DOCUMENT RESUME

ED 230 588 TM 830 368

AUTHOR McMorris, Robert F.; And Others
TITLE Effects of Including Humor in Test Items.
PUB DATE Apr 83
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Grade 8; Grammar; *Humor; Junior High Schools; *Multiple Choice Tests; Questionnaires; Scores; Test Anxiety; Test Construction; Test Format; *Test Items
IDENTIFIERS Iowa Tests of Basic Skills; *Parallel Test Forms

ABSTRACT

Two 50-item multiple-choice forms of a grammar test were developed differing only in humor being included in 20 items of one form. One hundred twenty-six (126) eighth graders received the test plus alternate forms of a questionnaire. Humor inclusion did not affect grammar scores on matched humorous/nonhumorous items nor on common post-treatment items, nor affect anxiety. Students favored humor inclusion on tests, judged effects of humor positively, and estimated humorous items to be easier. Humor did not lower performance but was sought by the students. Potential for more valid and humane measurement is discussed. (Author)

EFFECTS OF INCLUDING HUMOR IN TEST ITEMS

Robert F. McMorris, Paul D. Kaiser, Sandra L. Urbach,

Eileen P. Handin, and Michael C. Connor

State University of New York at Albany

ABSTRACT

Two 50-item multiple-choice forms of a grammar test were developed differing only in humor being included in 20 items of one form. One hundred twenty-six (126) eighth graders received the test plus alternate forms of a questionnaire. Humor inclusion did not affect grammar scores on matched humorous/nonhumorous items nor on common post-treatment items, nor affect anxiety. Students favored humor inclusion on tests, judged effects of humor positively, and estimated humorous items to be easier. Humor did not lower performance but was sought by the students. Potential for more valid and humane measurement is discussed.

Paper presented at the meeting of the National Council on Measurement in Education, Montreal, April 1983.

The past 15 years have been witness to a reincarnated interest in the use of humor as an aid for both teaching and evaluation. Typically, commentaries have emphasized that positive results can be obtained by incorporating it into instruction (Ball & Bogatz, 1971; Earls, 1972) and tests (Adams, 1972; Monson, 1968). Yet, as noted in a review by Goldstein and McGhee (1971), published literature pertaining to humor has not been voluminous enough to permit many firm generalizations or conclusions. Similarly, the suggestion that it facilitates learning has failed to gather unequivocal support in recent research efforts (Davies & Apter, 1980; Ziv, 1976).

Intuition allows the formulation of two equally sound arguments to account for this inconclusiveness. On the one hand, some well directed levity in instructional techniques and devices may generate positive affect and serve to attract student interest on selected topics. On the other hand, it can be stated that humor produces an "easy-going" and "loose" atmosphere which inhibits the realization and acceptance of the popular tenet that learning is the result of hard work.

Incorporation of humor into learning situations has been supported by early psychological accounts of its key role as a mechanism for reducing anxiety (Freud, 1928; Keith-Spiegel, 1972; Spencer, 1960). This purported property stimulated several more contemporary empirical efforts directed toward highlighting its value as a tension reliever. Studies investigating humor-mediated tension reduction have demonstrated that subjects exposed to manipulations of emotional arousal, leading to reported stages of anger or anxiety, show a significant decline in these states following subsequent exposure to humorous material (Dworkin & Efran, 1968; Singer, 1968).

Such conclusions seem relevant to educational testing, where anxiety has been shown to influence test performance (Hill & Sarason, 1966). It is currently generally accepted that high test anxiety has an inhibiting effect on performance. Since humor has been conceptualized as having tension-relieving properties, it is logical that several research efforts focused on its interjection into test situations.

Several investigators have considered the effects of humorous modifications of test material on the test performances of students differing in level of anxiety (e.g., Smith, Ascough, Ettinger, & Nelson, 1971; Terry & Woods, 1975; Townsend & Mahoney, 1981). Using a sample of undergraduate university students, Smith, et al. (1971) showed that humorous manipulations of the stems of multiple choice items significantly improved test performance for a group of "highly test-anxious" students. The test was a mid-term examination containing 30 items administered under standard classroom conditions. The humorous form contained a modification of every third item. The anxiety measure was the Test Anxiety Scale (Sarason, Pederson, & Nyman, 1968).

