Journal of Educational Measurement
Winter 1998, Vol. 35, No. 4, pp. 328–347

Reviewing and Changing Answers on Computer-adaptive and Self-adaptive Vocabulary Tests

Walter P. Vispoel
University of Iowa

Results obtained from computer-adaptive and self-adaptive tests were compared under conditions in which item review was permitted and not permitted. Comparisons of answers before and after review within the "review" condition showed that a small percentage of answers was changed (5.23%), that more answers were changed from wrong to right than from right to wrong (by a ratio of 2.92:1), that most examinees (66.5%) changed answers to at least some questions, that most examinees who changed answers improved their ability estimates by doing so (by a ratio of 2.55 to 1), and that review was particularly beneficial to examinees at high ability levels. Comparisons between the "review" and "no-review" conditions yielded no significant differences in ability estimates or in estimated measurement error and provided no trustworthy evidence that test anxiety moderated the effects of review on those indexes. Most examinees desired review, but permitting it increased testing time by 41%.

The effects of reviewing and changing answers on fixed-item paper-and-pencil tests has been studied for at least 70 years (e.g., Crawford, 1928; Crocker & Benson, 1980; Cummings, 1981; Matthews, 1929; McMorris & Weiderman, 1986; McMorris, DeMers, & Schwarz, 1987; Mueller & Wasser, 1977; Smith, White, & Coop, 1979; Waddell & Blankenship, 1995). In summarizing much of this research, Benjamin, Cavell and Shallenberger (1984) noted several consistent findings: (a) only a small percentage of answers are changed, (b) more answers are changed from wrong to right than from right to wrong, (c) most examinees change answers to at least some questions, and (d) most examinees who change answers improve their scores by doing so. Although examinees are often unaware of the benefits of review and answer change, they are certainly accustomed to having and taking advantage of these options on paper-and-pencil tests.

Opportunities for item review and answer change are far less common on computerized tests, especially when those tests are administered adaptively. Allowing item review can complicate item administration algorithms, prolong testing

I am very grateful to the Iowa Measurement Research Foundation for supporting this project; to Robert Forsyth for granting me permission to computerize the *Iowa Tests of Educational Development*; to Timothy Bleiler, Karen Steger-May, Bethany Brunsman, Ellen Forte Fast, Huey-ing Tzou, and Jaeyool Boo for their assistance in collecting and analyzing data, and to Rebecca Zwick and the anonymous reviewers for their helpful comments about the original manuscipts.

The present article is an extension of a paper presented at the April, 1996 meeting of the National Council on Measurement in Education in New York City.

time, lower measurement precision, and increase the likelihood of examinees obtaining artificially inflated scores. However, most examinees desire review options because such options permit reconsideration of previous answers and corrections of typing and other careless errors. Moreover, disallowing review might have especially detrimental effects on the test performance of certain examinees (those at certain ability levels, those from particular ethnic backgrounds or cultural groups, etc.).[1]

Empirical research into the effects of item review on results yielded by computerized tests is very limited with most studies focused on answer review within adaptively-administered tests (either computer-adaptive or self-adaptive). Computer-adaptive tests (CATs) are ones in which examinees receive items from a large pool based on their responses to previous items. The examinee's ability is estimated after each item is answered, and new items are chosen to maximize the measurement precision associated with the examinee's most recent ability estimate. The typical benefit of CATs over ordinary fixed-item tests (FITs) in which everyone gets the same items is a substantial reduction in number of items and time it takes to reliably measure most examinees' ability levels (see, e.g., Lord, 1977; Urry, 1977; Vispoel, Wang, & Bleiler, 1997; Weiss, 1982).

Self-adaptive tests (SATs) are variations of CATs that allow the examinee to chose the difficulty level of each item. The main benefit of SATs is thought to be a decrease in the effects of text anxiety on ability estimates. This might occur because the act of selecting item difficulty enables examinees with high anxiety to feel more control over the testing situation, which in turn may increase their attention to the test and reduce negative self-defeating thoughts.[2] Support of this idea comes from studies in which SATs yielded higher ability estimates or ability estimates less correlated with test anxiety than did either CATs or FITs (Rocklin & O'Donnell, 1987; Rocklin, O'Donnell & Holst, 1995; Roos, Wise, & Plake, 1997; Wise, Plake, Johnson, & Roos, 1992; Vispoel & Coffman, 1994; Vispoel, Rocklin, & Wang 1994; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992).

Live testing research into the effects of review on results from CATs and SATs is limited (Ferrara et al., 1996; Lunz, Bergstrom, & Wright, 1992; Stone & Lunz, 1994; Vispoel et al., 1992). All of the studies cited here included CATs, and one (Vispoel et al., 1992) included SATs and FITs as well. Findings from these studies replicated those from previous paper-and-pencil testing research in that only a small percentage of answers was changed, more answers were changed from wrong to right than from right to wrong, most examinees changed answers to at least some questions, and most examinees who changed answers improved their scores by doing so. The studies also revealed that review increased testing time from 37% to 61% (Vispoel et al., 1992; Ferrara et al, 1996); that ability estimates before and after review were highly correlated (.98 or higher; Lunz et al., 1992; Stone & Lunz, 1994; Vispoel et al., 1992), that differences in measurement precision before and after review were of little practical significance (after/before test precision ratios were .97 or higher; Lunz et al., 1992; Stone & Lunz, 1994; Vispoel et al., 1992), that concurrent validity coefficients showed little change (Vispoel et al., 1992), and that review was desired by a substantial percentage of examinees (87% overall;

Vispoel et al., 1992). Examinees' strong desires for review opportunities also has been verified in other studies of computerized tests that did not focus directly on answer changing behavior (Baghi, Gabrys, & Ferrara, 1991; Gershon & Bergstrom, 1991; Legg & Buhr, 1992; Moe & Johnson, 1988; Schmitt, Urry, & Gugel, 1978; Vispoel, 1993; Vispoel, Rocklin, & Wang, 1994).

