DOCUMENT RESUME

ED 387 528 TM 024 079

AUTHOR Morse, David T.
TITLE The Relative Difficulty of Selected Test-Wiseness Skills among College Students.
PUB DATE Nov 94
NOTE 16p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Nashville, TN, November 9-11, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Age Differences; Children; *Cues; *Difficulty Level; Grammar; Higher Education; Response Style (Tests); *Skill Development; Student Reaction; *Test Wiseness; *Undergraduate Students; Young Adults
IDENTIFIERS *Gibb Experimental Test of Testwiseness

ABSTRACT

The relative difficulty of the seven test-wiseness skills measured by the Gibb Experimental Test of Testwiseness, a measure of cue-using skills, was studied. Participants were 243 undergraduates from 3 universities, 79% of whom were Caucasian. Participants reported a mean grade point average of 3.0 on a 4-point scale. Results suggest that some of the test-wiseness skills identified by the Gibb measure do differ significantly in how easy they are to apply. The easier skills were observed to be the use of grammar cues, choosing the correct alternative when it was notably longer than other choices, and eliminating absurd or unrelated alternatives. These skills were found to be easier than avoiding alternatives containing specific determiners, such as all, everyone, or never. Alliterative association was the second most difficult of the skills to demonstrate on the Gibb test. Previous research suggests that young adults are able to respond to measures of test-wiseness in ways that may differ from those of children. Further research should consider age differences as well as a wider range of test-wiseness skills. (Contains 3 tables and 19 references.) (SLD)

The Relative Difficulty of Selected Test-Wiseness Skills Among College Students

David T. Morse
Mississippi State University

A Paper Presented at the Annual Meeting of the Mid-South Educational Research Association, Nashville, TN, November 1994.

Test-wiseness was introduced as a construct at least four decades ago. Thorndike (1951), discussing sources of variation entering into observed test score differences, identified test-wiseness as a persistent, general attribute of the examinee that would contribute in part to differences among individuals. In their seminal article, Millman, Bishop and Ebel (1965) identified the concept of test-wiseness, articulating their explanation with a proposed taxonomy of test-wiseness skills. Millman et at. defined test-wiseness as "a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score" (p. 707). They further asserted that test-wiseness "is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures" (p. 707). A refinement offered by Millman et al. was that of separating test-wiseness skills into two broad domains, skills that are logically independent of the test purpose or test constructor (class I), and skills that are dependent on the test purpose or test constructor (class II).

Research on test-wiseness suggests that: (a) differences in test scores do correlate, to varying degree, with test-wiseness (Sarnacki, 1979); (b) test-wiseness skills can be taught and learned by examinees as young as upper elementary school grades (Sarnacki, 1979; Samson, 1985; Dolly & Williams, 1986); and (c) many times teacher-made tests include cues that would make the items artificially easier for test-wise examinees (Brozo, Schmelzer & Spires, 1984). However, there has been very little research focused upon the degree to which different test-wiseness skills might well not be of equal difficulty to learn or to apply in a testing situation. To date, the related literature has been scarce.

Differences Among Test-Wiseness Skills for Adults

The studies discussed in this section stemmed from investigations of whether poor item-writing practices, as stated in textbooks on measurement or test construction, actually influenced examinee performance on or the technical characteristics of tests. The basic research design was that of taking what were considered acceptable test items and rewriting them to reflect various of the item writing flaws pointed out in texts and administering the

items in counterbalanced fashion to examinees. The degree to which average performance differed on the items written to include flaws versus the original versions was thought to be indicative of the impact of the poor item-writing practices. Generally, only a few of the test-wiseness skills were incorporated in these studies.

Dunn and Goldstein (1959) tested 832 Army trainees during the eighth week of basic training on four-option multiple choice items covering four subject areas. Twenty-five items were written to reflect each of various three test-wiseness cues: inclusion of irrelevant cues or specific determiners (the Millman et al. category for specific determiners is II.B.3), grammar cues (II.B.1.h), or having the longest alternative be the correct choice (II.B.1.a). The mean item p-value (proportion of examinees correctly responding) for each type of item was, in order, .54 for cues/specific determiners, .55 for grammar cues, .59 for length, and .61 for items having both length and grammar cues. When compared with "unflawed" items, the mean differences in p-values were .03, .03, .07, and .09 for cues, grammar, length, and length plus grammar. Thus, in this study, length (in this case the longest answer being correct) would appear to be the type of cue most easily detected by examinees uninstructed in test-taking skills.

