DOCUMENT RESUME

ED 358 155 TM 019 954

AUTHOR Schaeffer, Evonne L.
TITLE Toward an Understanding of Context Effects: Test-Taker Processes and Test Situation Demands.
PUB DATE Apr 93
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cognitive Processes; *Context Effect; Difficulty Level; High Schools; *High School Students; Item Response Theory; *Pacing; Reading Tests; Resource Allocation; Test Construction; Test Items; *Test Wiseness; Timed Tests
IDENTIFIERS Self Monitoring; *Testing Effects

ABSTRACT

Context effects in test taking were explored, paying attention to the psychological processes that occur during test taking, and modeling context effects for each individual at the item block level. A sample of 279 high school students (140 females and 139 males) was chosen to yield adequate power for detecting interactions. Reading test forms were developed, and a measure of the various test taking processes thought necessary for test performance was developed. Students were administered the reading test in unpaced (1 total time) or paced (each passage separately timed for 3.5 minutes) conditions, with paced conditions considered to demand less student responsibility for self-monitoring. The overall relationships of performance consistency and the cognition monitoring composites suggest relationships that may not occur in a typical testing situation, and that monitoring ability is positively related to performance consistency. Context effects, viewed as performance inconsistency, are perhaps an indication that the examinee is not able to meet the monitoring demands of the task. By imposing a resource allocation structure, pacing appeared to benefit those who were not effective resource allocators. To fully understand context effects, it is important to take into account the demands of the situation and test takers' abilities to meet those demands. Six graphs and three tables present study data. (SLD)

Toward an Understanding of Context Effects: Test-Taker Processes and Test Situation Demands

Evonne L. Schaeffer
Pacific Graduate School of Psychology
April 1993

Paper presented at the annual meeting of the American Educational Research Association
Atlanta, Georgia

Both prior learning and the psychological processes that operate during testing affect test performance (e.g., Embretson, 1985). One view of test performance is that it represents a "sampling of mental organizations, not just bits and pieces, as well as their dynamic assembly or reassembly to meet task demands. The sampling and assembly operations shift during tests and learning tasks as a function of item variations" (Snow & Lohman, 1989, p. 317). Item variations result from different contents and difficulties, as well as from intra-test context factors such as item order and item format. To the extent that test performance is influenced by these incidental context effects, the test's construct validity may be compromised (Messick, 1989).

Understanding why context effects occur and controlling for their possible effects is important for both classical and modern measurement applications. For example, comparing total scores across examinees taking the same items in differing orders is fair only when context has no effect on performance (or, at least, has a similar effect across examinees). When context effects operate, the local independence assumption in item response theory (IRT) has not been strictly met, perhaps weakening the validity of inferences based on test scores.

Under the IRT model assumptions, each item "stands alone" in the sense of having unchanging characteristics regardless of where it may appear on the test. Most IRT applications, including customized testing, adaptive testing, and matrix sampling, fundamentally depend on this local independence assumption. Although items are individually characterized, they rarely reside in isolation. Thus we must ask whether this theoretical assumption is realistic: Does an item present the same task regardless of its relation to other items?

This context effect question has been recast from a measurement perspective: Do item parameters remain invariant regardless of the context in which the item appears? Whitely and Dawis (1976), Yen (1980), Kingston and Dorans (1984), and Wise (1986), among others, have compared performance on items when they appeared in different

locations and provided some evidence that item parameters can depend on the location of the item in the test. The extent and type of dependencies differed for various test contents and/or samples.

Comparing IRT person ability or item parameter estimates across different intra-test contexts is one way to describe context effects, but several other approaches have also been considered. Leary and Dorans (1985) reviewed and categorized much of the context effects literature as to whether the studies examined the main effect of intra-test context on performance or the aptitude-treatment interaction (ATI) of intra-test contexts with some examinee characteristic. They found few consistencies.

One wave of context effects research sought main effects on total test scores associated with different item orderings. The only consistent result was that for relatively speeded tests, the easy-to-hard item ordering elicited better overall performance than the hard-to-easy ordering. This trend was not observed when tests were administered under nonspeeded conditions. Test speededness apparently mediated the effects of item order.

Another wave of context effects research examined interactions of intra-test context with examinee characteristics. The rationale here is that test-takers vary in their susceptibility to changes in item context. Part of the challenge with these ATI models is to identify a dimension of examinee variability that systematically predicts intra-test context susceptibility. In the work reviewed by Leary and Dorans, no examinee characteristic was found that consistently interacted with intra-test context to affect performance. Anxiety was examined most frequently, without much empirical success. Whether examinee ability or gender interact with changes in context also is not clear.