This Investigation

The purpose of the study reported here was to expand upon previous research into effects of item review on results obtained from computerized tests. Important features of the study included: (a) use of a well calibrated pool of vocabulary items from a nationally standardized test, (b) examination of two types of adaptively administered tests (CATs and SATs), (c) comparison of independent review and no-review testing conditions, (d) comparison of answer changes before and after review in the review condition, and (e) inclusion of test anxiety as an individual difference variable to explore possible interactions between it and the manipulated variables (item review and administration mode).

Three research questions were addressed.

1. What are the main and interactive effects of item review (allowed vs. disallowed), administration mode (CAT vs. SAT), and test anxiety on estimated vocabulary ability, estimated measurement error, and testing time?
2. Are examinees answer changing behaviors (percentage of answers changed, percentage changed from wrong to right vs. right to wrong, etc.) consistent with results reported in previous research?
3. What are examinee's attitudes concerning the importance of including review options on computerized vocabulary tests?

Method

Participants and Measures

The participants were 379 University of Iowa students who volunteered for the study to receive points toward their final grades in introductory educational psychology courses. Each participant completed the Test Anxiety Inventory (TAI; Spielberger, 1980), one form of a computerized vocabulary test (either CAT or SAT), and questionnaires assessing demographic information and attitudes about the computerized tests.

TAI

The TAI is designed to measure individual differences in "test anxiety as a situation-specific personality trait" (Spielberger, 1980, p. 1). It has twenty items that are answered using a 4-point scale (1 = almost never, 4 = almost always). Examples of items include "I feel very jittery when taking an important test," and "During tests I find myself thinking about the consequences of failing." Evidence supporting the construct validity of TAI Total scores as reported in the test manual includes alpha-reliability estimates ranging from .92 to .96, test-retest correlations of .80 or higher over 1 to 2 week time intervals, a test-retest correlation of .62 over a 6 month interval, and logical patterns of relations with other measures (e.g.,

positive correlations with other anxiety measures, and negative correlations with test scores and class grades). For the present examinees, the alpha-reliability estimate for the TAI Total score was .95.

Computerized Tests

The computerized tests were constructed using vocabulary items from five equated paper-and-pencil forms of the *Iowa Tests of Educational Development* (ITED; Feldt, Forsyth, & Alnot, 1986a, 1986b; Feldt, Forsyth, & Lindquist, 1979a, 1979b; Lindquist & Feldt, 1972). Item parameters were calibrated using the LOGIST V computer program (Wingersky, Barton, and Lord, 1982) based on responses of over 20,000 examinees to each item. A modified three-parameter model was used with *c* parameters fixed at .15. A total of 275 non-overlapping items was calibrated and a subset of 200 of them was selected for the CAT and SAT item pool. The items eliminated were those found to be too easy for college students who had participated in prior CAT and SAT studies that included the full item pool (see Vispoel, 1993; Vispoel et al., 1994). The mean for item difficulty parameters (b values) for the present item pool was 0.70 (*SD* = 0.80), and the mean for item discrimination parameters (a values) was 1.07 (*SD* = 0.41; see Vispoel (1998) for more detailed information concerning the distribution of these item parameters).

The CATs and SATs were scored using Bock and Mislevy's (1982) expected a posteriori (EAP) Bayesian ability estimation procedure. The initial prior distribution for EAP ability estimates was assumed to be normal with a mean of 0.7 and a standard deviation of 1.00. The prior mean was set higher than the original calibration sample mean of 0.00 because the college students sampling here were expected to be higher in average vocabulary ability than the Iowa high school students who were used for item calibrations.

Directions that preceded the administration of the CATs informed examinees that they were going to take a test in which items were chosen based on their performance on previously answered items. They were told that more difficult items typically would be administered after correct responses and that easier items typically would be administered after incorrect responses. The actual CATs began with an item from the pool that provided maximum information at the initial prior mean for the EAP ability estimates (0.7). The computer then administered the most informative item remaining in the pool based on the examinee's ability estimate after each subsequent item was scored.

Directions that preceded the SATs informed examinees that they were about the take a test in which they were to choose the difficulty level of each item administered. They were told that they could choose items from any difficulty level, but they were urged to choose the hardest items that they thought they could answer correctly. Examinees chose among eight available difficulty categories (1 = very easy, 8 = very difficult) before answering each item. There were 25 items available in each category, and the most discriminating remaining item from a given category always was administered. If examinees exhausted all items within a given category, they were asked to choose another category.

In the no-review conditions, examinees were told to answer each item to best of their abilities and that they *would not have* the opportunity to review and change their answers to a given item after they went on to the next item. In the review conditions, examinees also were told to answer each item to the best of their abilities, but that they *would have* the opportunity to review and change their answers after they had answered all items. On all computerized tests, examinees were asked to press a verify key before their answers were officially recorded. On the SATs, examinees also were asked to confirm that their intended level of item difficulty was correct. Answer feedback was not provided on any test, and the CAT and SAT item selection algorithms were not adjusted to balance content or to control item exposure rate. All tests were terminated at 30 items.

After each CAT and SAT item was administered, the computer recorded the item sequence number, the item pool number, the SAT item difficulty category, the examinee's answer, the correct answer, the item score (1 if correct, 0 if incorrect), the examinee's current EAP ability estimate, the posterior variance for that ability estimate, and the time spent answering the item. For examinees in the review conditions, the computer also recorded changed answers, time spent on each item in review, and the ability estimate and posterior variance after all answer changes were made. At the end of the test, each examinee's final ability estimate, the final posterior variance, and total testing time (excluding directions) were summarized and recorded. All tests were administered on IBM-AT microcomputers and scored using the Computerized Adaptive Testing System (CATSYS; de la Torre & Vispoel, 1991)—a testing program that yields results comparable to MicroCAT™ (Assessment Systems Corporation, 1989).