Board and Whitney (1972) re-wrote acceptable items from an undergraduate course in American Politics to reflect various item writing flaws; those that related to test-wiseness skills included: (a) keyed responses noticeably longer or shorter (II.B.1.a), and (b) grammatical cue to keyed response (II.B.1.h). Fifteen items for each skill (or flaw) were used, and were tested in both flawed and unflawed form. Each set of items was administered to 80 undergraduates (160 total) who had been blocked into five levels based on their performance on an unadulterated test of course content. Overall, the mean p-value for the length cue items was .64, while that for grammar cue items was .67. The difference between the flawed and unflawed versions of items was negligible, however, about .01 for length cue and .00 for grammar cue. Board and Whitney noted a statistically significant item version by ability level interaction for the length cue items. By quintile, the mean p-value difference for the flawed

items was .08, .04, .10, -.15, and .00 for the fifth to the first ability level, respectively. Thus the lower ability students were better able to capitalize on this type of cue, but their performance was lower than that of the upper ability students.

Weiten (1984) generated eight "flawed" or test-wiseness cue-laden and eight "good" items for each of four test-wiseness cues: (a) Item stem-answer resemblance (Millman et al. category II.B.4), (b) grammar cue (II.B.1.h), (c) implausible or absurd alternatives (I.D.1), and length of (longer) correct alternative (II.B.1.a). When these items were administered to 54 undergraduate students enrolled in a child psychology course, the mean p-values were, in order, .59 for stem-answer similarity, .60 for grammar and for length, and .71 for absurd alternatives. The mean differences in p-value across flawed and good versions of the items were .03 for grammar, .09 for length, .10 for stem-answer resemblance, and .18 for absurd alternatives.

In all, the studies involving adult respondents have addressed from two to four test-wiseness skills; in all five such skills were investigated. The findings of these studies are summarized in Table 1.

Differences Among Test-Wiseness Skills for Children

With one exception, the studies cited in this section were not specifically oriented toward evaluating the relative difficulty of test-wiseness skills. As was the case for studies involving adults, most studies involving children only addressed a small number of test-wiseness skills.

Slakter, Koehler, and Hampton (1970a, 1970b) administered a test measuring four of the Millman et al. skills, each by four multiple choice items: (a) stem-answer resemblance, (b) options known to be incorrect (I.D.1), (c) similar options (I.D.2), and (d) specific determiners in options (II.B.3). In the first study (Slakter et al., 1970a), these items were administered to approximately 2360 students in grades 5-11 in small school districts in western New York and northern Michigan. Though means were not given, the relative order of mean p-values was

stable across the skills for five of the seven grades. From easiest to most difficult, the sequence was known incorrect options, stem-answer resemblance, similar options, and specific determiners.

In the second study (Slakter et al., 1970b), the authors administered the same test to 76 high school seniors who had been trained in test-wiseness skills and to 85 seniors who had not been trained. On the immediate posttest after training, the mean item p-values for the trained students were: .75 for specific determiners, .80 for similar options, .82 for stem-answer resemblance, and .86 for known incorrect options. These values were in the same order as found for the students in grades 5-11. However, the difference in mean p-values between the trained and untrained groups was just the reverse; the largest difference (.33) was observed for specific determiners, followed by .28 for similar options, .15 for stem-answer resemblance, and .02 for known incorrect options. Thus, what apparently were the more difficult skills to demonstrate were those on which the greatest gains due to training were observed.