Shortcomings common to many of these studies may account for the lack of progress in our understanding. For example, total test score was often used as the dependent measure and this level of analysis may obscure trends at the item level. In at least two studies (Kingston & Dorans, 1984; Newman, Kundert, Lane, & Bull, 1988), no context effects were apparent at the level of total test scores, but context trends were evident

on scores at the item or item-type level. For example, Kingston and Dorans found that reading comprehension items were subject to fatigue effects (i.e., an item was more difficult when it appeared later, compared to earlier) whereas antonym items were subject to practice effects (i.e., an item was easier when it appeared later, compared to earlier). These contrasting influences were masked at the level of overall verbal ability score, which appeared not to change as a result of changes in item order.

Another shortcoming, especially for the ATI studies, was lack of power. ATI studies are different from traditional experiments in that the sample size necessary to detect an interaction is larger than that needed to detect main effects. A study with inadequate sample size lacks the power to detect an interaction, even if a strong one is present in the population (Cronbach & Snow, 1981). For example, Hambleton and Traub (1974) tested the hypothesis that different item orders (easy-to-hard and hard-to-easy) had different effects for those high and low on a debilitating test anxiety scale. They reasoned that the hard-to-easy item ordering would be most stressful for highly test anxious examinees. They predicted that the performance of high and low test anxious students would differ more under the hard-to-easy item ordering compared to the easy-to-hard item ordering. Using an extreme-group blocked analysis of variance design, the study included approximately 25 students in the upper and lower quartiles of the anxiety scale. No interaction between anxiety and item order was found. "The data of this study provide no evidence to support the hypothesis that the difference in performance between high and low test anxious subjects would in general be greater on the difficult-to-easy order than the reverse order" (p. 45). It is important to keep in mind that the absence of a significant interaction in low power studies is not evidence that the interactions do not exist. Cronbach and Snow (1981, p. 60) suggest for such extreme-group designs (under some guiding assumptions) when the top and bottom quartiles are compared, the sample size needed in each quartile to reach a power of .80 in detecting an interaction is 59. This is more than twice that used in the Hambleton and Traub study.

The failure to identify examinee characteristics that interact with changes in item context may also contribute to the lack of consistencies in the ATI studies. Noticeably missing from previous ATI research was an analysis of examinee processing during testing. Identifying salient examinee characteristics that actually operate during test taking seems fundamental to the successful explanation of why performance may differ as a function of variations in intra-test context. For example, it seems reasonable to expect that anxiety level may affect performance; however, if anxiety was not actually operating during the test administration (e.g., the students knew that the test results would be used only for research purposes and had no real reason to be concerned about their scores), the two variables would probably not correlate.

In an attempt to reconcile some of the inconsistencies of previous research, this study recognized that some test-takers may be more susceptible than others to item context and that some testing situations may be more likely than others to evoke context effects. This study follows Snow and Lohman's (1989) view that test performance is the *match* of the test-taker's cognitive, metacognitive, and volitional abilities to the demands of the situation.

Test situation demands. Several previous studies have found that reading comprehension items tend to elicit item-order effects more so than do other contents (e.g., Yen, 1980; Kingston & Dorans, 1984; and Wise, 1986). Measures of reading comprehension have features that set them apart from other test content areas and it is likely that these features create a greater need for self-regulation (e.g., mindfulness, comprehension monitoring, etc.) for successful performance. For example, reading tests, by nature, cannot be content-free. Whether or not the material is familiar, active discipline is needed to maintain attention to the task (e.g., Farr, Pritchard, & Smitten, 1990). If reading tests tend to require more cognition monitoring during performance and those same measures tend to exhibit the most context effects, then it is proposed that *context effects are more likely to occur in situations that require more cognition monitoring.* Comparisons between aptitude and achievement measures support the proposition that test contents that

demand more active cognitive and metacognitive processing (e.g., aptitude tests or tests of fluid ability) appear to be most susceptible to item location effects (Leary & Dorans, 1985). But, just as some situations make more metacognitive demands, some individuals are more able than others to meet those demands.

Test-taker processes. An examinee characteristic thought to mediate the effect of item order is the ability to monitor one's thoughts and progress during test taking. As referred to here, *cognition monitoring* involves executive, metacognitive processes such as keeping track of progress in light of time remaining and adjusting test taking strategies accordingly, as well as maintaining attention and motivation. These metacognitive strategies are not different in kind from those studied by Kuhl (1986), Brown (1980), Forrest-Pressley and Waller (1984), and Jacobs and Paris (1987) among others. Cognition monitoring involves being sensitive to changing situation demands and being flexible in matching one's resources to those demands. At the same time, cognition monitoring involves conserving resources in anticipation of future demands. Cognition monitoring refers to the use one makes of the available resources, where resources in the testing situation include time, motivation, persistence, and concentration. Able monitors allocate their resources optimally. They are able to meet the attentional and strategic demands of the task and are thus free to perform to the best of their ability. Those less skilled at cognition monitoring are not able to make the most of their resources. Less able monitors may not adequately meet the attentional and strategic demands of the task; those demands may thus interfere with performance.