Several steps were taken in administering the computerized tests to approximate the climate of a realistic testing situation and to motivate examinees to do their best. First, a 20-minute time limit was imposed. Second, examinees were told that their percentile rank scores would be posted according to the last four digits of their student IDs after the experiment was completed. Third, instructions for the test emphasized the importance of trying hard. Finally, examinees were told that individuals who received one of the five highest vocabulary scores would each receive $20. Taking such steps has been shown to be effective in eliciting hypothesized interactions between administration mode and test anxiety in other studies (e.g., Rocklin, et al., 1995; Vispoel & Coffman, 1994; Vispoel et al., 1992).

Other Measures

The investigator-designed questionnaires included items assessing demographics and attitudes about computerized vocabulary tests (see Vispoel, 1993; Vispoel et al., 1994). The results reported here for attitudes are limited to examinees' perceptions of how important item review options are in creating computerized tests that accurately measure vocabulary skills. These perceptions were assessed using an 8-point scale (1 = extremely unimportant, 2 = very unimportant, 3 = moderately unimportant, 4 = slightly unimportant, 5 = slightly important, 6 = moderately important, 7 = very important, 8 = extremely important).

Procedure

Participants completed the measures in individual carrels at a computer lab on the university campus during a 50-minute research session. They were assigned at random to four testing conditions (CAT with review, CAT without review, SAT with review, SAT without review). This was accomplished by creating random orders of the four conditions (1234, 1324, 3241, etc.) using a random number table, and assigning participants sequentially to those conditions as they entered the computer lab. The demographics questionnaire was administered first, followed by the TAI, computerized test, and questionnaire about attitudes toward the computerized test, respectively.

Analyses

The effects of item review, administration mode, and test anxiety on ability estimates, measurement precision, and testing time (Research Question 1) were evaluated using heirarchical multiple regression procedures reminiscent of those commonly used in aptitude by treatment interaction studies (Cronbach & Snow, 1977). Here, however, test anxiety served as the individual difference variable instead of aptitude, and administration mode and item review served as the treatment variables. A separate regression analysis was run for vocabulary ability estimates, estimated measurement error (as indexed by posterior variance), and testing time (as indexed by seconds needed to complete the computerized tests excluding directions). In each regression analysis, the independent variables were entered one-by-one in a predetermined order beginning with test anxiety, followed by the treatment variables (administration mode, item followed by the two-way interactions (Mode × Review, Test Anxiety × Mode, Test Anxiety × Review), and ending with the three-way interaction (Test Anxiety × Mode × Review). The final residual term was used to test for statistical significance at each step. Because of the large number of significance tests and a concern about the practical significance of the findings, an alpha level of .001 was adopted for all reported analyses.

The congruence of the present answer changing results with those from prior studies (Research Question 2) was evaluated within the item review conditions by tabulating percentages of total answers that were changed, percentages of answers changed from wrong to right and from right to wrong, percentages of examinees who changed an answer to at least one item, and the percentages of examinees whose ability estimates increased, decreased, or remained the same after answer changing. These results were examined further by partitioning examinees into three vocabulary ability groups. Examinees' perceptions of the importance of including item review (Research Question 3) was evaluated by determining the extent to which the examinees felt that inclusion of item review was important or unimportant in obtaining accurate measurement of vocabulary skills from computerized tests.

Results

Effects of Item Review, Administration Mode, and Test Anxiety on Ability Estimates, Measurement Error, and Testing Time

The multiple regression results for the effects of administration mode, review, and test anxiety on vocabulary ability estimates, estimated measurement precision, and testing time appear in Table 1. Dependent variable means by administration mode and review condition appear in Table 2.

Ability estimates, estimated measurement error and testing time. The multiple regression results for ability estimates in Table 1 show that there was a significant test anxiety ($p < .0001$) main effect, which contributed .08 to the R^2 value. Individuals higher in test anxiety had lower ability estimates than did examinees lower in test anxiety ($r = -.28$). The results for estimated measurement error (i.e., posterior variance) yielded a significant administration mode main effect ($p < .0001$), which contributed .16 to the R^2 value. CATs provided more precise results than the SATs (posterior variance = .04 vs. .06), but the inclusion or exclusion of review had little effect on precision. Results for testing time revealed significant administration mode ($p < .001$) and review ($p < .0001$) main effects, which added .03 and .18, respectively, to the R^2 value. SATs took longer on average than did CATs (563 sec. vs. 488 sec.), and tests with review took longer than did tests without review (612 sec. vs. 436 sec.). Specifically, review options increased testing time across conditions by about 41%; using self adaptation rather than computer adaptation increased testing time by about 16%; and using both review and self-adaptation increased testing time by about 63%.

Answer Changing Behavior Within the Review Condition

Table 3 shows the results for answer changing behavior within the review condition. Consistent with previous research, examinees on average changed answers to a small percentage of items (5.23%); they changed more answers from wrong to right than from right to wrong (by a ratio of 2.92:1); they took advantage of their opportunity to review and change answers (66.49% changed answers to one or more items); and they generally benefited from changing answers (ability estimates improved by .04 on average, and 2.55 times as many examinees increased rather than decreased their ability estimates by changing answers). The correlation between ability estimates before and after review was .99 in both the CAT and SAT conditions. Measurement precision ratios (mean posterior variance before review divided by mean posterior variance after review) were .983 and .991 for the CAT and SAT, respectively.

Table 4 shows results for answer changing behavior for individuals in the lower, middle, and upper third of the vocabulary ability distribution after review.[3] The results reveal a series of systematic trends associated with increases in ability level. As ability level increases, testing time and number of answer changes decreases, but the benefits of answer changing increase in that ability estimate gains increase, item gain to loss ratios increase, and ability estimate gain to loss

Table 1
Multiple Regression Results for Vocabulary Ability Estimates, Posterior Variance, and Testing Time ($\underline{n} = 379$)

	Ability Estimate		Posterior Variance		Testing Time	
	F	R^2 dif.	F	R^2 dif.	F	R^2 dif.
Test Anxiety	33.57[b]	.08	3.68	.01	2.48	.01
Mode	2.84	.01	71.49[b]	.16	12.74[a]	.03
Review	0.09	.00	0.61	.00	68.14[b]	.18
Mode x Review	0.02	.00	2.47	.00	0.47	.00
Test Anxiety x Mode	5.24	.01	1.56	.00	2.51	.01
Test Anxiety x Review	0.04	.00	1.48	.00	1.40	.00
Test Anxiety x Mode x Review	0.80	.00	0.35	.00	0.00	.00

[a] $\underline{p} < .001$, [b] $\underline{p} < .0001$

ratios increase. Reflecting the more pronounced changes in ability estimates in the high ability group, measurement error (i.e., posterior variance) in that group also shows the largest increase following review. This result makes sense because items selected to maximize precision based on ability estimates before review are the most mismatched to the ability estimates after review in the high ability group.

