European Journal of Psychological Assessment, Vol. 16, Issue 1, pp. 3–12

The Choice of Item Difficulty in Self-Adapted Testing

Pedro Hontangas[1], Vicente Ponsoda[2], Julio Olea[2], and Steven L. Wise[3]

[1]Universidad de Valencia, Spain,
[2]Universidad Autónoma de Madrid, Spain,
[3]University of Nebraska-Lincoln, USA

Keywords: Self-adaptive testing, computerized testing, difficulty level choice

Summary: The difficulty level choices made by examinees during a self-adapted test were studied. A positive correlation between estimate ability and difficulty choice was found. The mean difficulty level selected by the examinees increased nonlinearly as the testing session progressed. Regression analyses showed that the best predictors of difficulty choice were examinee ability, difficulty of the previous item, and score on the previous item. Four strategies for selecting difficulty levels were examined, and examinees were classified into subgroups based on the best-fitting strategy. The subgroups differed with regard to ability, pretest anxiety, number of items passed, and mean difficulty level chosen. The self-adapted test was found to reduce state anxiety for only some of the strategy groups.

Progress in computer technology and the development of item response theory (IRT) have led to the appearance of new measurement instruments, the so-called "computerized adaptive tests" (CATs). Basically, a CAT is comprised of an IRT-calibrated item bank and a procedure for adapting item selection to each examinee (for example, an algorithm presenting easy or difficult items based on examinee's performance on items presented earlier in the test). CATs are more efficient than paper-and-pencil or computerized conventional fixed-item tests; they provide more precision with the same number of items or equivalent precision with fewer items (Lord, 1980; Wainer, 1990).

A CAT matches item difficulty to each examinee's ability, and it has been asserted that CATs provide a similar affective and motivational situation across examinees (Wainer, 1990). In a CAT, items challenge low- and high-ability examinees to the same degree, because the percentage of items passed is about the same (roughly half) for all examinees. However, some doubts have been expressed regarding whether CATs are appropriate from a motivational point of view. For example, Andrich (1995) stated that, "I wonder . . . whether the feature that people would only solve about 50% of the items correctly . . . is a good one. It seems to me that such a success rate may be too low to maintain the interest and enthusiasm of an examinee" (p. 617). Moreover, the motivational impact of a CAT may not be equivalent across examinees: Reactions to test difficulty have been shown to depend on examinees' motivational style (Atkinson & Litwin, 1970), test anxiety levels (Rocklin & Thompson, 1985), and self-concept (Vispoel, Rocklin & Wang, 1994). Thus, the homogeneous item passing rate that characterizes CATs may be incongruent with the affective and motivational states of examinees (Rocklin, O'Donnell, & Holst, 1995). Self-adapted testing (SAT) was proposed by Rocklin and O'Donnell (1987) as a method for allowing examinees to adapt their tests to their affective and motivational states as well as to their abilities.

In a SAT, the examinee takes an active role in the item-selection procedure which, in principle, allows the examinee to adapt the assessment process to his or her affective and motivational state. The item bank used in SAT is subdivided into between 5 to 9 difficulty levels,

or strata. Prior to each item, the examinee chooses the difficulty level from which the item to be administered will be drawn. This is in contrast with a CAT, in which no direct examinee participation is allowed and item selection is driven by an algorithm that depends solely on the examinee's responses to earlier items. Instructions in a SAT encourage the examinees to choose the most difficult items they think they can answer correctly, and feedback after each response is normally provided.

Most SAT research has focused on the advantages and disadvantages of SATs compared to CATs and fixed-item tests (see Rocklin, 1994; Wise, 1994). Despite some minor divergences, some general results are that self-adapted tests (a) decrease state anxiety, (b) reduce the correlation between anxiety and ability, (c) sometimes increase mean ability, (d) increase testing time, and (e) increase the standard error of ability estimation. In short, SAT tests yield scores that are less affected by anxiety at the cost of lower efficiency and precision than CATs.

SAT research to date, however, has not paid much attention to the item-selection process and the consequences it may have on precision and individual well-being. In SAT, the examinee is given the responsibility to choose the item difficulty, and psychometric consequences of this unique test type should be studied. For instance, an examinee may self-administer items that are too easy or too difficult – in which case, items will not be well matched to ability level, precision will suffer accordingly, and the estimated ability may be biased (Rocklin, Vispoel, & Wang, 1995).

Two previous reports have studied the item-selection process in SATs (Rocklin, 1989; Johnson, Roos, Wise, & Plake, 1991). These were their main conclusions:

a) Examinees selected easy items at the beginning of the test and increasingly difficult items as the test progressed.
b) Difficulty choices were correlated with anxiety, self-efficacy (perceived capability and perceived confidence), and score on the previous item (correct or incorrect).

Anxiety and self-efficacy were primarily related to initial item choices; examinees high in anxiety and/or low in self-efficacy chose easier initial items. The score on the previous item was increasingly related to difficulty choice as the test progressed (Johnson et al., 1991).

