ED 431 813	TM 029 933
AUTHOR	Gadalla, Tahany M.
TITLE	Multiple-Choice versus Constructed-Response Tests in the Assessment of Mathematics Computation Skills.
PUB DATE	1999-04-00
NOTE	12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Computation; *Constructed Response; Elementary Education; *Elementary School Students; Foreign Countries; *Mathematics Tests; *Multiple Choice Tests; *Test Construction; Test Results
IDENTIFIERS	Canada

ABSTRACT

The equivalence of multiple-choice (MC) and constructed response (discrete) (CR-D) response formats as applied to mathematics computation at grade levels two to six was tested. The difference between total scores from the two response formats was tested for statistical significance, and the factor structure of items in both response formats was compared. Responses of 1,028 students in grades 2 through 6 on the Canadian Achievement Tests were analyzed. Stem-equivalent and scoring-equivalent MC and CR-D forms were used, and each student was tested twice with a time lapse between testings. Analyses show that response format has a significant effect on the performance of students in grades 2 and 3 with the mean score on the MC format higher than the mean score on the CD-R format. Response format is not found to have a significant effect on the performance of students in grades 4, 5, and 6. (SLD)

Multiple-Choice versus Constructed-Response Tests in the Assessment of Mathematics Computation Skills

Tahany M. Gadalla

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in Montreal Canada
April 1999

Multiple-Choice versus Constructed-Response Tests in the Assessment of Mathematics Computation Skills

Abstract for AERA 1999

Tahany M. Gadalla

Ontario Institute for Studies in Education/U of Toronto

1. Introduction

Most standardized educational testing has used multiple-choice (MC) response format. However, in recent years, there has been increasing concern that MC tests may be too limited in the skills they tap. As a consequence, alternative response formats have been developed and many are being used in several educational contexts (Bennet and Ward, 1993), a fact which sparked renewed interest in the enduring question of whether tests of the same content that employ different response formats measure the same traits. For example, many empirical studies on the equivalence of multiple-choice and constructed response (Discrete) (CR-D) formats have been reported. However, their results have not been conclusive and many were seriously flawed in design and analysis (Traub and MacRury, 1990). In general, these results suggest that MC and CR-D tests of the same content cannot be assumed to be equivalent and that format effect is not uniform across subject matters. It is also conceivable that format effect is not uniform across ages of examinees. With regard to subject matter, Traub (1993) concludes that for the quantitative domain, the two formats probably do not measure different traits.

In the math computation domain, it is hypothesized that, regardless of the format, items will require the calculation of the answer and that answers to math computation items will not be recognized by most examinees when answering a MC test. In other words, a MC and a CR-D forms of the same stem will be processed in the same way. Nevertheless, it has generally been assumed that correct answers to MC items can be guessed at more readily than CR-D items, it is thus expected that MC tests are less difficult, less discriminating and less reliable than CR-D tests of the same content. In addition, having multiple answers - one of which is the correct one - may alert the examinee who makes a mistake in the computation and ends up with an answer which is not on the list of choices, to check and/or redo the computation. Such guidance is not available with the CR-D format and can result in the MC format to have reduced relative difficulty. However, these expectations are not consistently supported by findings of empirical research (Traub & MacRury, 1990).

2. Objectives

The main purpose of this study is to test the equivalence of MC and CR-D response formats as applied to mathematics computation at grade levels two to six. This is carried out in two steps. First, the difference between total scores from the two response formats is tested for statistical significance and the factor structure of the items in both response formats is compared. Second, if, based on results obtained in the first analysis, we fail to reject the hypothesis that the two response formats measure the same traits, their relative difficulty and reliability will be compared.

3. Data and Methods

Data for this study consist of the responses of 1028 students in grades two to six to the mathematics computation component of the Canadian Achievement Tests, Second Edition (CAT/2) (Canadian Test Centre, 1992). Stem-equivalent and scoring-equivalent MC and CR-D forms were used and each student was tested twice with a time lapse of two to four weeks between the two testings. Students at each grade level were divided into four groups; two groups were retested with the alternate format and the other two groups were retested with the same format. That is, response formats were used in four testing sequences; namely, CR-D/CR-D, CR-D/MC, MC/MC and MC/MC. Nine schools from across Canada, five of them located in rural areas, participated in the study. Teachers administering the test were instructed not to review the first test material with their students, not to teach them to the test, and not to tell students that they will be writing a second test of the same content.