Test-taker processes and test situation demands. By considering testing demands and test-taker processes simultaneously, this study attempts to describe the effect of intra-test context and to predict when and for whom item context has its greatest influence. This study compared two reading comprehension test administration conditions (*unpaced* and *paced*) that, due to their different metacognitive demands, were thought to be more and less likely to elicit order effects. The aptitude-treatment interaction methodology (ATI;

Cronbach & Snow, 1981) was used to model context effects as a function of the interaction between situational demands and test-taker cognition monitoring abilities.

Insert Figure 1 about here

Figure 1 displays the research hypothesis under study. Specifically, it was predicted that examinees who are able cognition monitors would show minimal context effects regardless of situational demands (*unpaced or paced*), whereas examinees who are not able cognition monitors would produce more context effects, particularly in situations that make greater self-regulatory demands (*unpaced*).

The current study tried to incorporate the lessons learned from the previous inquiries: Attention was paid to the psychological processes that occur during test taking; context effects were modeled for each individual at the item block level; and the sample size yielded adequate power for detecting interactions.

Method

The first step of the research was to develop reading test forms. One way to study the effects of reading block position is to construct the test with reading blocks (a *block* is a passage with its associated multiple-choice questions) that are functionally equivalent. That way, if contexts effects are not operating, one would expect consistent performance across the equivalent blocks. Any within-person variation in performance these blocks, then, can be attributed to the fact that the blocks differ (almost) only in their positions.

Of the reading blocks in the Nelson-Denny Reading Tests Forms C, D, E, and F (Brown, Nelson, & Denny, 1973; Brown, Bennett, & Hanna, 1981), seven were identified to be reasonably similar on the following characteristics: length, genre, and vocabulary level of the reading passage; level of processing required to answer the questions; and, based on pilot data, empirical block difficulty; time to complete the block,

and rated levels of interest. Four reading test forms were then constructed, each containing the seven similar reading blocks, but in different orders. Table 1 shows the order of reading blocks for each of the four reading test booklets.

Insert Table 1 about here

The next step was to develop a measure of the various test taking processes thought necessary for successful test performance. One hundred and fifty high school students participated in several cycles of the development, piloting, and revising of the preliminary versions of the cognition monitoring assessment. Because cognition monitoring is a complex construct, multiple measures, methods (i.e., self-report and performance), and response formats (e.g., Likert, true/false, and fill-in) were included. The final Cognition Monitoring Battery (CMB) included exercises adapted from Ferrell (1972) on testwiseness, Kuhl (1985) on action versus state-orientation, and from Sarason (1980) and Sarason and Sarason (1987) on test anxiety. The CMB contained six sections which produced a total of 12 variables. (A more detailed description of the development of the CMB can be found in Schaeffer, 1991).

Data were collected over two days in each class during May and June of 1990. Over 300 high school students participated by completing the Reading Comprehension Test on the first day of data collection and the Cognition Monitoring Battery on the second. The administration order of these two measures was the same across classes because it was thought that the CMB exercises may enhance one's awareness of test taking skills, possibly influencing subsequent test taking behavior, and also because taking the reading test first provided all of the students with a recent testing example from which to draw when responding to those CMB questions that related to the testing experience.

On the first day of data collection, students in each of the classes were administered the reading test in one of two conditions. Students in the *unpaced* classes were

administered the reading test in the conventional way, under one total time limit (24 1/2 minutes). Students in the *paced* classes were administered the same forms of the reading test (slightly different cover pages), with each passage separately timed for three-and-a-half minutes. Time allocations were determined from the pilot data and from those used with the Nelson-Denny Reading Tests. Although total testing time was the same in both conditions, the unpaced condition was considered more demanding because the students needed to regulate their own progress. It was thought that the paced administration relieved the students of that responsibility. The four reading test forms were randomly assigned within each class by spiralling. On the second day of data collection in each class, the CMB was administered. At the start of both days of data collection, the students were encouraged by the researcher and by their teachers to approach the exercises "as if they counted towards their grades."

Results

These analyses were based on a working sample of 279 students with complete data records. The sample consisted of 140 females and 139 males, with an average age of 15.14. Most students were in grades 9 (N=195) or 10 (N=74). Their reported ethnicities (45% white, 29% Hispanic, 10% Black, and 9% Asian or Pacific Islander) were representative of the diverse community in which the data were collected.