Rocklin (1989) proposed item-selection strategies, each based on the score from the previous item. First, an examinee using a "flexible" strategy would select the next higher difficulty level following a correct response and the next lower difficulty level following an incorrect response. Second, in the "failure-tolerant" strategy, the examinee would select the next higher difficulty level after a correct response, but the same level following an incorrect response. Finally, in the "failure-intolerant" strategy, the examinee would select the same difficulty level following a correct response, but would select the next lower difficulty level following an incorrect response. Rocklin (1989) found that most examinees applied the flexible strategy. Johnson et al. (1991) found, however, that these strategies were not strictly followed by most of their subjects. They instead found evidence of a strategy in which examinees tended to move to a more difficult item only after several correct responses and to a less difficult item only after several incorrect responses. Rocklin (1989) classified his subjects according to the three strategies cited above and carried out differential analyses among them. Failure-tolerant examinees had a lower ability mean than their flexible and failure-intolerant counterparts, although the three groups did not differ on standard error or pretest anxiety.

In the current study, item difficulty choices made by examinees in an actual classroom assessment situation were investigated. Examinees were administered an SAT as a class requirement and their test scores determined 20% of the second-term grade. This involves a difference regarding previous studies of SAT, in which there were little or no consequences associated with test performance. Specifically, we investigated

– the relationship between ability and difficulty choices,
– the changes in difficulty level choices throughout the testing session,
– the relative contributions of several predictors of difficulty level choice.

In addition, a new difficulty level strategy was proposed and its personal (anxiety) and psychometric consequences studied.

Method

Subjects

A total of 171 high-school students (54.4% senior-year students, 45.6% junior-year; 43.9% boys, 56.1% girls) took part in the study. Subjects age ranges from 16 to 19 years. One goal was to use an SAT in a situation as similar as possible to an ordinary class examination. To accomplish this goal, teachers agreeing to include the SAT score as part of the second-term marks were contacted. Their students were told that they would have their ordinary exam and a special exam (the SAT) both contributing to the second-term grade (80% from the ordinary exam, and 20% from the SAT). Thus, examinees were required to take part in the SAT just as they had to participate in any other course assessment.

Instruments

State anxiety was measured by the Spanish version of the State Anxiety scale of State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970), which is comprised of 20 Likert-type four-option items (scored from "0" to "3"). Reliability coefficients in the range of .90 to .93 and concurrent validity coefficients in the range of .82 to .88 have been reported in different Spanish samples (TEA, 1988). The scale was split into two 10-item subscales and administered by computer to evaluate an examinee's state anxiety immediately before (pretest) and after (posttest) the SAT test. To assess the equivalence of the two half scales, the responses of a heterogeneous sample of 1414 students and health workers were analyzed (Hontangas, Canela & Agustin, 1996). Equipercentile equating (Kolen, 1980) was applied to make comparable scores from the pretest (mean = .98, SD = .52) and posttest scales (mean = 1.00, SD = .49). In the current study, alpha coefficients for the two subscales were found to be .86 (pretest) and .87 (posttest).

The SAT used in the study measured English vocabulary in a Spanish-speaking population. The 221 multiple-choice items in the pool were each composed of six words: one English (the stem) and five Spanish (the options). The items were calibrated by applying the program ASCAL (Assessment Systems Corporation, 1989), and best fit was provided by the three-parameter logistic model. Means and standard deviations of the item parameters were –.08 and 1.5 (difficulty), 1.09 and .37 (discrimination) and .22 and .06 (pseudo-guessing). Validation studies revealed that the ability obtained from the entire bank showed significant correlation (.61) with the Oxford Placement Test (Allan, 1992), and that ability means were significantly different among groups with diverse English training (Ponsoda, Olea, & Revuelta, 1994; Olea, Ponsoda, Revuelta, & Belchí, 1996; Ponsoda, Wise, Olea, & Revuelta, 1997).

The item pool was divided into seven ordered difficulty levels, based on the difficulty parameters (from extremely easy to extremely difficult): $b \leq -2.5$ (13 items), $-2.5 < b \leq -1.5$ (30 items), $-1.5 < b \leq -.5$ (39 items), $-.5 < b \leq .5$ (58 items), $.5 < b \leq 1.5$ (47 items), $1.5 < b \leq 2.5$ (29 items) and $2.5 < b$ (5 items). Before the current vocabulary test started, examinees had a seven-item training phase – consisting of one item from each difficulty level – to allow them to experience the difficulty of items from each level (as in the "Route" condition of Plake, Wise, & Roos, 1995, but with no proficiency information provided based on the training items). When the test started, the examinee had to choose – prior to each item – one of the seven levels, after which the computer administered the most informative available item (relative to the examinee's last ability estimate) from the selected level. If the examinee selected a difficulty level for which no items remained, he or she was informed that no more items were available and was instructed to select another level. After responding to each item, an examinee received feedback regarding whether his or her response was correct or incorrect. Test length was fixed at 20 items, which is consistent with the test length used in several previous SAT studies (Rocklin, 1989; Johnson et al., 1991; Plake et al., 1995). The test administration program computed conditional maximum-likelihood estimates of ability, standard errors of estimation, number of correct responses, and response times (i. e., seconds elapsed from item appearance to examinee response). Two measures of item difficulty were collected – the difficulty level selected and the administered item's difficulty parameter.

