The Canadian Journal of Program Evaluation Vol. 13 No. 2 Pages 123–128
ISSN 0834-1516

MEASURING THE CONSISTENCY IN CHANGE OF HEPATITIS B KNOWLEDGE AMONG THREE DIFFERENT TYPES OF TESTS: TRUE/FALSE, MULTIPLE CHOICE, AND FILL IN THE BLANKS TESTS

Vic Sahai
Petra Demeyere
Sheila Poirier
Felice Píro
Sudbury District Health Unit
Sudbury, Ontario

Abstract: This Research and Practice Note assesses grade 7 students' recall of hepatitis B information using three different types of tests: short answer, true-false, and multiple choice. It concludes that short answer testing is the most reliable and recommends that evaluators be consistent with the type of test used to assess pre- and post-knowledge of students.

Résumé: Cette Note de recherche et de pratique évalue les connaissances d'élèves de septième année sur l'hépatite B à l'aide de trois types de contrôle: réponses courtes, vrai-faux et réponses à choix multiples. La conclusion est que les interrogations demandant des réponses courtes constituent le contrôle le plus fiable et que les évaluateurs devraient utiliser le même type pour évaluer les connaissances préalables et subséquentes des élèves.

Evaluators of school health education need to know if teaching time spent is justified and if resources are effective. Often pre- and post-tests are used to evaluate change in knowledge. Many different types of tests exist, each with their own strengths and weaknesses. Varying test types may assess knowledge differently and may therefore produce different results. For example, true/false questions assess the abilities of the students to indicate the correctness of statements or facts (Midwood, O'Connor, & Simpson, 1994). Likewise, multiple choice questions require the student to recognize the appropriate answer out of a selection. In both of these circumstances, the student is not required to produce information, but rather to judge provided answers as correct or incorrect. Also, in both of these tests, students can achieve correct answers just by guessing. This,

then, may not reflect knowledge learned. Short answer tests, on the other hand, mainly evaluate factual recall that may reflect a greater understanding of the material learned (Midwood et al., 1994). Due to different test characteristics, before and after differences in test scores may be partially explained by the type of test rather than by the improvement in knowledge.

As part of the Ontario Ministry of Health Hepatitis B Vaccine Program, public health nurses are required to prepare and deliver health information to Grade 7 students. To evaluate the effectiveness of the educational component, the nurses used three types of tests: multiple choice, true/false, and short answer. The objective of this paper is to determine if different types of tests given to Grade 7 students involved in a Hepatitis B educational program are consistent in measuring the students' pre- and post-knowledge of Hepatitis B, and to determine whether different types of testing show differing results in short-term recall. It is hoped that the results of this study will assist evaluators in choosing between testing options to measure change in knowledge.

METHODS

The educational component alluded to above involved the presentation of a standard video, "Hep B is a Bad Rap," as well as a standardized information session, prepared and given by six public health nurses. To test short-term recall about Hepatitis B, the Grade 7 students were given one of three types of tests: multiple choice (MC), true/false (TF), or short answer (SA). The multiple choice test consisted of 10 questions of which 4 questions offered three answer choices and the remaining 6 offered four answer choices. Likewise, the true/false test was made up of 10 questions with three answer choices in each, namely: "true," "false," and "don't know." The short answer test was made up of 10 questions, each requiring one-word answers or a short explanation of one sentence.

Each class was randomly assigned one of the three test types. Within a time span of approximately 1.5 hours, the students participated in a pre-test, followed by the instructional session and the post-test. The pre- and post-tests were not identical, but of the same type. For example, a class given a multiple choice pre-test would later receive a multiple choice post-test. To preserve confidentiality, the students were assigned a number to be entered on the pre- and post-test documents.

Reliability Testing

In order to ensure that the pre- and post-tests were of equal difficulty, the test-retest method was used to assess the reliability of the MC, TF, and SA tests. Six classes of 30 students were chosen to pilot the three test types. The test types were assigned randomly to the classes. No educational component was given to these students. If the tests were equally difficult, the students would be expected to score the same on the pre- and post-tests. The students were asked to write both the pre- and the post-tests, one after another, in any order they chose (i.e., they could write the post-test first and immediately thereafter the pre-test, or vice versa). The pre- and post-test types were consistent, that is, a student writing an SA pre-test also wrote an SA post-test. The pre- and post-tests were found to be of equal difficulty as the students scored similarly on both tests. Table 1 gives the correlation coefficients for each test.

An ANOVA test was used to determine if the difference in the mean scores of the three pre-tests was statistically significant. This same procedure was then applied to the three post-test scores. Scores were considered to be different from one another if the difference was statistically significant at the 95% confidence level. Finally, paired T-tests were used to examine the mean differences between the pre- and post-tests for all three test types.

RESULTS

Justification for Combining Linguistic Groups

Pre- and post-test results were analyzed separately at first for English and French students. The results are presented in Table 2. As indicated there, the scores for the English and French students differed significantly. The trends, however, are the same for both

Table 1
Correlation Coefficients for Each Test Type

Test type	Correlation coefficient	*p*-value
Multiple choice	0.41	0.023
True/false	0.38	0.08
Short answers	0.83	0.0001

populations. It is for this reason that we were able to combine English and French student results and thereby allow for a greater sample size in the final analysis.