Evidence of convergent and discriminant validity was provided by the interpretable patterns of intercorrelations among the CMB variables, total reading score, and a measure of general ability. The 12 CMB variables were combined based on theoretical and empirical grounds to form two composites. The RESOURCE composite refers to the effective use of available resources during testing and included exercises that measured the appropriate use of one's time and the use of incidental clues to obtain correct answers. The CONTROL composite refers to the ability to maintain attention and to control the direction of one's thoughts, and included exercises that measured selective attention, frequency of

distracting task-irrelevant thoughts, and persistence. The CMB composite reliabilities, though not strong (lower bound estimates: RESOURCE = .37 and CONTROL = .53), appear to be adequate for these purposes.

Students in the paced (N=149) and unpaced (N = 130) conditions were compared on several background variables (e.g., age, grade, a measure of general ability), as well as on the CMB composites of CONTROL and RESOURCE. No significant differences were found, suggesting that any observed differences in performance are likely due to the different test administration conditions and not to any preexisting differences. Likewise, no significant differences on these background variable were found among students taking different reading comprehension booklets.

Each student received a score from 0 - 4 on each reading block indicating the number of correct responses. The average number of correct items per block ranged from 2 to 2.5, suggesting the empirical similarity of these reading blocks. Position difficulties were calculated as the average number of correct responses across blocks appearing in each position. Figure 2 shows that while the paced and unpaced groups performed similarly on blocks appearing early in the booklet, the paced group showed better performance for blocks appearing later.

Insert Figure 2 about here

Several individual blocks also displayed the pattern that their block difficulties changed as a function of their positions in the booklet more so for the unpaced group than for the paced group. Figure 3 shows performance on reading block F5, for example, when it appeared in booklet positions 2, 3, 5, and 6 for the unpaced and paced groups. The decline in performance as the block appeared later in the booklet seems more pronounced for the unpaced group.

Insert Figure 3 about here

Because the reading blocks were relatively similar to one another it was thought that if block order had no effect, then a person's performance should be consistent across blocks. Inconsistency in an individual's performance across blocks, therefore, provides an indication of the effect of block order. This was estimated by each individual's standard deviation across the reading blocks. However, because students in the paced condition were experiencing a novel test administration which would become more familiar with practice, the within-person standard deviation of reading block scores (RCSD6) was based on the last six blocks.

A higher standard deviation indicates greater inconsistency and therefore more within-person context effects. Student # 720, for example, showed rather inconsistent performance across the reading blocks with block scores of 0, 4, 4, 2, 3, and 2. In contrast, Student # 586 showed consistent performance, obtaining scores of 2, 3, 2, 2, 2, and 2. Their within-person standard deviations of 1.38 and 0.37, respectively, help to describe their different performance consistencies.

The influence of pacing on context effects was first examined by comparing these within-person consistencies for paced and unpaced students. Students in the paced condition performed significantly more consistently ($t = 2.63$, $p < .01$). Overall, pacing seemed to lessen the effects of block order. To see whether pacing increased performance consistency equally for students who differed on their cognition monitoring abilities, interaction analyses were performed.

A central hypothesis of this study was that pacing, by lessening the task demands, should benefit poor cognition monitors more so than able ones. This benefit would surface as more consistent performance across the equivalent blocks, indicating that the effects of block order on performance are minimized. The following analyses addressed whether the data supported this aptitude (i.e., cognition monitoring) by treatment (i.e., paced vs.

unpaced) interaction. The fundamental question is whether pacing altered the relationship between cognition monitoring and performance consistency.

One way to test whether pacing influenced the relationship between performance consistency and cognition monitoring is to compare the within-group regressions for the unpaced and paced groups. The ATI model may be expressed as:

$$RCSD6 = \alpha + \beta_1 CM + \beta_2 RCGROUP + \beta_3 (CM*RCGROUP),$$

where RCSD6 is the measure of within-person consistency, CM refers to cognition monitoring ability, and RCGROUP represents a dummy variable indicating paced or unpaced administration conditions. The null hypothesis to test the equivalence of the within-group regressions was $H_o: \beta_2 = \beta_3 = 0$. If both β_2 and β_3 are equal to zero, then the within-group regression equations coincide, indicating that pacing does not affect the relationship between performance consistency and cognition monitoring. On the other hand, if one of those two parameters is not equal to zero then the lines do not coincide, indicating that pacing does have an effect. Specifically, if $\beta_2 \neq 0$ and $\beta_3 = 0$ then the within-group lines would be parallel, indicating that pacing had a constant effect across all levels of cognition monitoring; whereas if $\beta_3 \neq 0$ the within-group lines would not be parallel, indicating that the effects of pacing differed for different levels of cognition monitoring.

