Commentary

Low-Confidence Responses on the Vocabulary Size Test

T. P. Hutchinson

Centre for Automotive Safety Research, University of Adelaide
doi: http://dx.doi.org/10.7820/vli.v04.2.hutchinson

Abstract

McDonald and Asaba (*Vocabulary Learning and Instruction*, 2015) reported an administration of the *Vocabulary Size Test* that was modified to include "I don't know" as a fifth response option on all items, and in which participants later responded to the items originally marked as "I don't know". McDonald and Asaba were inclined to favour the score calculated without the later (reluctant or low-confidence) responses. It is argued here that this goes beyond the data. In many educational contexts, strong encouragement is given to respond when unsure, as examinees to have a better-than-chance probability of being correct, and will be disadvantaged if they do not respond.

In a vocabulary or other educational test, how important are correct responses given with low confidence? I would like to add a few comments to those of McDonald and Asaba (2015) in contrasting the performance of two students who they referred to as Rena and Risako. The test was the *Vocabulary Size Test* (Nation and Beglar, 2007) with responses in Japanese, modified to include "I don't know" as a fifth response option on all items. The four participants were instructed not to guess if unsure, but to select "I don't know" instead. On a second pass through the test, participants answered the items originally marked as "I don't know". It seems reasonable to refer to this second set of attempts as low-confidence or reluctant responses. McDonald and Asaba also interviewed the four participants about their reasoning in selecting responses in the second pass. I regard empirical studies using unusual response formats as very valuable. The research by McDonald and Asaba qualifies for two reasons—distinguishing responses according to level of confidence, and introspection by the participants.

1 Reconsideration of Results

Rena scored 84 without guesses, and 91 with all responses included. Risako scored 54 without guesses, and 89 with all responses included. Two further participants, Rika and Mari, respectively scored 51 and 50 without guesses, and 83 and 81 with all responses included. McDonald and Asaba are rightly cautious about giving too much weight to one type of score (e.g., without guesses) rather than another (e.g., all responses), but nevertheless they rather favour the score

without guesses. In both the Abstract and the Conclusion of the paper, they say there are much greater differences between types of score for the lower proficiency learners, and so they must be accepting the scores (without guesses) of 54 for Risako, 51 for Rika, and 50 for Mari as valid indicators of lower proficiency.

My opinion is that in giving precedence to the score without guesses, McDonald and Asaba are going beyond their data. In many educational contexts, there is strong encouragement of examinees to respond when unsure, because examinees usually have a better-than-chance probability of being correct when unsure, and they will be disadvantaged if they do not respond. This disadvantage may be greater for some personality types than others. It is widely thought, I believe, that strong encouragement to examinees to respond to all items is a simple way of minimising the penalty for not guessing that is otherwise imposed on some people. The arguments otherwise of McDonald and Asaba are not strong.

Table 1 is a reorganisation of results in Tables 2 and 3 of McDonald and Asaba (2015). It shows, for each of the participants, the numbers of responses of each type. In particular, Rena confidently answered 33 items wrongly. (This is calculated as 140 items, minus 23 "I don't know" responses, minus 84 confident correct responses = 33 confident wrong responses.) With the data arranged like this, a striking story can be seen.

(1) Confident responses. Rena gets a lower proportion correct than the other three participants.
(2) Confident responses. For all participants, some were wrong, despite the instruction to select "I don't know" if unsure.
(3) Guess responses. Rena gets a lower proportion correct than the other three participants. Rena's responses were correct barely above the chance level (7 out of 23 is 30%, compared with a chance level of 25%). Other participants showed partial knowledge, in the sense of a better-than-chance probability of being correct (47%, 38%, and 38%).
(4) Rena responded at first pass (i.e., confidently) more frequently than the other participants.

Table 1. Confident Responses and Guesses: Correct and Wrong Answers

	Confident responses		Guess responses	
	Correct	Wrong	Correct	Wrong
Rena	84	33	7	16
Risako	54	11	35	40
Rika	51	4	32	53
Mari	50	9	31	50

Note: Confident refers to responses given on the first pass through the test; guess refers to responses given on the second pass to items initially marked as "I don't know". A guess in this sense may be based on some knowledge, it is not necessarily a random guess.

Above-chance performance in items answered with low confidence is a common finding in the literature (Hutchinson, 1982).

Rena was the participant who scored highest, yet at each level of confidence, her probability of being correct was the lowest of all four participants. It may be that Rena differed from the other participants in how she used the term "unsure"—McDonald and Asaba say that "Participants were explicitly instructed not to guess on items they were unsure about, but to select 'I don't know' in these instances instead". That is, she may not be more proficient than Risako, but instead be more confident (or more willing to claim confidence) in her responses. Educational testing is usually intended to measure proficiency (or achievement, ability, aptitude, etc.), and not personality traits and states. I expect that there are some occupations for which Rena's personality and behaviour are advantageous, but I also expect there are others for which people like Risako are better suited.

2 Discussion

McDonald and Asaba had participants express a level of confidence in their responses, and also report on their thought processes. Various other formats have been used from time to time to try to estimate either partial knowledge or willingness to respond (e.g., answer-until-correct, "None of the above" as a response option, and nonsense items). Data from such formats are used best if there is some theory available—a theory with more psychological content than Item Response Theory has, and yet is broad-brush rather than requiring detailed analysis of each item individually (Hutchinson, 1982). It may be over-ambitious to seek quantitative comparison of data with theory, but possible to look for qualitative features in the data, such as above-chance probability of correctness with reluctant responses (as in McDonald and Asaba) and with second attempts after a first response is wrong.

Acknowledgements

References

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*(1), 71–89. doi:10.1111/j.2044-8317.1982.tb00642.x

McDonald, K., & Asaba, M. (2015). "I don't know" use and guessing on the bilingual Japanese Vocabulary Size Test: A preliminary report. *Vocabulary Learning and Instruction*, *4*(1), 16–25. doi: 10.7820/vli.v04.1.mcdonald.asaba

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. Retrieved from http://jalt-publications.org/tlt/issues/2007-07_31.7