Attitudes About the Importance of Including Item Review

Means for examinees' attitudes about the importance of including review options on computerized vocabulary tests ranged from 5.97 to 6.41 on an 8-point scale (1 = extremely unimportant, 8 = extremely important; see Table 2) across the test type and review conditions. Over the entire sample, review was desired (i.e., given a rating of 5 or higher) by 85% of the examinees, ranging from 82% in the CAT and SAT no-review conditions to 89% in the CAT and SAT review conditions. In addition, the desire for item review options did not vary significantly with level of test anxiety ($r = .07$, $p = .17$, two-tailed) or with final vocabulary ability estimates ($r = -.09$, $p = .08$, two-tailed).

Discussion

No single investigation is likely to provide an adequate evaluation of the effects of reviewing and changing answers on computerized tests. The results reported here are limited to a low-stakes testing situation involving a modest number of college students who were permitted to review and change answers only after they had completed all items. Nevertheless, the results are consistent with findings for

Table 2
Means and Standard Deviations for Dependent Variables by Administration Mode and Revie
Condition (n = 379)

	Ability Estimate		Posterior Variance		Testing Time (sec)		Perceived Importance of Review Options		
Condition	M	SD	M	SD	M	SD	M	SD	n
CAT review	1.43	.79	.04	.02	567.1	200.9	6.41	1.75	96
CAT no review	1.41	.72	.04	.01	403.6	129.7	6.06	1.66	95
SAT review	1.58	.77	.06	.02	656.7	257.0	6.41	1.52	95
SAT no review	1.50	.86	.06	.04	467.4	111.4	5.97	1.81	93
CAT	1.42	.75	.04	.01	482.6	186.4	6.24	1.71	191
SAT	1.54	.81	.06	.03	560.7	218.2	6.19	1.68	188
Review	1.50	.78	.05	.02	611.6	234.0	6.41	1.63	191
No review	1.46	79	.05	.03	434.6	124.9	6.02	1.73	188
Total sample	1.48	.79	.05	.03	521.0	206.0	6.21	1.69	379

answer changing behavior on objective tests that date back at least 70 years. When given the opportunity to review and change answers, most examinees will do so, but they will generally change answers to only a small proportion of items. These changes typically benefit the examinees because their scores or ability estimates improve slightly on average and their wrong to right answer changes outnumber their right to wrong answer changes. A new finding observed here for adaptively administered tests was that individuals with high estimated ability benefited the most from having review opportunities.

The decision to include versus exclude review is not a difficult one to make for computerized FITs in which all examinees get the same items. Because these options are generally available and difficult to exclude on paper-and-pencil FITs, it seems reasonable to allow them on computerized versions of those tests. Many test developers and researchers, in fact, have advocated the inclusion of review options on computerized tests as a way to enhance the comparability of computerized and paper-and-pencil FIT scores (e.g., Mazzeo & Harvey, 1988; Spray, Ackerman, Reckase, & Carl, 1989). Programming computers to allow review, answer change, and skip options is also much easier to do with computerized FITs than with CATs or SATs because the same items are always administered, and ability estimates do not have to be made on an item-by-item basis. Issues concerning possible context effects or score inflation strategies on the validity of results are irrelevant because the context is the same for all examinees, and the score inflation strategies (to be discussed shortly) only apply to adaptively administered tests. Consequently, allowing item review on computerized FITs seems appropriate unless test efficiency is seriously undermined or the construct in question is measured more accurately with only one exposure to each item.

Table 3
Descriptive Statistics for Answer Changing Behavior in the Review Conditions

Index	CAT	SAT	Total
Ability estimate means			
Before review	1.39	1.54	1.47
After review	1.43	1.58	1.51
Difference (after-before)	0.04	0.04	0.04
Number correct score means[a]			
Before review	21.28	20.64	20.96
After review	21.60	21.02	21.31
Difference (after-before)	0.32	0.38	0.35
Item changes			
Mean number	1.32	1.81	1.57
Mean percent[b]	4.41	6.04	5.23
Item gains (W to R)			
Mean number	0.60	0.85	0.73
Mean percent[b]	2.01	2.84	2.43
Item losses (R to W)			
Mean number	0.28	0.47	0.38
Mean percent[b]	0.94	1.58	1.27
Item (W to W)			
Mean number	0.44	0.48	0.46
Mean percent[b]	1.46	1.61	1.54
Item gain to loss ratio	2.15:1	1.80:1	1.92:1
Examinees who changed at least one answer			
Number	58	69	127
Percent	60.4	72.6	66.5
Ability estimate gains			
Number	40	44	84
Percent	41.7	46.3	44.0
Ability estimate losses			
Number	13	20	33
Percent	13.5	21.1	17.3
Ability estimates unchanged			
Number	43	31	74
Percent	44.8	32.6	38.7
Ability estimate gain to loss ratio	3.08:1	2.21:1	2.55:1

[a]Number correct score equals the sum of correctly answered items that each examine to
[b]Mean percent = (mean number/30) * 100

Decisions about whether or when to allow item review on adaptively administered tests, however, are far less clear because there are trade-offs associated with allowing versus disallowing review as well as important questions about review options on such tests that have not been adequately resolved. The two most commonly cited reasons for allowing review on adaptive tests are increased exam-