To test the null hypothesis that pacing had no effect, full and reduced regression models were fit and compared separately for the RESOURCE and CONTROL CMB composite variables, where the full model includes β_2 and β_3 and the reduced model does not (Chatterjee & Price, 1977). For each composite, the full regression model accounted for significantly more variance than its corresponding reduced model, providing evidence that pacing influenced the aptitude-outcome relationships.

Table 2 summarizes the regression results for the RESOURCE composite. Each of the parameters in the full model was significant, resulting in the disordinal interaction shown in Figure 4. The relationship between effective resource allocation and performance

consistency was steeper in the unpaced group. Evidently the pacing intervention compensated for those students who were not able resource allocaters. More able resource allocaters, however, tended to perform more consistently on their own; perhaps the pacing interfered with their naturally effective test taking rhythms.

Insert Table 2 and Figure 4 about here

Table 3 shows the regression results for the CONTROL composite. The full regression model accounted for significantly more variability than the corresponding reduced model. The within-group regressions are displayed in Figure 5. Because the interaction term in the full model is not significant, the two lines are approximately parallel, suggesting the similar effect of pacing across all levels of CONTROL. Pacing appeared to contribute to more consistent performance regardless of one's ability to control the direction of his or her thoughts.

Insert Table 3 and Figure 5 about here

There was an interesting trend in one of the CMB section variables (*selective attention*) that contributes to the CONTROL composite. Figure 6 suggests that selective attention may mediate the effectiveness of the paced intervention. That is, those who are able to selectively attend showed the most increase in consistency due to pacing, while those less able to selectively attend showed no increase in consistency. This raises the concern that perhaps the novel paced instructions may have distracted those students who were not adept at controlling their attention.

Insert Figure 6 about here

Although the RESOURCE and CONTROL ATI full models were significantly better than their corresponding reduced models, indicating the effect of pacing, the model R^2s were quite low (.06), suggesting that a great deal of RCSD6 variation was not

accounted for by the current models. Several sources of instability may contribute to these low R^2s, including the unreliability of the CMB composites and of the criterion measure (RCSD6) itself.

Discussion

The overall relationships of performance consistency and the cognition monitoring composites in the unpaced group may suggest the relationships that are likely to occur in the typical testing situation: monitoring ability is positively related to performance consistency. Context effects, as viewed in this study as performance inconsistency, are perhaps an indication that the examinee is not able to meet the monitoring demands of the task. By imposing a resource allocation structure, pacing appeared to benefit those who were not effective resource allocaters, presumably by lessening the monitoring demands of the task.

Two unanticipated patterns emerged. The performance of those who were effective resource allocaters was more consistent in the unpaced condition, suggesting that the pacing may have interfered with their naturally effective rhythms. Also, some results suggest that students who were not adept at directing their attention may have been distracted by the paced instructions. While pacing was meant to reduce the task requirement of monitoring one's time during performance, it may have actually increased the task requirements (e.g., performing under novel conditions). Thus pacing may have added to the task burden, particularly for those students for whom it was most intended to help.

The ideal intervention would lessen the incidental task demands during test taking, not only to diminish order effects, but, more importantly, to cleanse the resulting individual differences of variability due to unintended factors (e.g., cognition monitoring). Providing practice with individually timed reading blocks would help to reduce the novelty of the paced intervention, and thus its possible distraction, and, at the same time, may encourage

better self-regulation. Another unintended effect of pacing was an increase in inconsistency for students who were good resource allocaters, perhaps because the pacing interfered with their naturally effective rhythms. An intervention that might overcome this difficulty would be one that controls for the maximum time per block. This would allow students to progress more quickly if they desired.

That some of the variation due to reading block position has been shown to relate to the demands of the test administration and to individual differences in monitoring ability may help to explain why there has been lack of agreement across context effects studies. It appears as though we need to take into account both the demands of the situation, and the test-takers' abilities to meet those demands, if we are to fully understand the implications on test performance that context effects may have.

References

Brown, A. L. (1980). Metacognitive development and reading. In R. J. Spiro, B. C. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, J. I., Nelson, M. J., & Denny, E. C. (1973). *The Nelson-Denny Reading Test, Forms C and D*. Boston: Houghton Mifflin Company.

Brown, J. I., Bennett, J. M., & Hanna, G. (1981). *The Nelson-Denny Reading Test, Forms E and F*. Chicago: The Riverside Publishing Company.

Chatterjee, S., & Price, B. (1977). *Regression analysis by example*. New York: John Wiley & Sons.

Cronbach, L. J., & Snow, R. E. (1981). *Aptitudes and instructional methods*. New York: Irvington Publishers, Inc.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design--Developments in psychology and psychometrics*. San Diego, CA: Academic Press.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27* (3), 209-226.