A study by Terry and Woods (1975), using third- and fifth-grade students, contrasted the performance of these two groups on matched humorous and nonhumorous versions of mathematical and verbal tests. Humorous questions were shown to restrict the third graders' performance on the mathematical test but had no significant effect on their verbal performance. Humor did not significantly alter the fifth graders' mathematical performance and had mixed effects on their verbal performance, improving it on the first task and inhibiting it on the second. The authors assumed that the third graders were less anxious than the fifth graders because the importance of educational

evaluation increases with age. They reasoned that the third graders could have started out below the optimal level of tension for all three tasks. It was supposed that the fifth graders started out with a higher level of tension. As humor became noticed, the tension reduction resulted in heightened task performance as the optimal tension level was approached. Further tension reduction then went beyond that crucial level and lower performance ensued.

In both of the above studies the authors supported the notion that humor reduces high levels of anxiety to more moderate levels and that these moderate levels must be reached in order to facilitate cognitive functioning and test performance. Methodological concerns can be raised concerning both studies (Townsend & Mahoney, 1981). It is difficult to know whether the matched versions of the test in the Smith et al. (1971) study were equivalent forms, although the difference between the mean scores for the matched humorous and nonhumorous items was not significant. In the Terry and Woods (1975) study the inferred relationships between humor and anxiety are questionable without an anxiety measure.

It has also been shown that humor does not always serve to reduce tension. In a Levine and Abelson (1959) study, it was shown that groups of psychiatric patients, beset with anxiety and other symptoms of psychopathology, reacted more negatively than a control group of Naval enlistees in their judgments of appreciation of popular cartoons. The authors concluded that for highly anxious persons, some humorous stimuli may evoke a painful rather than a gratifying response. Two more recent studies have failed to support the supposed tension-relieving properties of humor. Hedl, Hedl, & Weaver (Note 1) investigated the appreciation of humor under achievement-oriented vs. non-stressful conditions, and Townsend and Mahoney (1981) investigated the effects of matched humorous-nonhumorous test forms.

The present study was developed for considering the contribution of humor and anxiety to test performance. For example, if a developer incorporates humor in test items, does it tend to improve or interfere with the testing? Does the humor reduce debilitating anxiety or create additional anxiety? Does humor facilitate or reduce concentration? Is humor appreciated without having a negative effect on test performance?

Because of the conflicting results of previous studies, directionality has been avoided in specifying the following research questions:

- Does performance on a humorous form of a test differ from performance on a nonhumorous form?
  - Does mean performance on humorous items differ from mean performance on nonhumorous items?
  - Does the inclusion of humorous items have an effect on mean performance for post-treatment items?
  - Is the reliability of a test affected by the inclusion of humor?
- Does the inclusion of humor in test items affect students' anxiety level?
  - What is the interaction of test performance, anxiety, and (humorous/nonhumorous) treatment?
- What do students perceive as the effects of humor on their test performance?
  - Do students wish to have humor included on tests?
  - Is the perceived easiness of the test related to the inclusion of humor in the items?

Method

Sample

One hundred twenty-six (126) students in the eighth-grade English classes of a suburban-rural school district participated in the study. All students were taught by one of two teachers and divided homogeneously into six sections of two advanced, three average, and one skills-level classes.

Instruments

Grammar Test. The test consisted of 50 items based on grammar topics outlined in the eighth-grade syllabus and corresponding to topics covered for the eighth-grade level of the Iowa Test of Basic Skills (ITBS). Topics included subject-verb agreement, comparative adjectives, homonyms, general usage, punctuation, and capitalization.

The items were judged for clarity and appropriateness for the grade level by five graduate students in education and reviewed by two English teachers in a suburban school district. Items were written in two multiple-choice formats: a stem with four options (MC), and sentence broken into three lines (Sentence). On the MC items, the stems differed for matched humorous-nonhumorous pairs, while the options were identical. For the Sentence items, humorous modification took place within the options, since for each item the options combined to form a sentence. Examples of the items appear in Table 1.