Table 4
Answer Changing Results By Ability Group

	Low Ability (n = 64)	Middle Ability (n = 64)	High Ability (n = 63)
Mean ability estimate change (after-before)	.02 (.07)	.03 (.09)	.05 (.09)
Mean number of changed answers	2.05 (2.06)	1.58 (1.72)	1.06 (1.70)
Mean number of wrong to right answer changes	0.78 (0.97)	0.77 (0.99)	0.63 (0.77)
Mean number of right to wrong answer changes	0.52 (0.69)	0.36 (0.57)	0.25 (0.58)
Item gain to loss ratio	1.52:1	2.13:1	2.50:1
% of examinees with at least one answer change	70.3	65.6	63.5
% of examinees with ability estimate gains	40.6	45.3	46.0
% of examinees with ability estimate losses	23.4	17.2	11.1
Ability estimate gain to loss ratio	1.73:1	2.64:1	3.14:1
Mean posterior variance change (after - before)	0.000 (0.009)	0.000 (0.003)	0.002 (0.006)
Mean testing time (sec)	634.7 (300.4)	620.2 (198.3)	579.5 (182.1)

Note: Values in parentheses represent standard deviations

inee satisfaction and increased test validity. Research to date clearly demonstrates that most examinees want to be able to review and change their answers and that they generally will take advantage of these opportunities. As a result, providing review opportunities would seem to provide a less stressful testing environment compared to when review opportunities are denied. In addition, some examinees may view having review options as an inherent right when taking any kind of test. (Lunz et al., 1992; Stocking, 1997; Stone & Lunz, 1994).

Allowing answer changes following review also could increase test score validity if the changes reflect corrections of typing errors, misreadings of items, temporary lapses in memory, or reconceptualizations of answers to previously administered items. Under these conditions, item review would yield more valid scores because the scores would represent the examinee's skill level at the end of the test more accurately, and the scores would not be contaminated with clerical or other inadvertent errors. Support for the increased validity of scores following review comes from studies in which examinees have specified the reasons why they changed answers (Ferrara et al., 1996; McMorris, DeMers, & Schwarz, 1987; McMorris & Weiderman, 1986; Schwarz, McMorris, & DeMers, 1991; Shatz & Best, 1987). In these studies, examinees typically cite learning answers from subsequent items as a reason for answer change less frequently than other reasons such as clerical errors, rereading items, rethinking and conceptualizing better answers, and guessing other answer alternatives. McMorris et al., (1987), for example, found that only one out of 76 answer changes from wrong to right was attributed to learning from subsequent items. Ferrara et al. (1996) also found little evidence of answer changing due to learning effects from subsequent items in the only study to date that has focused on reasons for answer changes on CATs.

Another important issue related to validity concerns the possible adverse impact of disallowing review on scores for certain individuals. This issue was addressed to

some extent in the study reported here with regard to individual differences in test anxiety and vocabulary ability. The absence of a Test Anxiety × Review interaction for vocabulary ability estimates in the between-group analyses involving review and no-review groups provided no trustworthy evidence that examinees with higher test anxiety profited any more from answer review than did examinees with lower test anxiety. However, when results within the review group were inspected further by dividing examinees into low, middle, and high vocabulary ability categories, it was apparent that review was most beneficial to examinees with high estimated vocabulary ability. As a result, excluding review options on adaptively administered test may have its most adverse impact on scores for those examinees. This conclusion, however, needs to be confirmed through further research, particularly under higher-stake conditions and within other content domains and item banks.

Although additional research into validity-related issues is clearly needed, it is apparent that some empirical support exists for allowing review on computerized tests. Nevertheless, many test developers and researchers remain unconvinced that review is appropriate on adaptively administered tests. Problems associated with allowing review include complications in item administration and scoring algorithms, increases in testing time, reductions in measurement precision, and reductions (rather than improvements) in test score validity.

There is little doubt that the inclusion of review options complicates the algorithms for administering and scoring CATs and SATs. These algorithms would have to be expanded to allow examinees to revisit any items chosen from the item pool, change answers to those items, and incorporate any answer changes into final ability estimates and measurement precision indices. The algorithms would be complicated further when fixed measurement precision rules are used to terminate the tests. When answer changes reduce the precision of ability estimates to a point where they no longer meet the termination criterion, additional items would need to be administered. The answers to those items might also be changed, and still additional items may need to be administered.

The notion that allowing review on adaptively administered tests will increase testing time received strong support here and in two other studies (Vispoel et al., 1992; Ferrara et al., 1996), with increases in testing time ranging from 37% to 61%. As a result, the inclusion of review options would seem to undermine one of the main advantages of using adaptive tests. Decreases in measurement precision resulting from the inclusion of review also has received some empirical support but this evidence is far weaker than that concerning testing time. No trustworthy differences in test precision between the review and no review groups were noted in the study reported here, and the significant differences found in other studies of adaptively administered tests have not been of great practical importance (Lunz et al., 1992; Stone & Lunz, 1994; Vispoel et al., 1992). Across studies of adaptively administered tests, for example, measurement precision ratios based on Fisher test information or Bayesian posterior variance have equaled or exceeded .97. However, these studies were conducted under conditions in which examinees did not seem to use strategies that might lead to inflated ability estimates.

Examinee's abilities to artificially inflate scores on adaptively administered tests can seriously undermine test score validity. One way that review could lead to

inflated scores is when clues or other information from subsequent items helps examinees answer previous items correctly. Possible clues are of greater concern in tests with review because examinees can revisit items several times and thereby increase their changes of spotting the clues. If such clues were present in some items but not in others, some examinees taking adaptively administered tests would benefit and others would not because all examinees do not receive the same items. Although results from studies cited earlier (Ferrara et al., 1996; McMorris et al., 1987; McMorris & Weiderman, 1986; Schwarz et al., 1991; Shatz & Best, 1987) do not completely rule out the use of clues in changing answers, they do indicate that most answer changes are made for more legitimate reasons. In addition, clues could be eliminated or at least significantly reduced by carefully screening and checking items in the pool for overlap. Developers of item response theory-based adaptive tests, in fact, routinely assume that item responses are locally independent or free from such overlap. Furthermore, the screening done for item overlap on operational CATs and SATs is likely to be more extensive than the screening done on the classroom tests used in most studies that have addressed reasons for answer changes.