Procedure

Tests were administered in a computer room containing eleven computers. Three research assistants met with each 11-student group, formed three small subgroups, and gave to them – in Spanish – a set of verbal directions. A translated version of the directions is as follows:

> "You are going to take part in an English test based on a new type of test developed at the university. 20% of your second-term grade in your English course will depend on your test score. So, please, do your best and try to get a good mark.
> Before and after the English test you will have to answer a few questions about how you feel at the moment. Please give honest answers to these questions; nobody will make improper use of them.
> The computer screen will give you more precise directions about how to provide your answers.
> In the English test you will have 15 seconds to answer each item. The computer will give you feedback about whether your response is correct or not. Before the English test, you will see seven training items. Please, pay attention to them as you will have to choose the difficulty of the items you have to answer. Each training item comes from a different difficulty level to allow you to learn the difficulty of the items in each category. It is very important for you to understand that in this test a good score does not necessarily mean a high number of items correct. It is quite possible to get a high score if a few correct responses are given to difficult items. Your best strategy should be to select items that you feel are challenging. Should you select easy items, then you will have a high number of items correct, but very likely your final score will be lower than you deserve."

Each testing session consisted of four phases. First, demographic data were collected. Next, the pretest anxiety subscale was administered, followed by the English test. Finally, the posttest anxiety subscale was administered.

Item-Selection Strategies

The three-item selection strategies proposed by Rocklin (1989) were considered, along with a new one. Each strategy is based on the relationship between the score on the previous item (correct or incorrect response) and the difficulty category selected. The proposed new strategy was termed "inflexible." In this strategy the examinee selects the same difficult category for almost every choice (at least 90%).

Data Analysis

At each of the 20-item positions, the relationship between ability and item difficulty was assessed, using both a Pearson correlation coefficient and the square root of the mean squared difference (RMSD) between the two variables. Item difficulty levels were used rather than difficulty parameters (though both indices basically provided the same results). To compute RMSD index, ability estimates were transformed into integers in the range from 1 to 7, based on the same classification scheme used to classify items into difficulty levels.

Mean item difficulty level choice for each item position was computed for the entire sample, and linear and nonlinear regression analyses were fitted to these pairs. Autoregressive components will be added to the regression equations because the independence of residuals cannot be taken for granted in this case.

In addition to these regression analyses, hierarchical multiple regression was applied to explore the contribution of several predictors to mean difficulty level choices made by examinees. Four blocks of predictors were used to predict the mean difficulty choice for item *i*:

- Block 1: gender, pretest anxiety, estimated ability.
- Block 2: item *i*-1 score, item *i*-1 difficulty level, item *i*-1 score by difficulty interaction.
- Block 3: item *i*-2 score, item *i*-2 difficulty level, item *i*-2 score by difficulty interaction.
- Block 4: item *i*-3 score, item *i*-3 difficulty level, item *i*-3 score by difficulty interaction.

The goal of these regression analyses was to identify which factors influenced the difficulty level choices made at various points during a testing session. Block-1 variables were introduced in the regression equation in the first place; block-2, block-3, and block-4 variables were consecutively introduced then, and the increase in variance explained for by each block was computed.

The procedure used by Rocklin (1989) was also applied to identify the strategies used by examinees. First, examinees whose level choices (at least 18 out of 20) were the same were classified as inflexible. For the remaining examinees, a theoretical 20-element vector was generated for the flexible, failure-tolerant, and failure-intolerant strategies. Each vector contained the expected difficulty levels that each examinee would have produced if he or she had strictly applied a particular strategy. Next, theoretical and actual observed choices were compared using the RMSD index. Each examinee was then classified into a strategy group, according to the strategy that best fit his or her pattern of difficulty choices. Finally, analyses of variance were performed to compare strategy groups on several variables, including: estimated ability, standard error of ability estimation, number of items passed, response time, mean difficulty level choice, mean item difficulty parameter, and anxiety (both pretest and posttest).

Results

Descriptive Statistics for the Sample

Means and standard deviations of the variables used in this study are shown in Table 1. Some general comments can be made regarding these results. First, the mean pretest anxiety slightly exceeded that found in our previous reports (1.08 in Olea, Ponsoda & Wise, 1995; .89 in Ponsoda et al., 1997) and in similar studies using the entire 20-item State Anxiety scale (.94 in Johnson et al., 1991; .94, .90 and .97 in Plake et al., 1995). The higher mean anxiety in the current study is likely due to the broader consequences associated with test performance. Some subjects reported higher anxiety levels, as 20% of the sample had pretest anxiety above 1.5 (the scale midpoint) and a 7% above 2 points. Mean pretest and posttest anxiety levels showed a significant decrease (see below), which suggests a positive effect of SATs on examinees' affective and motivational states.