Two parallel forms of the test were constructed. There were 15 identical nonhumorous items for the first subtest in both forms functioning as a pretest to compare the groups. The items on the second and third subtests were interspersed: the second had 20 humorous items on one form matched with 20 nonhumorous items on the other form functioning as treatment/control; the third part of the test had the remaining 15 nonhumorous items identical on both forms functioning as a posttest. The same item order was followed for

both forms, with the parallel nonhumorous items substituted in the comparison form (see Table 1).

Questionnaires. Students receiving the alternate test forms were asked to complete separate questionnaires. Both forms included items relating to anxiety; eight of these items were used to form an anxiety score. Two items related to student perception of "how easy" and "how much fun" the test was. Both forms also included an item to query whether students seek humor on tests.

Students receiving the humorous test form were also asked to respond to questions related to 1) whether they noticed the humor; 2) how funny they thought the humorous interjections were; 3) how they felt the humorous items caused them to react while taking the test (four items); and 4) whether they thought the humorous questions varied in difficulty from the nonhumorous questions (two items).

Iowa Tests of Basic Skills. Scores from the Iowa Tests of Basic Skills were summed for two composite scores: 1) a Grammar composite based on the Capitalization, Punctuation, and Usage subtests; 2) a Verbal composite including those three scores plus the Vocabulary, Reading, and Spelling subtests. The composite scores were used to check for equality of ability between treatment groups, to provide (criterion-related) validity information for the newly-developed grammar test, and to allow for further consideration of the grammar variable as related to other variables.

Procedure

Packets each containing a test and a questionnaire (the latter having been sealed in an envelope) were arranged so that the humorous form was alternated with the nonhumorous form and therefore distributed essentially randomly within each class. Because of the University's guidelines for human subjects research, the test was presented as an optional exercise rather than

as a regular classroom test. Students were requested to complete the test and then to respond to the optional questionnaire. They were instructed not to identify themselves by name on either the numbered test answer sheet or the questionnaire; however, the procedure allowed for matching scores with the ITBS scores.

Results

Judging from the means of the first subtest scores for the experimental instrument and of the ITBS scores, the groups of students in each treatment were comparable (see Table 2).

The same table (#2) is also relevant when considering the impact of the treatment on test performance. The inclusion of humor had no apparent effect on performance for the matched items (subtest 2) or for the common items (subtest 3) as summarized by the means and the reliability coefficients. Similarly, the inclusion of humor seemed to have no effect on the relationships among subtest scores or between the subtest and the Iowa scores (see Table 3). For example, the correlation between subtests 2 and 3 was .65 with the humorous form and .62 with the nonhumorous form--a difference representing neither practical nor statistical significance.

The inclusion of humor also had no apparent effect on the anxiety level, as the t value for the difference in the mean anxiety scores was nonsignificant at .11. The interaction of test performance based on subtest 2 or subtest 3 with anxiety and (humorous/nonhumorous) treatment was also nonsignificant.

Two of the questionnaire items were directed toward awareness and appreciation of the humor on the test. Fifty-seven out of 62 students noticed that humor was included in some of the questions. When they were asked about the funniness of the humor, one-third indicated "very funny" or "funny",

one-half indicated "funny, but not that funny", and one-sixth indicated "not funny at all".

Do the students favor the inclusion of humor? When asked on the questionnaire whether they would like most tests to include jokes, 12 responded "no" and 110 responded "yes". Inferring from the results in Tables 4 and 5, students who responded to the humorous form of the test perceived the jokes to be helpful and not harmful, and judged the funny questions to be easier and not harder. The responses concerning easiness were consistent: no student indicated that the humorous items were both easier and harder.

The inclusion of humorous items, then, seemed to have no deleterious effects on test performance but was supported by the students.

Discussion/Implications

As may be judged from the results, this study tends to support the inclusion of humorous items especially when considering student reaction. The apparent lack of harm to test scores (as evidenced by similarity in means, reliability coefficients, and intercorrelations) does not militate against the conclusion based on student preferences, although there was no pattern of facilitated performance to provide stronger support. Moreover, the fact that the humorous interjections were made in the form of legitimate test items rather than extraneous humorous insertions, provides reasonable evidence that such item alterations can be made without sacrificing acceptable standards for instrument construction and without lengthening the test. (Note that such lengthening would impose both a practical cost and a comparison-of-treatment compromise.)