Ferrell, G. M. (1972). *The relationship of scores on a measure ot test-wiseness to performance on teacher-made objective achievement examinations and on standardized ability and achievement tests, to grade-goint average, and to sex for each of five high school samples*. Unpublished doctoral dissertation. University of Southern California.

Forrest-Pressley, B., & Waller, T. (1984). *Metacognition, cognition, and reading*. New York: Springer-Verlag.

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22* (3 & 4), 255-278.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education, 43* (1), 40-46.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8* (2), 147-154.

Kuhl, J. (1986). Motivation and information processing. A new look at decision making, dynamic change, and action control. In R. M. Sorrentino & E.T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior*. New York: Guilford Press.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55* (3), 387-413.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (Third edition)*. New York: Macmillan Publishing Company.

Newman, D. L., Kundert, D. K., Lane, D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education, 1* (1), 89-97.

Sarason, I. G. (1980). Introduction to the study of test anxiety. In I. G. Sarason (Ed.), *Test anxiety: Theory, research, and applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sarason, I. G., & Sarason, B. R. (1987). Cognitive interference as a component of anxiety: Measurement of its state and trait aspects In R. Schwarzer, J. M. Van der Ploeg, & C. Spielverger (Eds.), *Advances in test anxiety research, Volume 5*. Berwyn: Swets North America Inc.

Schaeffer, E. L. (1991). Understanding context effects: Test-taker processes and test situation demands. Doctoral dissertation, Stanford University, Stanford, CA. (University Microfilms No. 92-06, 849).

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement (Third edition)*. New York: Macmillan Publishing.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36*, 329-337.

Wise, L. L. (1986, April). Latent trait models for partially speeded tests. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17* (4), 297-311.

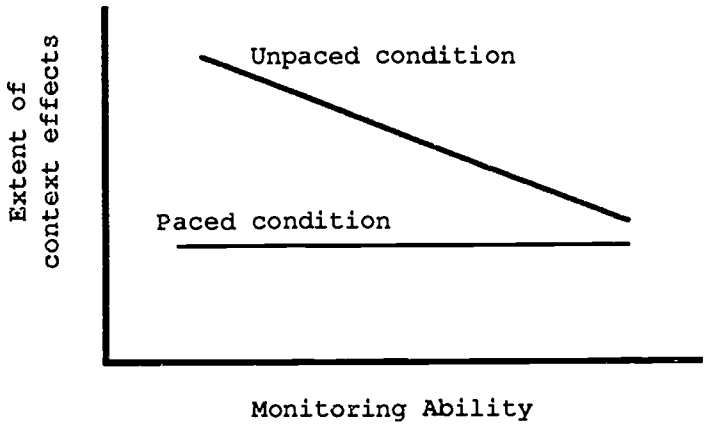

Figure 1. Research hypothesis indicating that less able cognition monitors, in general, will be more susceptible to context effects, and that pacing can help to reduce that susceptibility.

