DOCUMENT RESUME

ED 360 829 FL 021 379

AUTHOR Deville, Craig W.; Chalhoub-Deville, Micheline
TITLE Modified Scoring, Traditional Item Analysis, and Sato's Caution Index Used To Investigate the Reading Recall Protocol.
PUB DATE [93]
NOTE 16p.; Paper presented at the Annual Language Testing Research Colloquium (15th, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS German; *Item Analysis; *Reading Comprehension; *Reading Tests; *Recall (Psychology); *Scoring; Second Language Learning; *Second Languages; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Caution Index (Sato)

ABSTRACT

A study demonstrated the utility of item analyses to investigate which items function well or poorly in a second language reading recall protocol instrument. Data were drawn from a larger study of 56 learners of German as a second language at various proficiency levels. Pausal units of scored recall protocols were analyzed using both classical item analysis and the Sato Caution Index. The assumptions of classical local independence and unidimensionality were found to be largely fulfilled in this analysis. Pausal units did not fulfill the noninvasiveness condition, but this was not found detrimental to the investigation. Results concerning item difficulty, point biserial correlations, the discrimination index, reliability, item weighting, and use of the Sato caution index are summarized. Results of the Sato caution analysis were found to corroborate those of classical analysis, indicating that item analyses of dichotomously scored recall protocol units can be applied and yield interpretable results. It is not suggested that the Sato Caution Index supplant classical item analyses, but that it be used whenever integrative language tests using summated item scores are used to make theoretical, instructional, or evaluative decisions. A 30-item bibliography is included. One text used and its item analysis results are appended. (MSE)

Modified Scoring, Traditional Item Analysis, and Sato's Caution Index Used to Investigate the Reading Recall Protocol

Craig W. Deville & Micheline Chalhoub-Deville

Modified Scoring, Traditional Item Analysis, and Sato's Caution Index Used to Investigate the Reading Recall Protocol

Craig W. Deville & Micheline Chalhoub-Deville
The Ohio State University
287 Arps Hall
1945 N. High Street
Columbus, Ohio 43210
e-mail: CDEVILLE@MAGNUS.ACS.OHIO-STATE.EDU
MCHALHOU@MAGNUS.ACS.OHIO-STATE.EDU

I Background and Rationale

1 The Recall Protocol

The recall protocol is an assessment instrument which requires readers to read or listen to a short passage and then to write everything that they remember about it. The recall protocol is increasingly being used in second language reading research as a measure of comprehension (Bernhardt, 1983, 1991; Carrell, 1983, 1984a, 1984b; Connor, 1984; Lund, 1991). The procedure has been recommended as a measure of reading comprehension over other traditional instruments, such as multiple choice and cloze tests, because it "circumvents the pitfalls" associated with traditional tests, e.g., it provides no leading information pertaining to passage content, and requires the reader to integrate the components of the reading passage (Bernhardt, 1991: 28). Bernhardt reports that the recall protocol provides data that reflect the nature of the reading process in terms of encoding, restructuring, and analyzing information. In addition, she claims that the recall protocol is a more valid measure of reading comprehension because it conforms to current second language (L2) reading research-driven theories.

No doubt the most important characteristic of an instrument is its validity. Validity, however, is impossible to achieve without reliability (Bachman, 1990; Ghiselli, Campbell, and Zedeck, 1981). Although formal analyses of reliability are routinely performed on other reading tests, these procedures have been conspicuously absent with regard to the written recall protocol. Bernhardt and Deville (1991) discuss the importance of statistical analyses that would support the reliability of the recall protocol, yet a review of the studies in which the recall protocol has been used indicates that none of these analyses has yet been undertaken. Inter-rater reliability has been the only statistic reported. Inter-rater reliability, however, should not be confused with test reliability. An inter-rater reliability index simply indicates the extent of agreement between raters in their performance or assessment of a certain task, and test reliability pertains to the consistency of the test instrument. It is possible for inter-rater reliability on a test to be high but for test reliability to be low (Ebel and Frisbie, 1991). In other words, raters may agree in their assessment of subjects' ability on a specific instrument. The instrument, however, may not be tapping the particular trait in an accurate and a dependable fashion, thus rendering whatever assessment raters have agreed upon unreliable, and consequently invalid.