Second, mean standard error of ability estimation was .27, which would correspond to a reliability coefficient of .92 (Thissen, 1990). This suggests that short SATs may provide higher precise scores. Most examinees (90%)

Table 1. Descriptive statistics for the sample of examinees (n = 167).

Dependent variable	Mean	SD
Estimated ability	.39	.65
Standard error of ability estimation	.27	.05
Pretest anxiety	1.14	.51
Posttest anxiety	1.05	.54
Correct responses (%)	61.53	11.66
Mean item difficulty	4.33	.64
Item difficulty parameter	.33	.63

had standard errors ranging from .22 to .32, a small percentage (8%) of the sample had standard errors from .33 to .49, and only four examinees (2%) produced standard errors above .5 (.72, .90, 1.53 and 1.68). One of these examinees was very conservative in the selection of difficulties, always selecting the medium difficulty level and passing about all the items. For the other three examinees, the situation was the opposite – they selected difficulties well above their abilities and passed a very low number of items. These four cases illustrate drawbacks in SATs. Despite verbal directions asking examinees not to do so, they chose items that were not well matched to their abilities. In these situations, ability estimates lack of precision and are less useful. These four cases were consequently excluded from the final sample (thus reducing its size to 167 examinees).

There were several additional noteworthy findings in Table 1. The mean percentage of items passed was 62%, which was appreciably below that found in previous studies (75% in Johnston et al., 1991; 69% in Ponsoda et al., 1997). Mean response time per item was 8.16 seconds, and the 15-second time limits for each item was seldom reached. No examinees expressed complaints regarding time pressure.

Match Between Ability and Selected Difficulty Levels

Table 2 contains Pearson correlations between examinees' final ability estimates and difficulty levels chosen, at each item position. The correlations, which ranged from .35 to .56 with a mean of .46, showed a tendency to increase during the testing session. The magnitude of the correlations, however, was generally lower than those reported by Johnson et al. (1991; .64 and .83 for items 1 and 11, respectively) and Wise, Plake, Johnson, and Roos (1992; .68 for item 15).

The minimum (best possible fit) and maximum possible values for the RMSD index are 0 and 6, respectively. The second row of Table 2 shows that the RMSD values ranged from .46 to .75 (with a mean of .59). Consistent with the correlation coefficients, RMSD values indicated a trend toward better fit as the testing session progressed.

Initial Difficulty and Trend Throughout the Testing Session

Table 3 shows distributions of item difficulty choices for items 1, 11, and 20. The first choice for most examinees was a medium difficulty level. Comparisons of the three distributions reveals a general shift toward more difficult items being chosen as the test progressed. These results are in agreement with those found by Johnson et al. (1991).

Figure 1 depicts the mean difficulty level choices made at each item position. The empirical curve indicates increases in mean difficulty that diminished in size as the session progressed. When linear and nonlinear regression equations were fitted to this empirical data, a relationship between serial item position and difficulty mean appeared. As Table 4 and Figure 1 show, both linear and nonlinear regressions provided an adequate description, but significantly better fit was provided by the linear + quadratic equation. The percentage of variance accounted for by the linear + quadratic model was 95% (versus 86% for the linear model).

Predictors of the Difficulty Level Choices

Table 5 contains results of the hierarchical regression analyses of the difficulty level choices at each item position. Three variables consistently appeared as predictors of difficulty choices: examinee ability, the difficulty lev-

Table 2. Match between ability and difficulty level choice, by item.

	Item position																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Correlation	.35	.37	.36	.39	.47	.42	.44	.46	.46	.45	.50	.50	.56	.52	.49	.52	.50	.45	.39	.50
RMSD[a]	.75	.71	.72	.69	.59	.66	.57	.56	.55	.54	.54	.55	.46	.52	.53	.53	.53	.62	.60	.54

[a] Square root of the mean squared differences between ability level and difficulty choice level.

Table 3. Distributions of difficulty choices (percentages) for items 1, 11, and 20.

	Difficulty level choices								
	1	2	3	4	5	6	7	Mean	SD
Item 1	1.2	1.8	25.7	49.1	18.6	3.6	–	3.93	.88
Item 11	–	1.8	9.6	49.7	31.1	4.8	3.0	4.37	.90
Item 20	–	–	9.6	46.7	29.9	12.0	1.8	4.50	.89

Table 4. Linear and nonlinear regression equations between difficulty choices and serial item position.