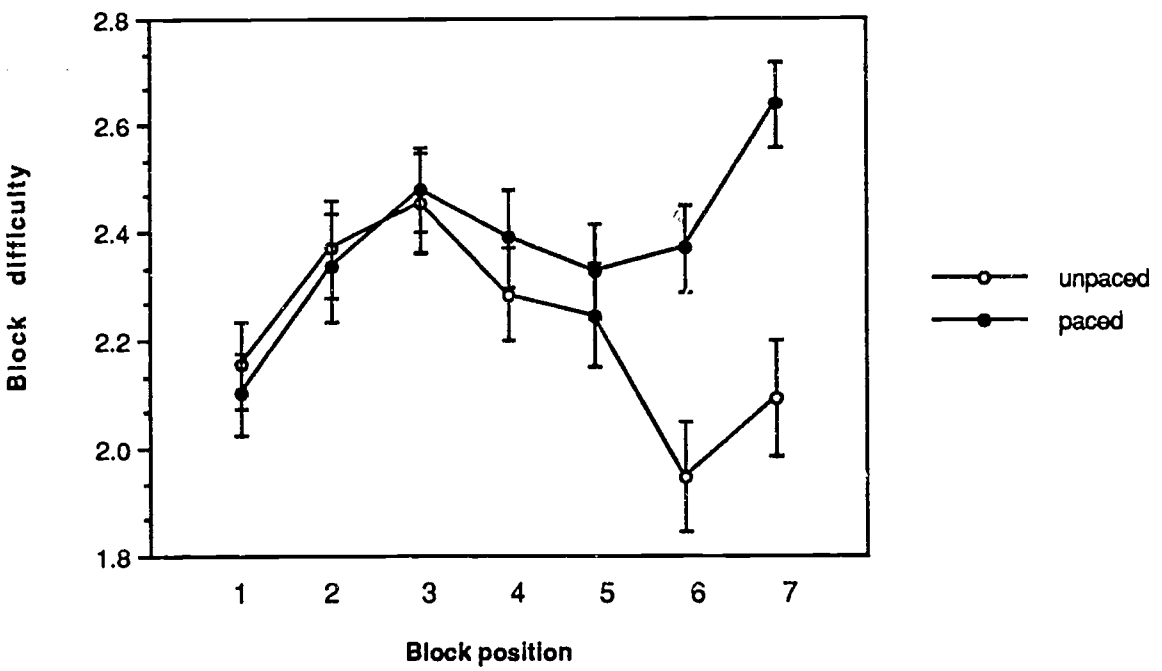

Figure 2. Unpaced and paced block difficulties, by position (plus or minus one standard error). The two administration conditions performed similarly at the beginning of the tests, but the paced group had a marked advantage toward the end of the tests.

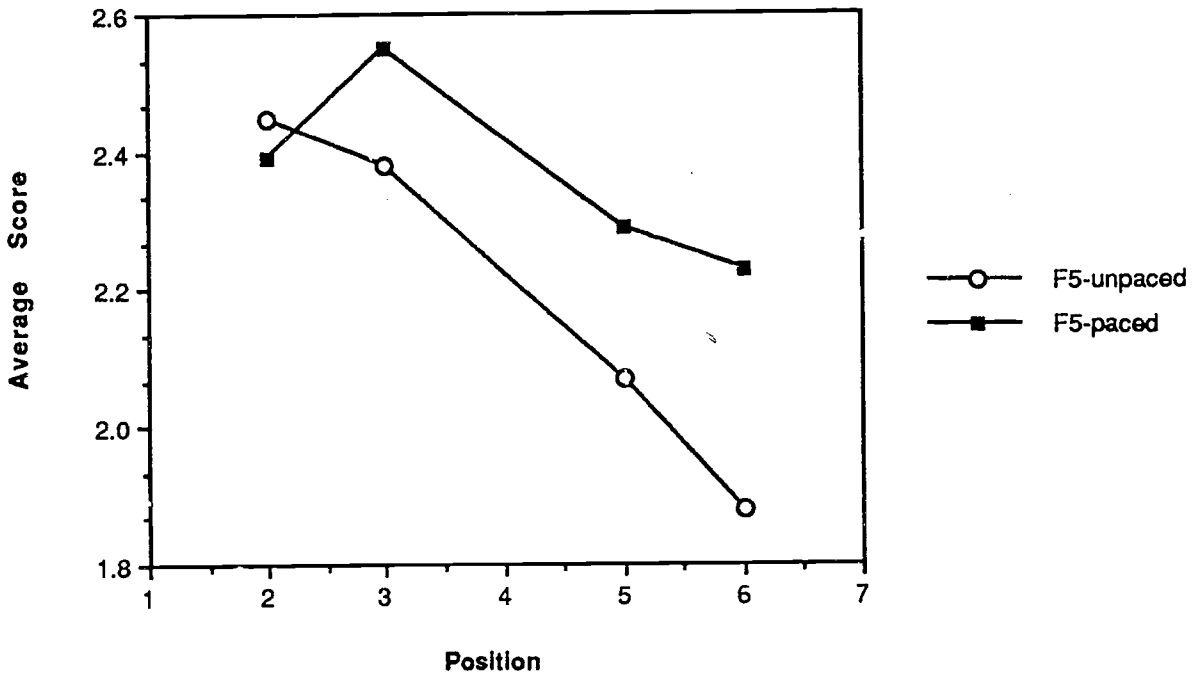

Figure 3. Reading block F5 average difficulty as a function of its block position across booklets and administration condition.

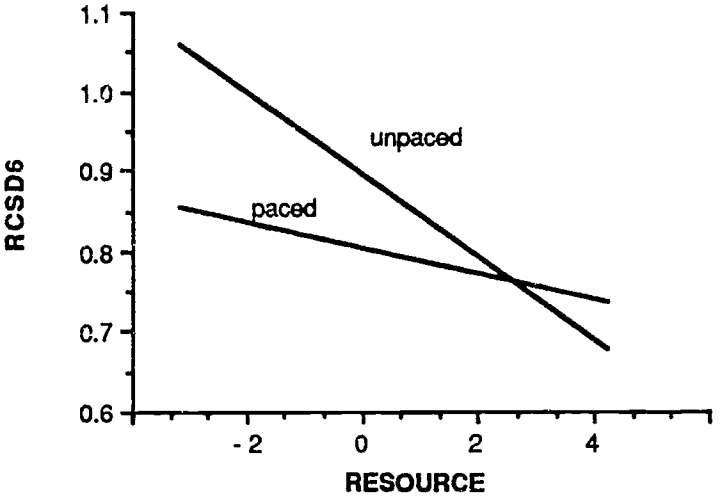

Figure 4. Fitted within-group regression lines for the relationship between performance consistency (RCSD6) and the ability to wisely allocate resources (RESOURCE).