Traditional item and reliability analyses, routinely performed on other measures, are needed in order to assess the internal consistency of

recall protocol instruments. Until such analyses are performed any claims as to the validity of the procedure are premature.

2 Scoring the Recall Protocol

One plausible reason why item and reliability analyses have not been performed on the recall protocol may be the essay-like nature of the measure. Although the recall protocol is essentially an essay, it is scored as though it were comprised of discrete units. According to the Meyer (1975) scoring/analysis system, a reading passage is arranged into a hierarchical tree structure. Information positioned at the top levels of the structure represents the main ideas of the reading passage, and information located at the low levels represents detail. A scoring template is thus formed to reflect the lexical units, often referred to as propositions, in the passage and the relationships between and among these units. A recall protocol is scored according to the presence or absence of the idea units represented in the scoring template. Following this scoring system, a total score on the written recall protocol is derived by summing the scores that correspond to the lexical units recalled.

Another frequently used scoring method is the Johnson (1970) system, whereby a reading passage is divided into pausal units based on normally paced oral reading. Each pausal unit is weighted on a scale of one to four, depending on its importance to passage meaning, one being least important and four most important. Once the pausal units are weighted, a scoring template is developed and followed when scoring readers' protocols. According to this scoring procedure, the total score on a recall protocol is the sum of the weighted pausal units recalled by the subject. It is clear then, that in both the Meyer and Johnson systems the propositions/pausal units are treated as discrete items for scoring purposes. In the remainder of this paper the terms "proposition," "pausal unit," and "item" will be used interchangeably. The authors believe this will clarify the measurement ideas put forth without doing excessive injustice to these concepts as used in the reading literature.

In summary, although the recall protocol is an essay-like instrument, the total score derived is based entirely on summing the discrete units correctly recalled. Consequently, item and reliability analyses comparable to those run on multiple choice tests can and should be performed on recall protocols and on other integrative measures. Cziko (1982) and Cziko and Nien-Hsuan, 1984), using a dichotomous scoring procedure to evaluate the reliability of essay-like responses to dictation and other tasks, found good estimates of internal consistency. They also determined that these integrative test items fit the Guttman cumulative scaling model, meaning among other things, that the items constituted a unidimensional scale. This aspect of unidimensionality will be discussed in more detail below.

II Purpose

The purpose of this study is to demonstrate how item analyses can be performed on the recall protocol instrument, and to illustrate how these analyses can provide information about both the pausal units of the reading text and the subjects. It is important to state that the function of the present study is not to uphold the recall protocol instrument over other measures of reading comprehension, but to

investigate the appropriateness and usefulness of information gained from item analyses of the reading recall protocol. Results from such analyses are often used to identify items functioning well and items functioning poorly. Poor items need to be revised or discarded. Revising and discarding pausal units, however, would diminish the authenticity of the reading text; nevertheless, the information provided by such item analyses could point to the inappropriateness of including poorly functioning pausal units in the final scores for subjects.

III Methodology

1 Data Collection and Subjects

The data used in the present study were originally collected as part of an investigation of the effect of anaphora on L2 reading comprehension (Berkemeyer, 1991), and kindly made available to the authors. The text, an excerpt of a German short story (Appendix A), is short enough to be read by most students of German in a few minutes. It is somewhat simple and straightforward in its content so that language learners from all levels can read and understand the main idea, yet it contains nuances in both language and interpretation.

Fifty-six subjects at various German proficiency levels provided the data employed in the present analysis. Berkemeyer used the Johnson (1970) weighted scoring system to analyze her subjects' protocols and reported a value of .99 for both inter- and intra-rater reliability.