A second and more serious threat to the validity of scores from adaptive tests with review is artificially inflated scores achieved through clever answer response strategies. One such strategy for inflating ability estimates on fixed-length CATs was suggested by Wainer (1993). In using this strategy, examinees would intentionally answer all items incorrectly before review, and thereby receive the easiest possible items that the adaptive testing algorithm could provide. During review, the examinees would answer the items to best of their abilities. If they could answer all of these items correctly, then they would receive unbounded high scores (e.g., positive infinity on the theta metric) if maximum likelihood ability estimation were used.

Research into the Wainer strategy has revealed that there are instances when it does lead to inflated ability estimates and that these ability estimates usually have large standard errors (Gershon & Bergstrom, 1995; Stocking, 1997; Vispoel, Rocklin, Wang, & Bleiler, in press). Such findings suggest that the Wainer strategy might be dealt with either by flagging unreliable ability estimates as invalid or by using a standard error stopping rule for the CAT. Unfortunately, both of these approaches have serious drawbacks. Flagging suspicious ability estimates would embarrass examinees, necessitate alternative retesting, and be open to legal challenges. Using a standard error stopping rule could expose large portions of the CAT item bank, yield unreasonably long tests for some examinees, complicate algorithms for balancing content and item exposure rate, require administration of additional items when ability estimates following review failed to meet the termination criterion, and would not be an option for fixed-length CATs such as the Graduate Records Examination (GRE). A third and perhaps more practical way to deal with the Wainer strategy is to limit review opportunities on CATs. Stocking (1997) found, for example, that restricting review to separately timed blocks of items or to items linked to a common stimulus (e.g, reading passages, tables, graphs) provided acceptable controls over bias and measurement precision.

A strategy aside from Wainer's that might be used to obtain inflated ability estimates on CATs was described by Kingsbury (1996; see also Green, Bock, Humphreys, Linn, & Reckase, 1984). This strategy is based on examinees' abilities

to infer whether they got items correct or incorrect based on the perceived relative difficulty of sequentially administered items. Because a CAT algorithm typically provides harder items after correct responses and easier items after incorrect responses, examinees can infer that they got the previous item wrong if the next item is easier and that they got the previous item right if the next item is more difficult. When reviewing the items, examinees would supply different answers to any items that were followed by easier ones. The success of this strategy would depend on the examinee's ability to recognize the relative difficulty levels of items and the extent to which the adaptive testing algorithm adheres to the rule of administering harder items after correct responses and easier items after incorrect responses. As a CAT progresses, differences in item difficulty will become more and more subtle, and the item selection algorithm is more likely to deviate from a correct-harder/incorrect-easier rule due to variations in item discriminations (e.g., items at ability extremes tend to be less discriminating than items at medium ability levels). As a result, easier items might be administered after correct responses or harder items might be administered after incorrect responses if those items provide more information about the examinee's ability level. In addition, even if the examinees are successful at using this strategy, they have only eliminated one incorrect answer choice for a given item.

Kingsbury (1996) evaluated the success of this strategy using computer simulation techniques based on the assumption that an examinee could recognize difficulty (i.e., b value) differences of .5 or higher. He found that the strategy led to high gains in ability estimates for low ability examinees, moderate gains for middle ability examinees and essentially no gains for high ability examinees. No study to date, however, has examined the effectiveness of this strategy on an operational CAT when examinees are taught to use the strategy. Consequently, the effectiveness of this strategy remains an unanswered question. There also are some safeguards that might to be used to reduce its potential effectiveness. One safeguard would be to employ a fixed-item routing test at the beginning of the CAT. After the routing test, differences between item difficulties would be less obvious. A second safeguard would be allow review on CATs in the ways suggested by Stocking (1997), as long as items are not administered adaptively within a given block.

Given the empirical findings about review discussed here, it seems reasonable to conclude that the evidence to date is not strong enough to rule out the inclusion of review on adaptively administered tests as long as certain precautions are taken. Item pools should be screened thoroughly for overlap and safeguards should be taken to reduce the possibility of examinees obtaining artificially inflated ability estimates. Strategies suggested by Stocking (1997) currently seem to offer viable compromises in integrating review into CATs under conditions in which willful manipulation of ability estimates might be attempted.

While acknowledging that review *can be* allowed on adaptively administered tests, the question still remains as to whether it *should be*. This is not an easy question to answer because compelling arguments have been made both for and against review. Although programming is more complicated, it is certainly possible to develop algorithms for CATs and SATs that allow review and answer change, and safeguards also can be taken to reduce the effectiveness of score inflation

strategies. Concerns about the measurement precision of scores following review also can be addressed. Evidence to date shows that under normal conditions the precision and magnitude of scores do not change markedly when review is allowed versus disallowed and that the rank ordering of scores before and after review changes little. Techniques for allowing review suggested by Stocking (1997) also seem to yield CAT scores with adequate degrees of measurement precision even when many answer changes are made. The slight reductions in precision that occasionally have been observed when review is allowed might be handled by extending test lengths by a small number of items.

Aside from the unresolved questions about test validity that need further investigation (i.e., reasons for answer change on CATs and SATs, adverse impact of disallowing review for certain individuals, effectiveness of score inflation strategies), the crucial factor that would lead one to allow or disallow review appears to be whether examinee satisfaction or test efficiency is of more concern to the test user. It is clear from the research to date that most examinees are likely to desire review options, but that such options can increase testing time rather substantially. Examinees' negative attitudes toward disallowing review might be dealt with by allowing them to choose between taking a fixed-item paper-and-pencil test in which review is allowed or a CAT in which review is disallowed. When examinees make such a choice, it may often be the case that positive feelings about computerized testing outweigh negative attitudes about the exclusion of review options. Alternatively, CATs could be constructed using techniques suggested by Stocking (1997) in which opportunities for review are more constrained.

