The Mathematics Enthusiast

Volume 6
Number 1 *Numbers 1 & 2*

Article 17

1-2009

COMPARISION OF HIGH ACHIEVERS WITH LOW ACHIEVERS: Discussion of Juter's (2007) article

T. P. Hutchinson

COMPARISON OF HIGH AND LOW ACHIEVERS: A Discussion of Juter's (2007) article[1]

T. P. Hutchinson[2]
Centre for Automotive Safety Research, University of Adelaide, Australia

Abstract: *The use of questionnaires and interviews to compare the responses to a mathematical task of high achievers with low achievers has limitations. The partial information that they have provides a way of comparing high and low achievers. Some references are given here to relevant task formats and theories. An example is given of how examinees' performance with unusual task formats (specifically, answer-until-correct) may be analysed to throw light on the mathematical description of partial information.*

Keywords: answer until correct tasks; research methodology; questionnaire analysis; task formats;

1. Introduction

Juter (2007) compared high achieving students with low achieving students in respect of performance on problems concerned with limits of functions. Juter made use of questionnaires and interviews, and results were presented in the form of examples of responses given by high achievers and by low achievers. Presenting results in that way, and concluding that "high achievers have richer concept images" and their "abstraction abilities were more highly developed" (Juter, 2007, p. 64) does have some interest. But these descriptions do not say much more than that students who knew more about limits did, indeed, know more about limits --- the attempt at analysis is almost circular. Section 2 below will suggest some ways of comparing high achievers with low achievers that avoid this circularity. Section 3 gives an example of how empirical results (specifically, in an answer-until-correct task) may be compared with theories.

2. Discussion of Juter (2007)

The concern noted above may also be expressed as follows. On any single performance measure, high achievers are likely to score better than low achievers. This is unlikely to be of great interest on its own. An interesting research question is likely to involve two measures, and to concern

[1] http://www.math.umt.edu/TMME/vol4no1/TMMEv4n1a3.pdf

[2] email paul@casr.adelaide.edu.au
phone +61 8 8303 5997
fax +61 8 8232 4995

how the between-group difference on one measure relates to the between-group difference on another measure. Exactly what these questions are and how they can be answered will naturally depend on what the different observed measures are; an example will be given in Section 3.

Instead of taking a numerical difference, one might put high achievers and low achievers on an equal footing in one respect, and then compare something else: instead of comparing responses by high achievers with responses by low achievers, compare wrong responses by high achievers with wrong responses by low achievers, and (separately) compare correct responses by high achievers with correct responses by low achievers.

- Do high achievers and low achievers differ in the wrong responses they give? In cases where one wrong response is considered less wrong than another, do high achievers tend to give the less wrong response? (Juter refers to embodied, proceptual, and formal modes of mathematical thinking. This might be the basis for classifying one wrong response as less wrong than another.) If a second attempt is permitted following a wrong response, do high achievers tend to do better than low achievers?
- If an explanation of a correct response is asked for, is its quality better for high achievers than for low achievers? If confidence in a correct response is asked for, is it higher for high achievers than for lows achievers?

Once like is being compared with like, then it is reasonable to ask about richness of concept or abstraction, provided they can be operationalized and measured.

Most research into different wrong responses, performance at second attempt, and supplementary questioning has been based on multiple-choice items. Selection of different wrong responses by different ability groups in multiple-choice tests is discussed by Green et al. (1989), Price (1964), Wainer (1983), Wainer et al. (1984), and Hutchinson (1991, Sections 5.17, 8.6, 9.3, 9.4). When responses are generated (constructed) by examinees, there are often so many possible wrong responses that it is difficult to aggregate and classify them. However, Cairns et al. (2002) found evidence that some wrong responses are disproportionately generated by examinees of high ability and others disproportionately generated by examinees of lower ability.