Generalized linear models (GLM) procedure in the statistical computing package SAS is used to carry out a repeated measures analysis of variance in which components of variance due to carryover effects and format effects are estimated and tested for statistical significance. Factors included in this model are entered in the following order: (1) testing sequence: which includes the four categories, CR-D/CR-D, CR-D/MC, MC/CR-D and MC-MC, (2) student: which is nested within testing sequence, (3) order: which indicates first versus second testing, (4) response format, and (5) carryover effect. Tests using the hierarchical and the unique sums of squares are carried out.

Scatter plots for the score on the first testing versus the score on the second testing are prepared and the linearity of their relationship examined for each testing sequence of each grade level. Paired t-tests of the difference between the scores on first and the second testing are carried out for each testing sequence at each grade level. Also, test-retest reliability and correlation coefficients corrected for attenuating effect of errors of measurement are calculated and compared to unity.

4. Significance of the Study

The question of equivalence of MC and CR-D response formats is far from being resolved. The present study is intended to shed some light on this enduring question. It is of importance to know whether or not different formats of the same stem measure different traits. It is of equal importance to know what traits are measured with each format. Traits measured in different tests inform students and teachers about the kind of knowledge and skills that are most important to learn and to teach and can thus have direct and indirect consequences for the educational system.

In studies with repeated-measurement design and one group of examinees, the CR-D format is usually administered first followed by the MC format. Thus, ignoring carryover effects from CR-D to MC format. On the other hand, studies in which examinees are divided into two groups, each tested with both formats in reversed orders are bound to make the assumption that carryover effects from MC to CR-D are equivalent to carryover effects from CR-D to MC. Such assumption is not supported in the literature. As Heim and Watts (1967) point out, carryover effects are likely to be asymmetrical with probably more carryover from MC to CR-D than from CR-D to MC. Only with a multiple group design in which all combinations of format sequences are present, such as the one used in this study, that a separation of carryover effects, order effects, and format effects can be achieved without having to impose such assumption.

The MC and the CR-D tests used in this study are stem-equivalent and scoring-equivalent. Thus, it can be safely assumed that the score scales are equivalent which makes it possible to compare their relative difficulty.

Most published reports describe studies on examinees at grade eight or higher. Studies on younger children are difficult to find. It is also conceivable that the magnitude of format effects and/or carryover effects varies with the age of examinees. This study covers a range of 5 years from grades two to six which makes it possible to infer about whether or not carryover effects and/or format effects are uniform across the age range under study.

5. Findings

Table 1 shows the distribution of the participating students by grade and test sequence. Table 2 includes correlation coefficients of total scores from the two testings for students who were retested with the alternate format, Cronbach's alpha coefficient for internal consistency and correlation coefficients corrected for the attenuating effects of errors of measurement. The correlations ranged between 0.7 and 0.85 and their corrected values ranged between 0.81 and 0.92 which are quite high. Reliability coefficients ranged between 0.83 and 0.9 for the MC tests and between 0.9 and 0.95 for the CR-D tests; being consistently higher than that of the

MC tests of the same content.

Differences between percent correct scores achieved in the first and the second testings were tested using paired t-tests. Table 3 includes only those test sequences in which differences were statistically significant. Three of four groups in grade 2 showed significant improvement in test scores, two groups in each of grades 3 and 4 and only one in grade 6. That is, effects of repeated testing (practice/recall/carryover effects) are greater at younger ages. Paired t-tests on scores from first and second testings for students in the second grade indicate a significant recall effect in both groups retested with the same response format, i.e. CR-D/CR-D and MC-MC, with mean scores in the second testing significantly higher than mean scores from the first testing (p-value < 0.0005 for each group). Recall effects are found to be statistically significant for those students in the third and fourth grades who took the testing sequence CR-D/CR-D. Put another way, the test sequence CR/CR resulted in significant carryover/recall effects at grade levels 2,3 and 4 while the test sequence MC/MC resulted in significant carryover effects in grade 2 only (It seems to me like 'I do and I remember'). The magnitude of the improvement however, is not the same across grade levels. Although mean scores of the second testing were also found to be significantly higher than mean scores of the first testing for students in grades two, three and six in the testing sequence CR-D/MC, the source of this difference can not be decided from this analysis.