Model	R^2	p	b_0	b_1	b_2	b_3	AC[a]
Linear	.862	< .001	4.173	.026			.576
Linear + quadratic	.949	< .001	3.841	.080	–.003		–.650
Linear + quadratic + cubic	.967	< .001	3.741	.120	–.001	.000	–.650

[a] Autocorrelation

Table 5. Hierarchical and simultaneous regression analyses of difficulty level choices.

		Item Position																			
Block	Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Sex										–				–		–	–	–		
	Anxiety																				
	Ability	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	R^2	.13	.15	.17	.16	.28	.24	.26	.26	.31	.26	.32	.27	.40	.33	.33	.36	.36	.28	.24	.35
2	Difficulty (i-1)	X	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Score (i-1)	X	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Interaction (i-1)	X					+				+							+			
	R^2 increment	X	.46	.31	.34	.18	.33	.22	.26	.24	.22	.24	.32	.21	.26	.28	.23	.16	.31	.23	.34
3	Difficulty (i-2)	X	X		+			+	+			+	+	+	+	+	+	+	+	+	
	Score (i-2)	X	X		+							+			+	+	+				
	Interaction (i-2)	X	X																		
	R^2 increment	X			.10			.04	.04			.04	.03	.04	.06	.03	.04	.06	.05	.04	
4	Difficulty (i-3)	X	X	X									+					+			
	Score (i-3)	X	X	X														+			
	Interaction (i-3)																				
	R^2 increment	X											.02					.04			
	Total R^2	.13	.61	.48	.60	.46	.57	.52	.56	.55	.48	.60	.64	.65	.65	.64	.63	.62	.64	.51	.69
	Ability	.36		.15		.26	.14	.19	.18	.24	.18	.22	.11	.29		.13	.23	.30		.16	.22
	Difficulty	X	.74	.63	.72	.53	.62	.60	.64	.57	.58	.63	.73	.59	.69	.72	.64	.47	.79	.61	.70
	Score	X	.41	.36	.33	.19	.33	.22	.21	.29	.18	.21	.21	.25	.20	.18	.22	.27	.20	.20	.15
	Total R^2	.13	.61	.47	.50	.45	.53	.46	.52	.53	.43	.56	.60	.61	.58	.60	.57	.46	.58	.46	.68

Note. The upper part of the table shows the hierarchical regression results. A "+" and "–" stand for positive and negative significant regression coefficients, respectively. Blanks indicate nonsignificant regression coefficients or nonsignificant increments in R^2. An "X" signifies that the corresponding predictor could not be entered in that regression equation. The lower part of the table shows the results of the simultaneous regression in which the predictors were ability, difficulty of the previous item, and the score on the previous item.

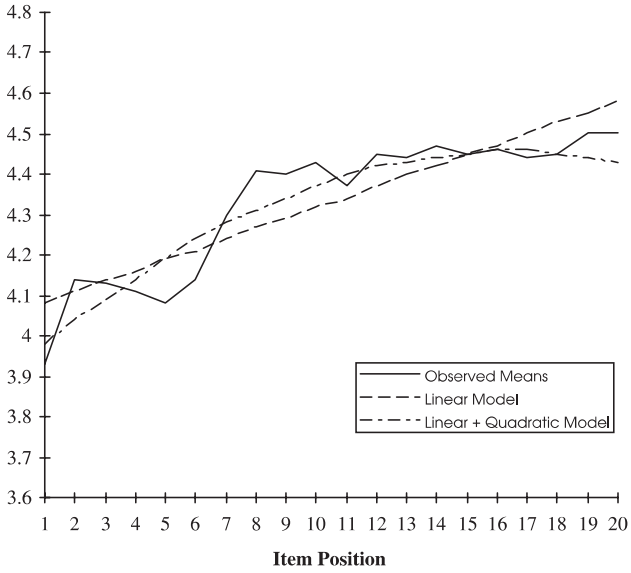

Figure 1. Actual mean item difficulty level choices and those predicted by linear and nonlinear regression models.

el of the previous item, and the score on the previous item. Their coefficients in regression equations were positive, indicating that (a) high ability students chose more difficult items, (b) an item difficulty choice tended to resemble the one of the previous item, and (c) examinees tended to choose a more difficult level after a correct response and a less difficult level following an incorrect response. The difficulty level and score on the penultimate item also contributed to prediction of many difficulty level choices, but their explained variance increases were modest (from 3% to 10%) and the statistical significance of their contributions was inconsistent across item position. Gender, pretest anxiety, and choices and score on the item preceding the penultimate item contributed little to prediction. Few significant interaction effects occurred at a rate consistent with that expected by chance; therefore, we interpreted these effects as probable Type I errors. The lower sector of Table 5 contains results from a second multiple regression analysis in which ability, previous item difficulty and previous item score were simultaneously introduced as the only predictors; these results were very similar to those from the hierarchical regressions. Percentages of variance accounted for by these three predictors ranged from 13% to 68%, with a mean percentage of 52%.