As a final note, it is important to emphasize that the present study and most previous ones (Ferrara et al., 1996; Lunz et al., 1992; Stone & Lunz, 1994; Vispoel et al., 1992) revealed little evidence of examinees attempting to use the Wainer or Kingsbury strategies to obtain positively biased ability estimates on CATs. If approaches such as those suggested by Stocking (1997) are effective in discouraging the use of these strategies, then results from the studies above may still reflect general patterns of answer changing behavior on CATs. Nevertheless, answer changing behavior is likely to be considerably different when examinees attempt to implement score inflation strategies (see, e.g., Vispoel et al., in press). The effects of such behavior on CAT results under high stakes live testing conditions is an important topic for future research.

Notes

[1]See the discussion section of this paper and papers by Lunz, Bergstrom, and Wright (1992); Stone and Lunz (1994); Vispoel, Wang, de la Torre, Bleiler, and Dings (1992); Wainer (1993); Wang and Wingersky (1992); and Wise (1996) for more detailed analyses of the pros and cons regarding answer review options on adaptive tests.

[2]See Rocklin (1996), Vispoel, Rocklin, and Wang (1994); Vispoel and Coffman (1994), and Wise (1994) for discussions of alternative explanations for why SATs might reduce the effects of test anxiety on ability estimates.

[3]The same pattern of results emerged when ability groupings were based on ability estimates before review. The results in Table 4 represent the lower, middle, and upper third of the ability distribution for the two administration modes pooled together.

References

Assessment Systems Corporation (1989). *User's manual for MicroCAT*. St. Paul, MN: Author.

Baghi, H., Gabrys, R., & Ferrara, S. (1991, April). *Applications of computer-adaptive testing in Maryland*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Benjamin, L., Cavell, T. A., & Shallenberger, W. R. III (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*, 133-141.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 37*, 431-444.

Crawford, C. (1928). *The technique of study*. Boston: Houghton Mifflin.

Crocker, L., & Benson, J. (1980). Does answer changing affect test quality? *Measurement and Evaluation in Guidance, 12*, 223-239.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York City: Irvington.

Cummings, O. W. (1981). Impact of response changes on objective test characteristics and outcomes for junior high school students. *Measurement and Evaluation in Guidance, 14*, 32-37.

de la Torre, R., & Vispoel, W. P. (1991, April). *The development and evaluation of a computerized adaptive testing system*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Feldt, L. S., Forsyth, R. A., & Alnot, S. D. (1986a). *Iowa Tests of Educational Development: Form X-8*. Iowa City, IA: The University of Iowa.

Feldt, L. S., Forsyth, R. A., & Alnot, S. D. (1986b). *Iowa Tests of Educational Development: Form Y-8*. Iowa City, IA: The University of Iowa.

Feldt, L. S., Forsyth, R. A., & Lindquist, E. F. (1979a). *Iowa Tests of Educational Development: Form X-7*. Iowa City, IA: The University of Iowa.

Feldt, L. S., Forsyth, R. A., & Lindquist, E. F. (1979b). *Iowa Tests of Educational Development: Form Y-7*. Iowa City, IA: The University of Iowa.

Ferrara, S., Frances, A., Gilmartin, D., Knott, T., Michaels, H., Pollack, J., Schuder, T., Vaeth, R., & Wise, S. (1996, April). *A qualitative study of the information examinees consider during item review on a computer-adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York City.

Gershon, R. C., & Bergstrom, B. (1991, April). *Individual differences in computer-adaptive testing: Anxiety, computer literacy and satisfaction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Gershon, R. C., & Bergstrom, B. (1995, April), *Does cheating on CAT pay: Not*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Kingsbury, G. G. (1996, April). *Item review and adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York City.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11*, 23-27.

Lindquist, E. F., & Feldt, L. S. (1972). *Iowa Tests of Educational Development: Form X-6*. Iowa City, IA: The University of Iowa.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16*, 33-40.

Matthews, C. O. (1929). Erroneous first impressions on objective tests. *Journal of Educational Psychology, 20*, 280-286.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

McMorris, R. F., DeMers, L. P., & Schwarz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement, 24*, 131-143.

McMorris, R. F., & Weiderman, A. H. (1986). Answer changing after instruction on answer changing. *Measurement and Evaluation in Counseling and Development, 18*, 93-101.

Moe, K. C., & Johnson, M. F. (1986). *Participants' reactions to computerized testing*. ERIC Document Reproduction Service No. ED 282935.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14*, 9-13.

Rocklin, T. (1996). Self-adapted testing: Improving performance by modifying tests instead of examinees. *Anxiety, Stress, and Coping, 10*, 83-104.

Rocklin, T. R., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.

Rocklin, T. R., O'Donnell, & Holst, P. M. (1995). Effects and underlying mechanisms of self-adapted testing. *Journal of Educational Psychology, 87*, 103-116.

Roos, L. L., Wise, S. L., & Plake, B. S. (1997). The role of feedback in self-adapting testing. *Educational and Psychological Measurement, 57*, 85-98.

Schmitt, F. L., Urry, V. W., & Gugel, J. J. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement, 38*, 265-273.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement, 28*, 163-171.

Shatz, M. A., & Best, J. B. (1987). Students' reasons for changing answers on objective tests. *Teaching of Psychology, 14*, 241-242.

Smith, M., White, K., & Coop, R. (1979). The effect of item type on the consequences of changing answers on multiple choice tests. *Journal of Educational Measurement, 16*, 203-208.

Spielberger, C. D. (1980). *Preliminary professional manual for the Test Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carl, J. E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261-271.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement, 21*, 129-142.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7*, 211-222.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Vispoel, W. P. (1993). Computerized adaptive and fixed-item versions of the ITED Vocabulary subtest. *Educational and Psychological Measurement, 53*, 779-788.

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement, 35*, 155-167.

Vispoel, W. P., & Coffman, D. D. (1994). Computerized adaptive and self-adapted tests of music listening skills: Psychometric features and motivational benefits. *Applied Measurement in Education, 7*, 25-51.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, adaptive, and self-adapted testing. *Applied Measurement in Education, 53*, 53-79.