Repeated measures analysis of variance indicate that, after adjusting for carryover effect and order effect, response format has a significant effect on the performance of students in grades two (p-value=0.0001) and three (p-value=0.0001) with the mean score on the MC format higher than the mean score on the CR-D format. Response format is not found to have a significant effect on the performance of students in grades four, five or six. The carryover factor has a significant effect on scores of students in grades two, three and four but not on scores of students in higher grades.

6. References

Bennet, R. E. & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Canadian Test Centre (1992). *Canadian Achievement Tests, Second Edition*. Canada.

Heim, A. W. & Watts, K. P. (1967). An experiment on the multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology*. 1967, 37, 339-346.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29-44). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Traub, R. E. & MacRury, K. (1990). Antwort-auswahl- vs freie-antwort-aufgaben bei lernerfolgs- tests [Multiple-choice vs. Free-response in the testing of scholastic achievement]. In K. Ingenkamp & R. S. Ja(..)ger (Eds.), *Tests und trends 8: Jahrbuch der pa(..)dagogischen diagnostik* (pp. 128-159). Weinheim, Germany: Beltz Verlag.

Table 1. Test Sequence versus Grade

Grade	CR/CR	CR/MC	MC/CR	MC/MC	Total
2	46	60	57	56	219
3	46	64	45	65	220
4	49	49	45	68	208
5	44	51	46	54	195
6	54	52	25	55	186
Total	239	273	218	298	1028

Computation component of the Canadian Achievement Test (CAT/2).

Table 2. Reliability and Correlations

Grade	Corr of MC and CR	Reliability		Corrected corr coefficient
		MC	CR	
2	0.70	0.83	0.90	0.81
3	0.82	0.86	0.93	0.92
4	0.79	0.88	0.94	0.87
5	0.85	0.90	0.95	0.92
6	0.74	0.88	0.94	0.81

* As one would expect, reliability of CR format is consistently higher than that of MC format of the same content.

* raw scores.

Reliability = Cronbach's alpha.

Table 3. Paired t-tests on Percent Correct Scores of 1st vs 2nd Testing, Significant Results Only

Grade	Test Sequence	Sign. level, P	N	Effect
2	CR(36.04)-CR(62.71)	<0.0005	46	carryover
	MC(69.02)-MC(77.40)	< 0.0005	56	carryover
	CR(56.03)-MC(62.24)	0.006	60	format and/or carryover
3	CR(43.03)-CR(54.41)	< 0.0005	46	carryover
	CR(60.66)-MC(69.81)	< 0.0005	64	format and/or carryover
4	CR(52.14)-CR(62.04)	< 0.0005	49	carryover
	MC(63.39)-CR(75.61)	< 0.0005	45	format and/or carryover
6	CR(70.96)-MC(77.02)	< 0.0005	52	format and/or carryover

Table 4. Results of GLM Repeated Measures ANOVA

(using adjusted sum of squares)

Grade	N	R^2	Effect	p-value
2	219	84.9%	format	0.0001
			carryover	0.0001
3	220	92.1%	format	0.0001
			carryover	0.0010
4	209	90.9%	format	0.4565
			carryover	0.0004
5	195	95.0%	format	0.9252
			carryover	0.7850
6	186	94.0%	format	0.3847
			carryover	0.6423

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029933

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Multiple-Choice versus Constructed-Response Tests in the Assessment of Mathematics Computation Skills.

Author(s): Tahany M. Gadalla

Corporate Source: Ontario Institute for Studies in Education / University of Toronto

Publication Date: April 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Tahany Gadalla

Printed Name/Position/Title: Dr. Tahany M. Gadalla

Organization/Address: 252 Bloor Street West
Toronto, Ontario M5S 1V6 Canada

Telephone: (416) 923-6641

FAX: (416) 926-4744

E-Mail Address: tgadalla@oise.utoronto.ca

Date: June 1, 99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com