Difficulty Selection Strategies

Table 6 shows, across examinees and items, changes in difficulty levels following correct and incorrect answers to previous items. Examinees repeated the previous item difficulty level on 63.4% of choices. On 36.6% of choices a change occurred. Consistent with a flexible strategy, after a correct response a more difficult item was selected far more often than a less difficult item (29.4 versus 2.8%). The opposite was true after an incorrect response: 5.2% of the choices were more difficult while 38.5% were less difficult. These findings did not differ appreciably from those reported by Johnson et al. (1991).

The inflexible strategy (same choice made in 18 or more choices) was found to be applied by 22 examinees (13.2%). 72 (43.1% of the sample) out of the remaining 145 examinees, were classified as using the failure-tolerant strategy (no change after an error and a more difficult choice after a correct response). A total of 64 examinees (38.3%) applied the flexible strategy (a more difficult choice after a correct item and easier choice after an error). Only 2 examinees were classified as using the

Table 6. Percentage of difficulty level choices, by score on previous item.

	Score on previous item		
Difficulty choice	Correct	Incorrect	Total
More difficult	29.4	5.2	20.0
Same difficulty	67.8	56.3	63.4
Less difficult	2.8	38.5	16.6
Total	61.4	38.6	

Table 7. ANOVAs for test characteristics, anxiety and fit, by item selection strategy.

	Flexible (n = 64)		Failure tolerant (n = 72)		Inflexible (n = 22)			
	Mean	SD	Mean	SD	Mean	SD	F	p
Ability	.24	.59	.60	.65	.32	.63	5.04	.020
Standard error	.26	.04	.27	.05	.27	.04	.93	.400
Subjective difficulty	3.99	.40	4.74	.54	4.07	.64	41.05	<.001
Correct responses (%)	65.94	9.12	55.56	10.53	66.14	14.05	19.15	<.001
Response time	8.41	3.29	8.36	2.36	6.90	2.26	2.85	.060
Pretest anxiety	1.20	.54	1.03	.42	1.30	.62	3.49	.030
Posttest anxiety	1.04	.53	1.01	.51	1.05	.50	.09	.920
Model fit	.89	.30	.80	.26	.44	.36	19.80	<.001

Note. Strategies with seven or fewer examinees were not analyzed. Each F-ratio was based on 2 and 155 df.

failure-intolerant strategy (no change after a correct response and an easier item selected after an error). For the remaining 7 examinees, two strategies fit the choice patterns equally; these examinees were classified as undefined.

Analyses of variance for several outcome variables were performed among the flexible, failure-tolerant and inflexible subgroups, as they had numbers of subjects deemed sufficient for comparisons of means. Results of these analyses are shown in Table 7. Three strategy groups differed in regard to ability, and a Newman-Keuls test revealed that failure-tolerant examinees had a higher mean ability than their inflexible and flexible counterparts. These groups were also found to significantly differ on both difficulty level choices and percentage of correct responses. A Newman-Keuls test found that failure-tolerant examinees chose more difficult items and had fewer correct responses than their flexible and inflexible classmates. However, no significant differences between groups were found regarding either standard error of ability estimation or mean response time.

Pretest and posttest anxiety measures were submitted to both cross-sectional and longitudinal analyses. A one-factor ANOVA revealed that the three strategy groups differed on pretest anxiety, with the highest mean pretest anxiety reported by examinees using the inflexible strategy, followed by the flexible examinees and then the failure-tolerant examinees. The three subgroups did not differ on posttest anxiety. An additional (longitudinal) analysis was also applied to anxiety data. A two-factor ANOVA, with strategy group and time (pretest, posttest) as factors, revealed a nonsignificant main effect for strategy ($F(2,155) = 1.36$, $p = .26$), a significant effect for time ($F(1,155) = 17.25$, $p < .001$), as well as a significant interaction effect ($F(2, 155) = 3.11$, $p = .05$). Inspection of the means (see Table 7) indicated that anxiety was reduced in examinees using the flexible and inflexible strategies, but not for those using in the failure-tolerant strategy.

The fit between actual and predicted choices for each strategy differed among the three strategies ($F(2, 155) = 19.80$, $p < .001$). A Newman-Keuls test revealed that the fit index corresponding to the inflexible strategy was lower than the given one by the other two strategies, who did not differ from each other.

Discussion

The choice of item difficulty levels is the distinctive feature of SATs. Rocklin (1994) stated that ". . . very little research . . . has addressed the factors involved in examinees' selections of item difficulty" (p. 12). The goal of the current work was to provide a better understanding of examinees' difficulty choices in SATs and their relationships with psychological and psychometric variables. This study is distinctive in the sense that it represents the first SAT study in which scores on the SAT had meaningful consequences for the examinees. Scores on the self-adapted test had a moderate impact on course marks; 20% of the second-term grade depended on a examinee's score on the SAT, and examinees were compelled to take the test. It was presented as an additional exam and was as compulsory as any other course assessment.