Diamond and Evans (1972) generated six four-option multiple choice items to measure each of five test-wiseness skills. These were administered to 95 sixth grade students from a suburban Philadelphia school. In order of increasing mean p-value, the skills were: (a) grammar cue (II.B.1.h) = .35, (b) overlapping distractors, such that the truth of one implies the correctness of several others (I.D.2) = .45, (c) specific determiners (II.B.3) = .50, (d) length of correct alternative (II.B.1.a) = .53, and (e) alliterative association (II.B.4) = .77. An interesting twist to this study was that these subscores were correlated with the Lorge-Thorndike IQ test, and observed moderate correlations for specific determiners (.43), grammar (.46) and alliterative association (.51), but not for length of correct alternative (.21) or overlapping distractors (.05).

Diamond, Ayrer, Fishman, and Green (1976) administered the same set of items used by Diamond and Evans (1972) to 40 fifth grade and 36 sixth grade students at an inner city school in Philadelphia. In order of difficulty, the overall mean p-values were: (a) grammar cue = .23, (b) specific determiners = .31, (c) overlapping distractors = .34, (d) length of

correct alternative = .39, and (e) alliterative association = .45. The ranking of these skills was exactly as was observed with higher ability students in Diamond and Evans. In general, the sixth grade students in Diamond et al. outperformed the fifth grade students, the two skills being exceptions in which roughly equal mean performance was observed included grammar cues and overlapping distractors.

McMorris, Brown, Snyder, and Pruzek (1972) constructed seven "flawed" and seven "clean" items for each of three test-wiseness skills to be administered to 494 eleventh grade students in a suburban New York school district. In order of mean difference in p-value favoring flawed over clean items, the skills included: (a) length of correct alternative = .03, (b) grammar cues = .07, and (c) stem-correct answer resemblance = .09.

Carter (1986) in an unusual study, administered one item for each of five test-wiseness skills: (a) having choice "C" be the correct answer (II.B.1.d), (b) length of correct alternative, (c) alliterative association, (d) grammar cue, and (e) "+/- options" in which one option was positively stated and the other three were negatively stated (II.B.1.f). The items were administered to 312 seventh grade students. In increasing order of p-value, the skills were grammar cue = .27, longer correct alternative = .50, alliterative association = .55, choice "C" = .69, and +/- options = .80. Subsequent interviews with some of the participants suggested, however, that the +/- options item suffered also from several absurd alternatives in the negatively stated choices.

The only study that explicitly appraised relative difficulty of certain of the Millman et al. test-wiseness skills was reported by Morse (1980). Twelve skills were appraised, each with from four to six items per skill, by administration to about 2900 fifth and sixth grade students in 30 Mississippi school districts. Using Rasch model logit scaling, the mean difficulty of the various skills was, in increasing order: (a) guess when there is no penalty (I.C.1, -2.15), (b) look over test before starting (I.A.2/I.A.3, -1.20), (c) read and follow directions (I.B.1, -.80), (d) Look for cues elsewhere in the test (I.D.5, -.23), (e) change your answer if you believe your first choice is wrong (I.A.5, -.02), (f) specific determiners (II.B.3, .04), (g) stem-item

resemblance (II.B.4, .10), (h) grammar cue (II.B.1.h, .14), (i) length of correct alternative (II.B.1.a, .22), (j) Budget your time and check your progress (I.A.2, .74), and (k) Don't choose your answer from a set of similar answers (I.D.2/I.D.4, .97). Morse reported that there was a statistically significant difference between the class I and class II skills in mean difficulty level, with the difference being nearly one-half a logit (standard deviation), such that the class I skills were the easier to demonstrate.

Across the studies reported, most may be characterized as not directly addressing the issue of relative difficulty of test-wiseness skills, and most involved only a few such skills. The purpose of the present study was to investigate the relative difficulty of the seven test-wiseness skills measured by the Gibb Experimental Test of Testwiseness.

Method

Subjects

Participants were 243 undergraduate students (62 men, 178 women, 3 unidentified by gender) from three universities. Forty-one (17%) were African-American students, four (2%) were Asian-American, 191 (79%) were Caucasian, three (1%) were Hispanic, and the other four were unidentified by ethnicity. The mean age of participants was 22.5 years (SD = 5.2). The self-reported mean grade point average was 3.0 on a four-point scale (SD = 0.5). All participants volunteered to enter the study; as an incentive to participate, test-taking skill workshops were offered after completion of the study.