In the present analysis the pausal units of the recall protocol were also scored dichotomously as though they were items in a multiple choice test with right and wrong answers. Subjects were given a one for every pausal unit they got "correct," i.e., that they included in their protocol, and a zero for "incorrect" responses, i.e., where that information was not included in the protocol, or where information not depicted in the scoring template was added. As such, a total score was obtained by summing the number of correct responses. The scores on the 46 pausal units in the reading passage ranged from 9 to 40, with a mean of 24.64 and a standard deviation of 9.00, which indicates that the sample included subjects with a wide range of language abilities.

2 Classical Test Theory Measures

In this study each pausal unit of the scored recall protocols will be analyzed separately using item analyses based on classical test theory. Classical test theory provides a structured analysis of test items, known as item analysis, that assists the test developer in evaluating the validity and reliability of a test as a measuring instrument. The present analysis incorporates item difficulty level, item discrimination index, point biserial correlations, and reliability estimates.

Recently, researchers (Blixt and Dinero, 1985; Harnisch, 1983; Tatsuoka and Tatsuoka, 1983) have looked beyond summated item scores to the relationship between subjects' response patterns and their total test score. These researchers point out that because a total score is the sum of correct responses, it is possible for different subjects to get different items correct and still receive the same total score. The total test score by itself, therefore, may not afford an accurate indication of differences in subjects' performances. Further analyses that examine subjects' response patterns are needed.

Sato (1975) developed a method that can be used to investigate and index response patterns, producing three indices: a disparity coefficient, a problem (item/pausal unit) index, and a subject index. The Sato Caution indices, along with classical item analysis, will be used to evaluate the quality of the recall protocol as a reading comprehension measure.

3 Local Independence and Unidimensionality

Performing reliability analyses requires fulfilling the local independence assumption. This is especially pertinent to integrative language tasks such as the recall protocol, cloze, dictation, and similar procedures, where a subject's responses may not be independent of each other (Bachman, 1990; Wainer and Lewis, 1990). In his review of the topic, Henning (1989) identifies three distinct characteristics of the local independence assumption:

> (1) classical local independence, where item responses are independent at fixed ability levels; (2) unidimensionality, where one and only one trait is required to define the latent space; and (3) noninvasiveness, where item responses are independent of the sequence in which items are encountered (p. 106).

It is obvious that responses to items are related, i.e., correlated. Yet, if the trait generating this relationship is "partialled out," then the items should be independent (Hambleton, Swaminathan, and Rogers, 1991). One approach to demonstrate whether the condition of classical local independence is satisfied is the "noncorrelation... among items for persons of the same ability level" (Henning, 1989: 104). In order to check this assumption in the present context, a correlation matrix of the recall protocol items for those subjects who were plus or minus one standard error of measurement from the mean was obtained. Results revealed that 93% of the inter-item correlations were not significant ($p<.05$) for this same-ability subgroup of subjects, providing support for the local independence condition.

Henning (1989) and Cziko and Nien-Hsuan (1984) maintain that the unidimensionality condition can be established through a Guttman scale approximation. The unidimensionality of the recall protocol pausal units was examined using the Sato (1975) procedure. The resultant disparity coefficient, which can be used as an index of departure from the Guttman model, was a low .31. This value is well below the value of .50 recommended by Sato.

A second, relatively new technique for examining test dimensionality was also employed (Chen and Davison, 1993). This approach can be used with small sample sizes, unlike most approaches (see Roznowski, Tucker and Humphreys, 1991, for recent discussion of approaches to unidimensionality), thus making it the logical choice in the present context. The procedure involves the multidimensional scaling (MDS) of a proximities matrix of paired comparisons of items. The proximities are based on conditional probabilities of answering a particular item correctly. (A program to obtain the proximities matrix can be made available by the first author upon request).