One characteristic of the current study which may be vital when interpreting the result is the extent to which the testing situation was likely

perceived by the students as relatively low in anxiety production. (One index of the anxiety level of the group was based on a set of 8 self-report items. The possible range of scores was from 8(low) to 32(high); the actual range was for 8 to 24, with a median of 12.8 and a mean of 13.9. Scores for the two treatment groups were essentially identical, with $\underline{t}$ = .11.) A replication of the study using appropriate but less extreme human subject guidelines might provide a different conclusion for research questions involving test scores, especially when relationships with anxiety are considered. Perhaps one reason for the equivocal results in the literature would be differences in anxiety among the samples from one study to another.

At a speculative level, another question emerges when considering how a test of grammatical usage differs from those designed to assess knowledge in most other academic subjects. It would appear the specific item content is more vital in tests of the latter variety. If humor must first be recognized and relegated to the background in order to complete an item, it intuitively seems reasonable to estimate that it could represent an additional source of complexity to any threatened examinee. When item content is less crucial, as in a test of grammatical relationships, it may be that there is less of a tendency to react negatively to the item if it has been modified. Further research could perhaps illuminate whether differences between perceptions or processes exist for these different types of items.

Testing has become a major component used in making evaluative decisions of countless types for both individuals and groups. Certainly efforts should be made to create testing situations leading to the best descriptive interpretations or decisions possible within the constraints of testing time. At the same time efforts should be made to create positive rather than negative

reactions to testing among the test takers.

If humor can be a source of positive affect, and humor is capable of reducing negative affective states, then humor, when introduced into the assessment process, could appeal to test takers without depressing scores. By minimizing some of the negative attitudes prompted by continuous testing and by the threat included in many testing situations, the progress of the test taker and the effectiveness of the instructional program might be depicted more accurately. With the inclusion of humor, the whole testing process could be a step more humane.

Art Linkletter, in A Child's Garden of Misinformation, includes a definition of a hypotenuse: a humane device for hanging hypotemusses. Perhaps we could also muse over our humane devices.

Reference Note

1. Hedl, J. J., Hedl, J. L., & Weaver, D. B. Test anxiety and humor appreciation. Paper presented at the meeting of the American Educational Research Association, Toronto, March, 1978.

References

Adams, R. C. Is physics a laughing matter? Physics Teacher, 1972, 10, 265-266.

Ball, S., & Bogatz, G. A. Summative research of Sesame Street: Implications for the study of pre-school-aged children. Princeton, N. J.: Educational Testing Service, 1971.

Davies, A. P., & Apter, M. J. Humor and its effect on learning in children. In A. J. Chapman & P. E. McGhee (Eds.), Children's Humor, London: John Wiley & Sons, 1980.

Dworkin, E. S., & Efran, J. S. The angered: Their susceptibility to varieties of humor. Journal of Personality and Social Psychology, 1967, 6, 233-236.

Earls, P. L. Humorizing learning. Elementary English, 1972, 49, 107-108.

Freud, S. Humour: International Journal of Psychoanalysis, 1928, 9, 1-6.

Goldstein, J. H., & McGhee, P. E. (Eds.). The psychology of humor. New York: Academic Press, 1972.

Hill, K., & Sarason, S. B. The relation of test anxiety and defensiveness to test and school performance over the elementary-school years. Monographs of the Society for Research in Child Development, 1966, 31, 1-76.

Keith-Spiegel, P. Early conceptions of humor: Varieties and issues. In J. H. Goldstein & P. E. McGhee (Eds.), The psychology of humor. New York: Academic Press, 1972.

Levine, J., & Abelson, R. Humor as a disturbing stimulus. Journal of General Psychology, 1959, 60, 191-200.

Monson, D. Children's test responses to seven humorous stories. Elementary School Journal, 1968, 58, 334-339.

Sarason, I. G., Pederson, A., & Nyman, B. Test anxiety and the observation of models. Journal of Personality, 1968, 36, 493-511.

Singer, D. L. Aggression arousal, hostile humor, and catharsis. Journal of Personality and Social Psychology Monograph Supplement, 1968, 8, 1-14.