Instrument

The Gibb Experimental Test of Testwiseness (Gibb, 1964) was designed to measure seven cue-using skills, each with 10 four-option multiple choice items: (a) alliterative association (II.B.4), (b) incorrect/absurd alternatives (I.D.1), (c) specific determiners (II.B.3), (d) precision or qualification of answer (II.B.1.b), (e) longer correct alternative (II.B.1.a), grammar cue (II.B.1.h), and (f) cues elsewhere in the test (I.D.5). Gibb found that the test could distinguish the test-wiseness performance of trained from untrained undergraduate

students. Sarnacki (1979) in his review on test-wiseness, declared the Gibb test to be the best available measure of test-wiseness. Miller, Fuqua and Fagley (1990) performed a principal components analysis on the seven subskills of the Gibb test and concluded that a two-factor structure seemed to represent the test well. Harmon, Morse and Morse (1994) reported on a confirmatory factor analysis of the Gibb test, showing that either a two-factor or a one-factor model could be asserted to represent the test.

Procedure

All participants were administered the Gibb Experimental Test of Testwiseness. There were no special instructions regarding guessing, nor did subjects receive any training in test-wiseness principles prior to completing the test. Gibb (1964) reported that undergraduate students could easily complete the 70-item test within 45 minutes, and that time seemed ample for all but a very few of the participants. Separate, machine-scoreable answer sheets were used to record the responses, which were then scanned for further analysis.

One-parameter logistic model (Rasch) item difficulty estimates were generated for each of the 70 items. These difficulty estimates were scaled as logits (log units), and arbitrarily centered at zero. On the logit scale, one logit is analogous to a standard deviation. Difficulty values below zero represent relatively easier items whereas values above zero represent relatively more difficult items. The Rasch difficulty values were then used as the data for a one-way analysis of variance (ANOVA), treating the seven test-wiseness skills as seven levels of the factor. Thus, the sample size for each skill was 10, representing the obtained Rasch difficulty estimate for each item measuring a given skill on the Gibb test. Significance tests were run at the .05 level.

Results

The mean Rasch difficulty estimates by test-wiseness skill measured on the Gibb test are presented in Table 3, and varied from -.33 (grammar cue) to .58 (specific determiners). Initial checks for homogeneity of variance (Levene's $F(6,63 = 1.09, p = .379)$ and normality

(Lilliefors' adaptation of the Kolmogorov-Smirnov test, D-max = .086, $p > .20$) suggested no apparent problems with the usual ANOVA assumptions. The one-way ANOVA yielded a statistically significant result, $F(6,63) = 4.47$, $p = .0008$. Follow-up testing via Tukey's HSD procedure indicated that the most difficult skill on average, specific determiners (M = .58) was statistically significantly more difficult than grammar cues (M = -.33), longer correct alternatives (M = -.28), and absurd or unrelated alternatives (M = -.25). No other difference was statistically significant.

Discussion

Some of the test-wiseness skills measured by the Gibb Experimental Test of Testwiseness do differ significantly in how easy they are to apply. Overall, the easier skills were observed to be the use of grammar cues, choosing the correct alternative when it was notably longer than other choices, and eliminating absurd or unrelated alternatives, which were found to be statistically significantly easier than avoiding alternatives containing specific determiners, such as all, everyone, or never. These findings are consistent with those that have been reported in other studies using young adults.

Dunn and Goldstein (1959) found that specific determiner cues were the most difficult to demonstrate of those that they compared, while grammar and length of correct response cues had mean p-values that differed only slightly. Board and Whitney (1972) observed very little difference between mean item p-values for grammar and length cues. Weiten (1984) reported that grammar and length cues had the same mean item p-value, and were somewhat more challenging than were items involving absurd alternatives. Weiten's study indicated that stem-answer resemblance, which alliterative association represents, was the most difficult of the four skills he compared, but was not very different from grammar or length mean item p-values. In the present study, alliterative association was the second most difficult of the skills to demonstrate on the Gibb test.