As well as the nature of wrong responses and performance at second attempt, other topics that have been studied include performance when “don’t know” is one of the available options, performance when “none of the above” is one of the available options, performance when none of the available options are correct, performance when more than one option is correct, the changing of responses by examinees, and the confidence that the examinee expresses. Unusual task formats are sometimes considered to have both practical utility and psychological interest, in that they reveal more about the examinee than “choose the one correct, or best, answer” does. But the practical utility is debatable, as administration time tends to be longer --- if extra time is feasible, it might be better to set more items of conventional format. (Concerning confidence, it may be noted that although there is some plausibility in the idea that high achievers will tend to know they are correct, and low achievers will not, there are great complications: people may differ in how they use the scale of confidence, how well they know themselves, and how honest they are in reporting. For example, among the four students discussed by Juter, 2005, it was the best student who was the only one who was unsure whether she had control over the notion of a limit.)

Two further comments are worth making. (a) Some researchers have asked examinees to describe their thought processes. Unfortunately, it is particularly difficult to say anything about low achievers. Williams and Jones (1972) reported an interview survey of 15 schoolchildren who had taken a mathematics test, but found that not much information could be gleaned from the weaker students. See also Section 5.18 of Hutchinson (1991). (b) As previously noted, examinees' performance with unusual task formats holds some value for psychological theory. I urge those who have used such task formats to look carefully at the resulting dataset for any implications it may hold. Examples of comparing datasets with a theory that seeks to operationalize the notion of partial information are in Chapters 6 and 7 of Hutchinson (1991), and another is given in Section 3 below. (However, the data typically need to be aggregated --- e.g., over all examinees within a certain band of abilities --- and it is not certain that what is seen at the aggregate level is also the case for an individual examinee. For other limitations of the approach, see Chapter 8 of Hutchinson, 1991.)

Thus it seems that methods concentrating on individual examinees (discussing responses to particular questions, as Juter did, or asking about thought processes), and methods that employ large samples and aggregated data, each have strengths and weaknesses.

3. Example of quantitative study of partial information

Suppose there is data on examinees' performance with an unusual task format. Sometimes a simple feature of the data is directly of interest. For example, is second-choice performance only at the chance level? Or, do the proportions with which different incorrect options are selected differ when examinees are grouped according to ability? On other occasions, a quantitative prediction is the centre of attention, as in the following example.

In answer-until-correct (AUC) tests, the examinee is given immediate feedback as to whether the response is correct; if it is wrong, then the examinee chooses another option, and again is given immediate feedback; the examinee continues until the correct option is chosen, then moves on to the next item. The dataset to be discussed is from Abplanalp (1995). That paper had much about the practicalities of AUC testing, and some interesting data, but lacked any theory to give context to the data. Consider the relationship between the number of errors when the test is scored conventionally and when using the AUC method. Figure 1 shows Abplanalp's data, which was from a test of 22 items having 5 options each, taken by 74 examinees. The horizontal axis shows the average number of wrong options chosen per item when using the AUC format, and the vertical axis shows the proportion of items answered correctly at first attempt.

Let y be the probability of answering correctly at first attempt, and x be the average number of wrong options chosen per item. Further, let m be the number of options per item. The limits on the relationship between x and y are as follows.

- If, whenever a second attempt is needed, the examinee is always correct at second attempt, $x = 1 - y$.
- If the examinee always chooses the correct option *last* whenever it is not chosen first, $x = (m-1)(1-y)$.

The two extremes are shown as dashed lines in Figure 1, as in Abplanalp's Figure 3.

The simplest theory for the relationship between x and y is based on assuming that the examinee's subsequent attempts are equivalent to random guesses whenever the first choice is wrong. Then $x = m(1-y)/2$. When m is 5 (as in Abplanalp's test), this leads to $y = (5 - 2x)/5$, and this is the straight line in Figure 1. It can be seen that most of the data points lie below this. That is, the examinees, if they are wrong at first attempt, take fewer attempts to find the correct response than they would if they had no knowledge. (At any given y, we can look across and see that the data points have a smaller x than would be expected if the examinees had no knowledge.) We might say the examinees have some degree of *partial information* about the item.