Smith, R. E., Ascough, J. C., Ettinger, R. F., & Nelson, D. A. Humor, anxiety, and task performance. Journal of Personality and Social Psychology, 1971, 19, 243-246.

Spencer, H. The physiology of laughter. Macmillan's Magazine, 1960, 1, 395-402.

Terry, R. L., & Woods, M. E. Effects of humor on the test performance of elementary school children. Psychology in the Schools, 1975, 12, 182-185.

Townsend, M. A. R., & Mahoney, P. Humor and anxiety: Effects on class test performance. Psychology in the Schools, 1981, 18, 228-234.

Ziv. A. Facilitating effects of humor on creativity. Journal of Educational Psychology, 1976, 68, 318-322.

Table 1

Illustrative Matched Humorous and Nonhumorous Multiple Choice and Sentence Items Used in Subtest 2

Format	Type	Items
Multiple Choice	Humorous	Mrs. Jones found Mr. Jones in the freezer. Apparently the kids put ____ father in ____ because they wanted to have a cold pop. A. there; their B. their; there C. their; they're D. there; they're
	Nonhumorous	Mrs. Jones heard barking in the basement. Apparently the kids put ____ dog down ____ because they wanted to play baseball. A. there; their B. their; there C. their; they're D. there; they're
Sentence	Humorous	A. The umpire's new glasses seemed to B. be helping him until he called C. a eagle that flew by a foul ball. D. No mistakes.
	Nonhumorous	A. The umpire's new glasses seemed to B. be helping him until he called C. a outside pitch a strike. D. No mistakes.

Table 2

Summary Statistics of Grammar Test and ITBS Scores for Humorous and Nonhumorous Groups

Variable Group	n	Statistics M	SD	t	Reliabilities KR_{20}	t	Split	t
GRAMMAR TEST[a]								
Subtest 1								
Humorous	64	6.05	2.12		.32		.16	
				.28		.37		-.78
Nonhumorous	62	5.95	2.00		.26		.29	
Subtest 2								
Humorous	64	14.58	3.13		.82		.84	
				.15		.48		.68
Nonhumorous	62	14.50	2.98		.79		.80	
Subtest 3								
Humorous	64	7.28	2.70		.67		.55	
				.21		.10		-1.07
Nonhumorous	62	7.18	2.73		.66		.67	
ITBS SCORES[b]								
Grammar								
Humorous	59	64.22	13.47		not available			
				.41				
Nonhumorous	57	63.14	14.78					
Verbal								
Humorous	58	162.66	35.14					
				.49				
Nonhumorous	57	159.37	37.31					

[a]The grammar test is composed of three subtests. Subtest 1 is a pretest of 15 items given to both groups. Subtest 2 is 20 humorous or matched nonhumorous items. Subtest 3 is a post-test of 15 items given to both groups.

[b]Two composites were formed by summing raw scores for the Iowa Tests of Basic Skills. Grammar = Usage, Capitalization, and Punctuation. Verbal = Vocabulary, Reading Comprehension, Spelling, and the three Grammar subtests.

Table 3

Correlations for the (Experimental) Grammar Test and the Iowa Tests of Basic Skills Composites

Test		Grammar Test			ITBS	
		Score 1	Score 2	Score 3	GRAM	VERBAL
Score	1		35	32	60	63
	2	42		65	60	65
	3	32	62		64	63
ITBS	GRAM	38	66	62		93
	VERB	44	69	67	92	

Note. The correlations above the diagonal are based on performance for students receiving the humorous subtest; the correlations below the diagonal are from students receiving the nonhumorous subtest. Decimals are omitted. Reliability coefficients for the Grammar Test are given in Table 2.

Table 4

Student Perceptions of Effects of Humor on the Grammar Test

Did the jokes on this test[a] ...	No	Yes
help you feel more relaxed?	14	47
help you to concentrate?	26	34
make you more tense?	60	2
make you confused?	57	4

[a]These four items, each including the heading, appeared consecutively within the form of the questionnaire received by the humorous treatment group.

Table 5

Perceptions of Easiness of the Humorous Items

Funny Items Easier	Funny Items Harder: Yes	Funny Items Harder: No
No	3	11
Yes	0	47

Corrected Chi Square = 6.51, $p \leq .01$
Pearson's $r = .42$, $p \leq .001$