Results from studies using children are somewhat different; one reason for this difference is that grammar cues appear to be relatively more difficult for children to use than

adults. Diamond and Evans (1972), Diamond et al. (1976), Carter (1986), and Morse (1980) reported that grammar cue items were either the most or among the more difficult test-wiseness skills to demonstrate. Several studies, though, including Slakter et al. (1970a, 1970b) and Diamond and Evans showed specific determiners to be the most difficult of the skills to apply. Morse noted that specific determiner items were close in mean difficulty to the average for the entire set of items, though as a set they were the fourth most difficult among the 12 skills examined.

These results suggest that not all test-wiseness skills are of equal difficulty. Further, the way young adults are able to respond to measures of test-wiseness may be qualitatively different, due perhaps to experience or cognitive strategies that have evolved over time, than the way children can respond. Researchers or trainers addressing test-wiseness should take into account such differences. Further research addressing the age factor as well as a wider sampling of test-wiseness skills from the Millman et al. (1965) taxonomy would aid in understanding how these findings might apply over a broader range of examinee characteristics and specific test-wiseness skills.

References

Board, C., & Whitney, D. R. (1972). The effect of selected poor item-writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 9, 225-233.

Brozo, W. G., Schmelzer, R. V., & Spires, H. A. (1984). A study of test-wiseness clues in college and university teacher-made tests with implications for academic assistance centers. (Report No. 84-01). GA: College Reading and Learning Assistance Technical Report. (ERIC Document Reproduction Service No. ED 240-928).

Carter, K. (1986). Test-wiseness for teachers and students. Educational Measurement: Issues and Practice, 5(4), 20-23.

Diamond, J. J., Ayrer, J., Fishman, R., & Green, P. (1976). Are inner city children test-wise? Journal of Educational Measurement, 14, 39-45.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 9, 145-150.

Dolly, J. P., & Williams, K. (1986). Using test-taking strategies to maximize multiple-choice test scores. Educational and Psychological Measurement, 46, 619-625.

Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 19, 171-179.

Gibb, B. G. (1964). Test-wiseness as a secondary cue response. Unpublished doctoral dissertation, Stanford University. Ann Arbor, MI: University Microfilms Document 64-76 3.

Harmon, M. G., Morse, D. T., & Morse, L. W. (1994, November). Confirmatory factor analysis of the Gibb Experimental Test of Testwiseness. Paper presented at the Mid-South Educational Research Association, Nashville, TN.

McMorris, R. F., Brown, J. A., Snyder, G. W., & Pruzek, R. M. (1972). Effects of violating item construction principles. Journal of Educational Measurement, 9, 287-295.

Miller, P. M., Fuqua, D. R., & Fagley, N. S. (1990). Factor structure of the Gibb Experimental Test of Testwiseness. Educational and Psychological Measurement, 50, 203-208.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. Educational and Psychological Measurement, 25, 707-726.

Morse, D. T. (1980, April). The relative difficulty of selected test wiseness skills. Paper presented at the National Council on Measurement in Education, Boston, MA.

Samson, G. (1985). Effects of training in test-taking skills on achievement test performance: A quantitative synthesis. Journal of Educational Research, 78, 261-266.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49, 252-279.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970a). Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 7, 119-122.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970b). Learning test-wiseness by programmed texts. Journal of Educational Measurement, 7, 247-254.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), Educational measurement (pp. 560-620). Washington, DC: American Council on Education.

Weiten, W. (1984). Violation of selected item construction principles in educational measurement. Journal of Experimental Education, 52, 174-178.

