1392,00.html

Steinle, V., & Stacey, K. (2003). Grade-related trends in the prevalence and persistence of decimal misconceptions. In N.A. Pateman, B.J. Dougherty & J. Zilliox (Eds.), *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 259 – 266). Honolulu: PME.

Steinle, V. (2004). *Changes with Age in Students' Misconceptions of Decimal Numbers*. Unpublished PhD, University of Melbourne, Melbourne.