Alternative predictions arise from the following approach. (For more details, see Hutchinson, 1982, 1991, 1997.) Suppose that the examinee considers each option within each item, and that each option within each item gives rise to some feeling as to the degree to which it fails to match the question posed. At first attempt, the examinee will choose the option generating the lowest feeling of mismatch. If the first choice turns out to be wrong, so that a second attempt is necessary, the option generating the second-lowest feeling of mismatch is chosen. And so on. Now suppose that the mismatch for the correct options is taken from some probability distribution, and that the mismatches for the wrong options are taken independently from some other probability distribution. The distribution for the wrong options will have a higher mean than that for the correct options. Indeed, the difference between the means is a measure of the examinee's ability. Let the probability of the mismatch exceeding z be F(z) for correct options and G(z) for wrong options. Further, let f be the probability density of mismatch for correct options, $f = -dF/dz$.

- The probability of being correct at first attempt is the probability that the mismatch from the correct option is some value z, multiplied by the probability of all of the mismatches from the wrong options (there are m - 1 of them) being greater than z, integrated over all z.
- If the mismatch from the correct option is z, the proportion of mismatches from wrong options that are less than z is 1 - G(z), and the expected number of them is (1 - G(z)) (m-1). Averaging over different values of z is achieved via another integration.

Let λ be some measure of how different G is from F, that is, a measure of the examinee's ability --- it might, for instance, be the difference between the means of the distributions. Once an assumption has been made about what F and G are, the integrations referred to above lead to an equation for y in terms of λ and an equation for x in terms of λ. Then the equation for y in terms of x is obtained by elimination of λ.

To get a definite prediction, it is necessary to make some specific assumption about what F and G are. Three examples that are easy algebraically are as follows.

- Exponential distributions. Here, mismatch is taken to have an exponential distribution with mean 1 in the case of a correct option, and an exponential distribution with mean λ (this being greater than 1) in the case of wrong options. In the case of m = 5, this leads to $y = (4 - x)/(4 + 3x)$.

- All-or-none knowledge. Mismatch is taken to have uniform distributions, with the upper end of the range being the same for the wrong options as for the correct option. For m = 5, y = (5 - 2x)/5, as given earlier.
- Recognisable distractors. Mismatch is taken to have uniform distributions, with the lower end of the range being the same for the wrong options as for the correct option. For m = 5, $y = [1 - (1 - x/2)^5]/(5x/2)$. This is quite the opposite to the previous model, in the sense that now a wrong option is sometimes recognised as being wrong, but the correct option is never positively identified as such. This assumption has found occasional application in the psychological literature (Murdock, 1963; see also Section 4.6 of Hutchinson, 1991).

It may be asked how different are the three models, and whether Abplanalp's data favour one in preference to the others. The relationships between y and x are plotted in Figure 1. It appears that examinees have some degree of information when they give a wrong answer initially, but that it is rather less useful than is implied by the "recognisable distractors" and "exponential" theories.