Table 1

Findings on Difficulty of Test-Wiseness Skills from Studies Using Adults

Study	Subjects	Millman, Bishop & Ebel Test-Wiseness Skill	Mean item p-value	Mean difference
Dunn & Goldstein (1959)	832 Army trainees, about 200 per test	Specific determiner (I.B.3)	.53	.03
		Length of correct alt. (II.B.1.a)	.59	.07
		Grammar cue (II.B.1.h)	.55	.03
		Length and Grammar cues	.61	.09
Board & Whitney (1972)	160 under-graduates	Length of correct alt. (II.B.1.a)	.68	.01
		Grammar cue (II.B.1.h)	.67	.00
Weiten (1984)	54 under-graduates	Stem-answer resemblance (II.B.4)	.59	.10
		Grammar cue (II.B.1.h)	.60	.03
		Absurd alt's (I.D.1)	.71	.18
		Length of correct alt (II.B.1.a)	.60	.09

Table 2

Findings on Difficulty of Test-Wiseness Skills from Studies Using Children

Study	Subjects	Millman, Bishop & Ebel Test-Wiseness Skill	Mean item p-value	Mean difference
Slakter et al. (1970a)	2361 students grades 5-11	Stem-answer resemblance (II.B.4)	2[a]	
		Incorrect/absurd options (I.D.1)	1	
		Similar options (I.D.2)	3	
		Specific determiners (II.B.3)	4	
Slakter et al. (1970b)	76 high school seniors given testwiseness training	Stem-answer resemblance (II.B.4)	.82	.15[b]
		Incorrect/absurd options (I.D.1)	.86	.02
		Similar options (I.D.2)	.80	.28
		Specific determiners (II.B.3)	.75	.33
Diamond & Evans (1972)	95 suburban 6th graders	Length of correct alt (II.B.1.a)	.53	
		Grammar cue (II.B.1.h)	.35	
		Specific determiners (II.B.3)	.50	
		Alliterative association (II.B.4)	.77	
		Overlapping distractors (I.D.2)	.45	
Diamond et al. (1976)	76 inner-city 5th and 6th graders	Length of correct alt (II.B.1.a)	.39	
		Grammar cue (II.B.1.h)	.23	
		Specific determiners (II.B.3)	.31	
		Alliterative association (II.B.4)	.45	
		Overlapping distractors (I.D.2)	.34	
McMorris et al. (1972)	494 suburban 11th graders	Stem-answer resemblance (II.B.4)		.09
		Grammar cue (II.B.1.h)		.07
		Length of correct alt (II.B.1.a)		.03
Carter (1986)	312 7th graders	Choice "C" keyed alt (II.B.1.d)	.69	
		Length of correct alt (II.B.1.a)	.50	
		Alliterative association (II.B.4)	.55	
		Grammar cue (II.B.1.h)	.27	
		+/- options (II.B.1.f)	.80	
Morse (1980)	2860 5th and 6th graders	Guess (I.C.1)	-2.15[c]	
		Look over test (I.A.2/I.A.3)	-1.20	
		Read & follow directions (I.B.1)	-.80	
		Cues elsewhere in test (I.D.5)	-.23	
		Change answer if wrong (I.A.5)	-.02	
		Known wrong answers (I.D.1)	-.02	
		Specific determiners (II.B.3)	.04	
		Stem-answer resemblance (II.B.4)	.10	
		Grammar cue (II.B.1.h)	.14	
		Length of correct alt (II.B.1.a)	.22	
		Budget time & check progress (I.A.2)	.74	
		Similar alternatives (I.D.2/I.D.4)	.97	

Note: [a] Exact values were not given, these represent rank order from easiest to hardest.
[b] These represent differences between the trained students and 76 untrained students
[c] These represent Rasch difficulty values, expressed as logits.

Table 3

Summary Statistics for Rasch Difficulty Estimates by Test-Wiseness Skill

Skill	Mean	SD
Grammar cue	-0.33	0.39
Longer correct alt.	-0.28	0.43
Absurd alternatives	-0.27	0.70
Cues elsewhere on test	-0.01	0.61
More precise/qualified alt.	0.03	0.32
Alliterative association	0.25	0.36
Specific determiners	0.58	0.52
Overall	0.00	0.56

Note: 10 items per skill.

Table 4

ANOVA Summary Table

Source	SS	df	MS	F	prob(F)
TW Skill	6.49	6	1.08	4.47	.0008
Residual	15.24	63	0.24		
Total	21.73	69			

Note: Eta squared = .30