Four examinees were dropped from the analyses because their standard errors were very high. They selected item difficulty levels which were either very easy or very difficulty for their ability levels. Practitioners should pay attention to this result when thinking about the type of test to apply; a small percentage of examinees receiving fixed-length SATs (2.3% in the current study) may not attain an acceptable level of precision for their ability estimates. If the test had used a fixed precision criterion as a stopping rule, it would have required a substantially longer test for these examinees. Developers of future SATs should pay attention to this issue. The problem was addressed by Wise, Kingsbury and Houser (1993), who proposed what they called "restricted" self-adapted testing (RSAT). In this type of SAT, an examinee is allowed a limited range of difficulty choices around his or her current ability estimate. Use of an RSAT would prevent an examinee from selecting item difficulties that are poorly matched to ability – thus providing examinees some control over item difficulty while ensuring that the items administered will yield an acceptably small standard error of ability estimation.

A positive correlation between ability and difficulty level choice was found; high-ability examinees tended to choose more difficult items and low-ability examinees tended to choose easier items. The relationship was less strong for the three first items, which may indicate that examinees used the initial items to learn the meaning of the difficulty levels through an exploration of different levels. Plake et al. (1995) found that the first difficulty selection depended on the prior information (if any) provided to the examinees about the actual difficulty of items assigned to each level. Our results suggest that not only the first item, but also the second and third was used by examinees to gain a better understanding of the difficulty levels.

Results from our regression analyses confirmed previous conclusions reached by Rocklin (1989) and Johnson et al. (1991). Items from intermediate difficulty levels were chosen for the first item, after which difficulty choices progressively increased up to the eighth item and

then did not appreciably change. A linear regression described the trend of the mean difficulty levels well, but a nonlinear (linear + quadratic) model provided better fit. It should be kept in mind that these results were obtained in 20-item tests. Other functions may result more appropriate for longer tests.

The strongest predictors of difficulty level choices were the examinee's ability, the difficulty level chosen for the previous item and the score on the previous item. Pretest anxiety was not found to be related to first item choice, as was reported by Johnson et al. (1991), but it was found to be related to the item selection strategies that the examinees applied.

In this study, 20% of second-term grade depended on the score obtained at the SAT test, which may explain the slightly higher pretest anxiety level found (as compared to previous American and Spanish studies) and the lower percentage of correct items. This consequential testing situation may have also affected the strategies that some examinees used. The flexible strategy was the most popular strategy found in previous studies (Rocklin, 1989; Johnson et al., 1991), but it was applied by only 38.3% of our sample. Most of the examinees in the current study (43.1%) preferred the failure-tolerant strategy, which indicates that they did not select items in a manner similar to a CAT algorithm. These examinees tended to choose a more difficult item after a correct response, but did not choose an easier category after an error. An inflexible strategy (two or fewer difficulty level changes during the test) was found to be used by some examinees (13.2%).

Examinees who applied different item difficulty strategies were found to differ on cognitive and stable characteristics (i. e., ability) as well as affective and situational (anxiety) characteristics. Comparisons of the strategy groups provided several interesting results.

First, examinees reporting low pretest anxiety tended to choose the difficulty selection strategy (failure-tolerant) that should lead to the most challenging tests. Those with high pretest anxiety tended to choose less risky strategies (flexible or inflexible). Examinees applying these latter two strategies appeared to be the most affected by the SAT format, as only in these cases was a significant reduction in anxiety observed. These results suggest that state-anxiety levels are related to the strategies applied. Under some strategies (those preferred by higher anxiety students) an anxiety reduction was witnessed. However, our data did not allow us to distinguish whether this reduction was due to the strategy applied, to the higher pretest anxiety level, or both.

Second, ability means (but not standard errors) differed among the three strategy groups. In the current study, the failure-tolerant subgroup had a higher mean ability than the other two groups, despite being the only group in which no significant change in anxiety was observed. Rocklin (1989) also found a significant difference in ability, but in his case the ability mean for the failure-tolerant subgroup was significantly lower that his other strategy groups. This would suggest that higher ability examinees tended to prefer the failure-tolerant strategy. However, recent results (Rocklin et al., 1995) may provide an alternative explanation. When item difficulties are not reasonably matched to examinee abilities, maximum likelihood estimates of ability may be biased. In our data, both the percentage of correct responses and the mean difficulty levels show that failure-tolerant examinees self-administered more difficult tests than the other two subgroups. Thus, a differential mistargetting between item difficulties and ability may also be related to the ability differences found among strategy groups. However, Rocklin et al.'s (1995) simulations also found that conditions showing bias showed higher standard errors. In the current study, the strategy groups did not significantly differ on precision – which renders unlikely ability estimation bias as an explanation for the ability main effect that we observed.