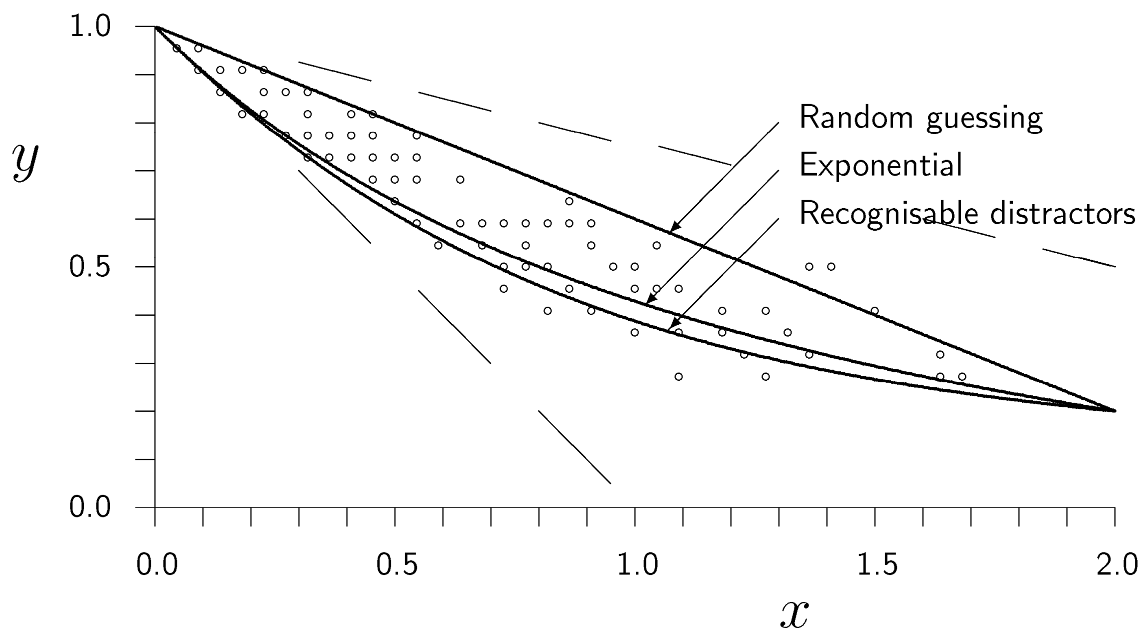

Figure 1. Data from Abplanalp (1995) compared with the predictions of three theories; y is the probability of answering correctly at first attempt, and x is the average number of wrong options chosen per item

Acknowledgements

The Centre for Automotive Safety Research, University of Adelaide, receives core funding from the Motor Accident Commission (South Australia) and the Department for Transport, Energy and Infrastructure (South Australia). The views expressed in this paper are those of the author, not necessarily of the University of Adelaide or the sponsoring organizations.

References

Abplanalp, P. (1995). Application of latent image technology to the multiple-choice test format. Optometric Education, 20, 118-123.

Cairns, D., Chekaluk, E., & Hutchinson, T. P. (2002). The wrong responses to a very difficult item: A comparison of high-scoring and low-scoring examinees. International Journal of Mathematical Education in Science and Technology, 33, 839-842.

Green, B. F., Crone, C. R., & Folk, V.G. (1989). A method for studying differential distractor functioning. Journal of Educational Measurement, 26, 147-160.

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. British Journal of Mathematical and Statistical Psychology, 35, 71-89.

Hutchinson, T. P. (1991). Ability, Partial Information, Guessing: Statistical Modelling Applied to Multiple-Choice Tests. Adelaide: Rumsby Scientific Publishing.

Hutchinson, T. P. (1997). Mismatch models that permit partial information to be shown. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory (pp. 481-494). New York: Springer.

Juter, K. (2005). Limits of functions --- How do students handle them? Pythagoras, 61, 11-20.

Juter, K. (2007). Students' conceptions of limits: High achievers versus low achievers. The Montana Mathematics Enthusiast, 4, 53-65. Retrieved from http://www.math.umt.edu/TMME/vol4no1/TMMEv4n1a3.pdf

Murdock, B. B. (1963). An analysis of the recognition process. In C. N. Cofer & B. S. Musgrave (Eds.), Verbal Behavior and Learning: Problems and Process (pp. 10-22). New York: McGraw-Hill.

Price, D. B. (1964). A group approach to the analysis of individual differences in the randomness of guessing behavior on multiple-choice tests and the development of scoring methods to take such differences into account. M.Ed. thesis, Rutgers, The State University, New Brunswick, New Jersey. Circulated as Research Bulletin 64-59, Educational Testing Service, Princeton, New Jersey.

Wainer, H. (1983). Pyramid power: Searching for an error in test scoring with 830,000 helpers. American Statistician, 37, 87-91.

Wainer, H., Wadkins, J. R. J., & Rogers, A. (1984). Was there one distractor too many? Journal of Educational Statistics, 9, 5-24.

Williams, S. I., & Jones, C. O. (1972). A comparison of interview and normative analysis of mathematics questions. Test Development Report 71-4, Educational Testing Service, Princeton, New Jersey.