The table also portrays a difference in test scores between the three groups ($p > 0.05$). Part of the difference in test scores may be explained by the ability of the test to measure the same level of knowledge; therefore, mixing test types is not advisable.

Combined Results for English and French Students

An Analysis of Variance was used to test the mean scores of the three different test types. As shown in Table 3, the mean scores of the pre-test for the MC, TF, and SA were 6.70, 4.70, and 1.71 out of 10 respectively ($p < 0.0005$). Similarly, the post-test scores of the MC, TF, and SA were 8.86, 8.69, and 8.20 out of 10 respectively ($p < 0.0005$). A multiple comparison of the group means using a Duncan's Multiple Range Test (SPSS for Windows Base System User's Guide Release 6.0, 1993) suggested that all the means were significantly different for the pre- and post-test scores.

Changes in the pre- and post-scores for each test type are also displayed in Table 3. The greatest improvement was found in the short answer test.

Table 2
Comparison of Pre- and Post-Test Mean Scores* Between French and English Students for Multiple Choice, Short Answers, and True/False Tests

		French (*N*)	English (*N*)	Percent difference out of 10
Pre-test	Multiple choice	5.56 (276)	6.72 (503)	11.6 % (S)
	Short answers	1.51 (212)	1.86 (576)	3.5 % (S)
	True/false	4.84 (244)	4.69 (681)	1.5 % (NS)
Post-test	Multiple choice	8.17 (276)	8.97 (503)	8 % (S)
	Short answers	7.40 (212)	8.25 (576)	8.5 % (S)
	True/false	7.89 (244)	8.64 (681)	7.5 % (S)

*All scores are out of 10
(S) T-test scores were statistically significant
(NS) T-test scores were *not* statistically significant

DISCUSSION

It should be recognized that a significant difference between pre-test and post-test scores occurred with all three test types. The short answer test had the best correlation coefficient (Table 1) and thus was the most reliable of the three tests. It is also apparent from the test results that the greatest difference in pre- and post-test scores occurred with the SA test. This type of test differs from the MC and TF tests in that it requires the generation rather than just the selection of an answer. The SA type of test reduces the possibility of the student simply guessing the correct answer. This is therefore the most difficult of all three tests. Multiple choice is the second most difficult test since it requires the student to know both what is right and what is wrong. In theory, the easiest of the three tests is the TF test, as the student has a 50% chance of guessing the correct answer (Midwood et al., 1994). Therefore, one would expect the change in knowledge in the MC test to be second largest and the smallest change in knowledge to be shown by the results of the TF test. For example, Harris and Changas (1994) found a significant difference between the true-false and multiple-choice test score. Out of a possible 25, the mean score for the group who took the true-false format was 14.83 (SD = 2.36), whereas the mean score for the group who took the multiple-choice version was 10.98 (SD = 2.65), t (373) = 14.81, $p < 0.01$. The authors suggest that the multiple-choice format was a more difficult test because a respondent's chances of answering a question correctly by guessing were greatly reduced. Our results show the reverse of Harris and Changas's findings. That is, the results actually point to a greater change in knowledge among the TF tests. This can be accounted for by the design of these two tests. In the TF test, the student actually had three choices, namely, "true", "false,"

Table 3
Pre- and Post-Test Scores for each Test Type (Linguistic Groups Combined)

	Pre-test scores out of 10	Post-test scores out of 10	Differences between pre- and post-test scores as a percent	*t*-test values (paired *t*-test)	Probability values
Multiple choice	6.70	8.86	21.6	30	$p < 0.0005$
True/false	4.70	8.69	38.3	50.6	$p < 0.0005$
Short answers	1.71	8.20	64.9	83.1	$p < 0.0005$

and "don't know." Any "don't know" answers were then lumped with the "incorrect" answers. The MC test, on the other hand, did not offer the "don't know" option to the student, therefore forcing the student to guess an answer when it was in fact "not known." This would have increased the pre-test score, thereby decreasing the overall change in pre- and post-test scores. Future studies should include "don't know" as an option in multiple choice tests in order to more accurately assess what the student knows or thinks he/she knows.

LIMITATIONS OF THE STUDY

There are limitations to this study. We cannot be sure that the nurses did not have an effect on the outcomes of the test scores. We tried to randomize the different tests, but with only six nurses this number was not enough to ensure that the nurses themselves did not have an effect on the outcomes.

Table 2 showed the number of students who took the three different types of tests by language. Although the schools and classes were randomly chosen, the representativeness of the Francophone population can be questioned because of the relatively small numbers.

CONCLUSION

The above results clearly show that there is a difference in pre- and post-test scores as evaluated by the three test types. Therefore, SA, MC, and TF tests are different and assess knowledge differently. Hence, mixing two different types of tests up for assessing pre- and post-knowledge would not provide the evaluator with a correct assessment of the actual knowledge gained by the student. Based on the findings of this study, it is advisable for evaluators to be consistent with the types of tests they use when assessing students' pre- and post-knowledge.

REFERENCES

Harris, D., & Changas, P. (1994). Revision of Palmore's second facts on aging quiz from a true-false to a multiple-choice format. *Education Gerontology, 20*, 741–754.

Midwood, D., O'Connor, K., & Simpson, M. (1994). *Assess for success* (p. 84). Toronto: Ontario Secondary School Teachers' Federation.

SPSS for Windows Base System Users' Guide Release 6.0. (1993). Chicago: SPSS Inc., p. 278.