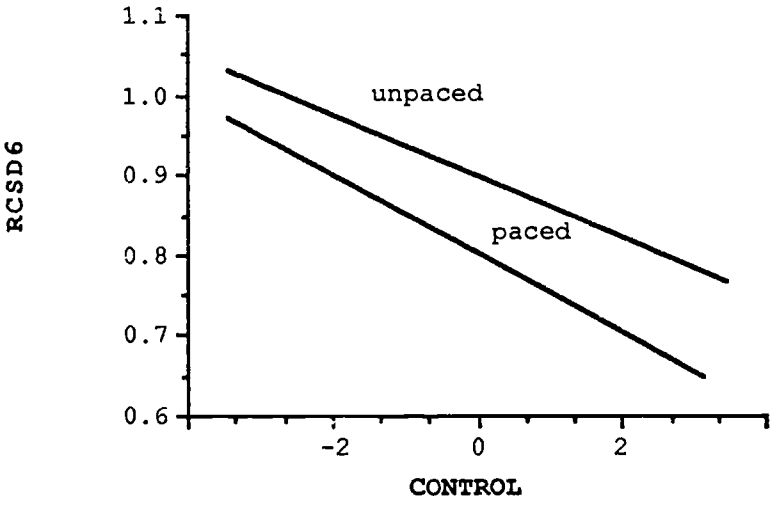

Figure 5. Fitted within-group regression lines for the relationship between performance consistency (RCSD6) and the ability to control the direction of one's thoughts (CONTROL).

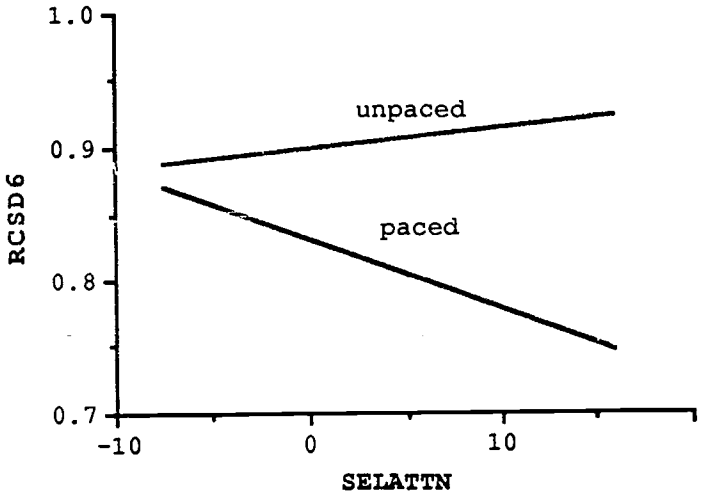

Figure 6. Fitted within-group regression lines for the relationship between performance consistency (RCSD6) and selective attention (SELATTN).

Table 1
Reading Comprehension Booklet Configurations

Test Booklet	Passage Position 1	2	3	4	5	6	7
I	E6	C7	E5	F2	F4	F5	E4
II	E6	F5	F4	C7	F2	E5	E4
III	E6	F2	F5	F4	E5	C7	E4
IV	E6	F4	C7	E5	F5	F2	E4

Table 2
Regression of RCSD6 on RESOURCE: Full versus Reduced Models

Variable	Coef.	s.e.	t	SSE	R^2
Full:				18.27	.06
INTERCEPT	.89	.02	39.39*		
RESOURCE	-.07	.02	-3.09*		
RCGROUP	-.09	.03	-2.75*		
RESOURCE*RCGROUP	.06	.03	2.03*		
Reduced:				19.06	.02
INTERCEPT	.85	.02	53.74*		
RESOURCE	-.03	.01	-2.43*		

Comparing full and reduced models:
F = 5.90*
df (2, 273)

Note: F-statistic to compare full and reduced models from Chatterjee & Price (1977), p. 88.

*Prob. < .05.

Table 3
Regression of RCSD6 on CONTROL: Full versus Reduced Models

Variable	Coef.	s.e.	t	SSE	R^2
Full:				17.34	.06
INTERCEPT	.90	.02	37.69*		
CONTROL	-.04	.02	-2.06*		
RCGROUP	-.09	.03	-2.72*		
CONTROL*RCGROUP	-.00	.03	-0.14		
Reduced:				17.84	.03
INTERCEPT	.85	.02	52.23*		
CONTROL	-.04	.01	-2.86*		

Comparing full and reduced models:
F = 3.69*
df (2, 256)

*Prob. < .05.