Previous comparisons between CATs and SATs provided a few divergent results (Rocklin, 1994). Results of the current study suggest that the effects of SATs are not uniform, as different strategy groups yielded different results – an issue that should be considered when comparing CATs with SATs. For instance, in some studies an anxiety decrease due to SAT was found (Olea et. al., 1995; Wise et al., 1992), but not in others (Ponsoda et al., 1997). If, as was found in the current study, the reduction in anxiety is related to the strategies used, then a differential use of strategies across studies may help explain these inconsistent findings.

Finally, at least two questions should receive further consideration in future investigations of difficulty selection strategies. First, the sampling distribution of the RMSD fit index distribution is not known, which hindered our assessment regarding whether a model or strategy provides an adequate description for data. Secondly, new strategies based on more elaborate empirical and theoretical criteria should be developed. As suggested by an anonymous reviewer, the use of multidimensional techniques (e. g., cluster and latent class analyses) may be proved useful for identifying new strategies and/or validating those already identified.

Acknowledgment

This research was partially supported by two DGICYT grants (PS94–0040 and PS95–0046).

References

Allan, D. (1992). *Oxford Placement Test 1.* Oxford: University Press.

Andrich, D. (1995). Review of the book *Computerized Adaptive Testing: A Primer*. *Psychometrika, 4,* 615–620.

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

Atkinson, J.W., & Litwin, G.H. (1970). Achievement motive and test anxiety conceived as motive to approach success and to avoid failure. *Journal of Abnormal and Social Psychology, 60,* 52–63.

Hontangas, P.M., Canela, A., & Agustín, C. (1996). *Medida de la ansiedad desde la teoría de la respuesta al ítem: estudio del Cuestionario de Ansiedad Estado-Rasgo* [Anxiety measure from TRI: a study of the state-trait anxiety inventory]. Paper presented at the I Congreso de la Sociedad Española para el Estudio de la Ansiedad y el Estrés, Benidorm.

Johnson, P.L., Roos, L.L., Wise, S.L., & Plake, B.S. (1991). Correlates of examinee item choice behavior in self-adapted testing. *Mid-Western Educational Researcher, 4,* 25–28.

Kolen, M.J. (1980). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25–44.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Olea, J., Ponsoda, V., Revuelta, J., & Belchí, J. (1996). Propiedades psicométricas de un test adaptativo informatizado de vocabulario inglés [Psychometric characteristics of a computerized adaptive test for the measurement of English vocabulary]. *Estudios de Psicología, 55,* 61–73.

Olea, J., Ponsoda, V., & Wise S.L. (1995, April). *Tests adaptativos y autoadaptados informatizados: Efectos en la ansiedad y en la precisión de las estimaciones* [SATs and CATs: Effects of anxiety on estimate precision]. Paper presented at the IV Symposium de Metodología de las Ciencias del Comportamiento, Murcia.

Plake, B.S., Wise, S.L., & Roos, L.L. (1995). Effects of informed item selection on test performance and anxiety for examinees administered a self-adapted test. *Educational and Psychological Measurement, 55,* 736–742.

Ponsoda, V., Olea, J., & Revuelta, J. (1994). ADTEST: A computer-adaptive test based on the maximum information principle. *Educational and Psychological Measurement, 54,* 680–686.

Ponsoda, V., Wise, S.L., Olea, J., & Revuelta, J. (1997). An investigation of self-adapted testing in a Spanish high school population. *Educational and Psychological Measurement, 57,* 210–221.

Rocklin, T.R. (1989). *Individual differences in item selection in self-adapted testing.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Rocklin, T.R. (1994). Self-adapted testing. *Applied Measurement in Education, 7,* 3–14.

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, A.M., & Holst, P.M. (1995). Effects and underlying mechanisms of self-adapted testing. *Journal of Educational Psychology, 87,* 103–116.

Rocklin, T.R., & Thompson, J.M. (1985). Interactive effects of test anxiety, test difficulty and feedback. *Journal of Educational Psychology, 77,* 368–372.

Rocklin, T.R., Vispoel, W., & Wang, T. (1995, April). *Can examinees manipulate their ability estimates in self-adapted testing? A simulation study.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Spielberger, C.D., Gorsuch, R.L., & Lushene, R.E. (1970). *Manual for the state-trait anxiety inventory.* Palo Alto, CA: Consulting Psychologist's Press.

TEA (1988). STAI. *Cuestionario de Ansiedad Estado-Rasgo* (3rd ed.). Madrid: TEA Ediciones SA.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161–186). Hillsdale, NJ: Erlbaum.

Vispoel, W.P., Rocklin, T.R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive and self-adapted testing. *Applied Measurement in Education, 7,* 53–79.

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

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

Wise, S.L, Kingsbury, G.G., & Houser, R.L. (1993, April). *An investigation of restricted self-adapted testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

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

Vicente Ponsoda
Facultad de Psicología
Universidad Autónoma de Madrid
Canto Blanco
E-28049 Madrid
Spain
E-mail Vicente.Ponsoda@uam.es
Tel. +34 91 397-5203
Fax +34 91 397-5215