Vispoel, W. P., Rocklin, T. R., Wang, T., & Bleiler, T. (in press). Can examinees use a review option to positively bias their scores on a computerized adaptive test? *Journal of Educational Measurement.*

Vispoel, W. P., Wang, T., & Bleiler, T. (1997). The efficiency, reliability, and concurrent validity of adaptive and fixed-item music listening tests. *Journal of Educational Measurement, 34*, 43-63.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). *How review options, administration mode, and anxiety influence scores on computerized vocabulary tests.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service, No. TM018547)

Waddell, D. L., & Blankenship, J. C. (1995). Answer changing: A meta-analysis of the prevalence and patterns. *Journal of Continuing Education in Nursing, 25*, 155-158.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*, 15-20.

Wang, M., & Wingersky, M. (1992). *Incorporting post-administration item response revision into a CAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

Wise, S. L. (1994). Understanding self-adapted testing: The perceived control hypothesis. *Applied Measurement in Education, 7*, 15-24.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and algorithmic adaptive achievement tests. *Journal of Educational Measurement, 29*, 329-339.

Author

WALTER P. VISPOEL is Associate Professor, University of Iowa, 361 Lindquist Center, Iowa City, IA 52242-1529; walter-vispoel@uiowa.edu. *Degrees:* BA, MEd, University of Illinois at Chicago, PhD, University of Illinois at Urbana-Champaign. *Specializations:* educational measurement, educational psychology.

Journal of Educational Measurement
Winter 1998, Vol. 35, No. 4, pp. 346–351

Handbook of Modern Item Response Theory

Wim. J. van der Linden and Ronald K. Hambleton
New York: Springer-Verlag, 1997

Reviewed by
Terry Ackerman
University of Illinois

Psychometricians and measurement experts working in either the private sector or in academia have been waiting for some time for more comprehensive texts on item response theory (IRT). Seminal sources to date that are written explicitly about IRT include Lord's *Application of item response theory to practical testing problems*, 1980, and Hambleton and Swaminathan's *Item response theory, Principles and applications*, 1985. Two relatively recent additions include Baker's *Item response theory, Parameter estimation techniques*, 1992, and Fisher and Molenaar's *Rasch Models: Foundations, recent developments and applications*, 1995. Over the past decade there have been many creative advances in IRT, but not all have appeared in traditional measurement journals. This is why van der Linden and Hambleton's book, *Handbook of Modern Item Response Theory*, is not only timely, but it also helps advance IRT research by providing another seminal source. This book represents an excellent compendium of current IRT research and covers a wealth of creative thought and applications of modeling with which psychometricians may not be familiar. Van der Linden and Hambleton have really done the measurement community a huge service by gathering this information into one text. It serves as a great resource that any serious measurement expert should read.

Wim Van der Linden and Ronald Hambleton begin the book by providing an historical overview of item response theory. This discussion opens with classical test theory, moves on to the one-parameter IRT model, the logistic Rasch model and Birnbaum's two and three parameter logistic models and then onto a brief discussion about polytomous and multidimensional models. This first chapter concludes with a brief commentary on the future directions of IRT.

Each chapter that follows has the same format. A model is developed from background material. The estimation of parameters and goodness-of-fit of the model are presented. Authors then present an example of an application of the model followed by a summary discussion section. Each chapter concludes with a complete list of references which can be used to further guide the reader who is looking for more information.

The book is divided into six sections, each dealing with a unique application of an IRT model or a family of models. A chapter by R. Darrell Bock describing the IRT model for nominal data begins the polytomous model section. A short develop-

ment of choice models and logit linear models leads to the nominal categories model. Bock illustrates a possible use by showing results of estimating parameters for ratings of 2000 examinee responses to three open-ended items from a human biology test.

David Thissen and Lynne Steinberg's multiple-choice model follows. As in the nominal case, this model has no a priori ordering of the response categories. The difference between the two models is that Thissen and Steinberg's model has an additional term for examinees who don't know the correct answer to the question and constitute a latent class by themselves. Illustrations provide more insight into the probability of selection for the different response categories.

The third chapter in this section, by Erling Andersen, details a rating scale model. This model is an extension of the Rasch model for polytomous data. Response data that fit this model are category scores that are assumed to be equally spaced. Andersen discusses both conditional and marginal maximum likelihood estimation (CML and MML) of the parameters. Residual analysis to determine data points contributing to lack of fit is outlined.

The fourth chapter is by one of the major founders of polytomous models, Fumiko Samejima. She presents the graded response model and distinguishes between the homogeneous case and the heterogeneous case. In the homogeneous case characteristic functions are identical in shape. Included in this group are both the normal ogive and logistic models. The heterogeneous case represents all probabilistic models that provide a set of cumulative response functions that do not have identical shapes (e.g., Bock's nominal response model). Given certain constraints, Masters' partial credit model and Muraki's generalized partial credit model can be considered to be special cases.

Geoffery Masters and Benjamin Wright develop their partial credit model next. Unlike Samejima's graded response model, this model belongs to the Rasch family of models and thus features separable person and item parameters and sufficient statistics. The authors provide an interesting graphical perspective about how to interpret the estimated parameters. Three different estimation approaches—CML, MML and joint maximum likelihood (JML)—are detailed.

A steps model to analyze partial credit is presented by Norman Verhelst, Cees Glas, and Hannah de Vries. This model provides a more mathematically tractable alternative to the Masters and Wright partial credit model. Item parameters in this model are unambiguously tied to the separate response categories and thus can be interpreted as difficulty parameters. A typical assessment scenario in which this model could be employed is one that involves multistage testing. One drawback to this model is the estimation of parameters: CML estimation cannot be performed and MML estimation is much more complicated than for the partial credit model.

Another variation of the partial credit model is the sequential model for ordered response developed by Gerhard Tutz. This model, the seventh of this section, is applicable in testing scenarios in which an examinee solves the problem in consecutive or sequential steps. The difficulty parameter in this model is not the conditional difficulty in the Rasch family of models but rather is a local step difficulty. Tutz discusses both JML and MML estimation methods.