The proximities matrix for the recall protocol items was submitted to MULTISCALE (Ramsay, 1977, 1991), a MDS procedure that makes use of

maximum likelihood estimation. Using MULTISCALE the chi-square difference between a one and two dimensional solution can be obtained and tested for statistical significance. The test for the recall protocol items was found to be significant, indicating that a second dimension improved the fit of the MDS solution (X^2=752.472,df=44). This test, however, is best used as a guideline, as it is notoriously stringent. Chen and Davison (1993) suggest that one examine the stimulus configuration to determine dimensionality. In the present context this did not provide a definitive answer to the question of dimensionality. Although most of the items clearly varied along the first dimension (as expected), there was some, albeit small, variability along the second as well. Nevertheless, the correlation of the item "loadings" (actually, coordinates) on the first dimension between the two solutions was .97, indicating that the second dimension might simply be error. This procedure is new and is still being refined, rendering any results as necessarily tentative.

In summary, the two assumptions of classical local independence and unidimensionality are largely fulfilled by the results from the present analysis of the recall protocol, and although the pausal units do not fulfill the noninvasiveness condition (to be discussed in more detail below), Henning maintains that if the first two assumptions are fulfilled, and test items retain the same sequence upon different test administrations, then "the various latent trait models and classical true-score applications can be said to apply" (1989: 103).

IV Results

1 Item Difficulty

P-values indicate the difficulty level of the particular item with respect to the group of examinees. The p-values for the recall protocol range from .04 to 1.00 (Appendix B), revealing that there was one very difficult item that only two subjects included correctly in their protocol (item #17), and one that everyone got correct (item #10). The average p-value across the 46 pausal units is .53, demonstrating that the test is indeed providing differential information on the examinees. One might, depending on the purpose of the test, consider not scoring item #10 in the final analysis because it provides no differential information on the test takers, i.e., it does not differentiate between readers at different ability levels.

Examining the p-values in Appendix B, it will be noted that identical pausal units appearing in various places in the passage have different p-values. The pausal unit "I want," for example, appears three times in the reading passage and has p-values of .84, .50, and .66. For the present discussion we will assume these different p-values are not the result of random error, but are due to other factors. Actually, the varied p-values should not be surprising because although these pausal units are comprised of the same words, their linguistic contexts are different. Different contexts will have their unique linguistic and semantic characteristics, making words or phrases -- although apparently alike -- acquire new dimensions that will consequently result in differential comprehension by readers (Brown and Yule, 1983). Brown and Yule quote Fillmore (1977), who writes "...I find that whenever I notice some sentence in context, I immediately find myself asking what the effect would have been if the context had been slightly different" (p. 35).

The recall protocol pausal units are embedded in particular discourse contexts within the text that constrain their comprehension and interpretation. As a result, several pausal units were recalled differentially according to their context in the sentence and text, a violation of the assumption of noninvasiveness discussed above. Although the item responses may not be independent of the sequence, i.e., context in which they are encountered, as long as the pausal units retain a fixed administration sequence, and both classical local independence and unidimensionality conditions are fulfilled, then the invasiveness "may be of little statistical consequence" (Henning, 1989: 107). Nevertheless, this issue deserves further attention.

2 Point Biserial Correlations

Ideally, there should be a positive relationship between all test items and the total test score. The results of the recall protocol examined yielded point biserials ranging from .12 to .78 (Appendix B), the average being .43, using a Fisher z' transformation. No point biserial estimate for item #10 was included because it was answered correctly by all subjects, and thus offers no variability. Otherwise, all point biserials are positive, indicating that the scores on the individual items are consistent with total test scores.

Similar to the p-values, pausal units with identical words resulted in different point biserial indices. These pausal units are really dissimilar because of their different contexts, and consequently, yield different comprehension results.

3 Discrimination Index

Another frequently reported item property is the discrimination index. The discrimination indices in the present analysis are based on the difference in the proportion of correct items between the top 26% and the bottom 26% scoring subjects. The resulting indices from the recall protocol range from 0.0 to 1.0, with a mean across all items of 0.54, and no negative values (Appendix B). Item #10 was answered correctly by all students, hence no discrimination is provided. Item # 17 also fails to provide discrimination because no subjects from either the high or low scoring groups included the pausal unit in their protocols. Several items were answered correctly by all subjects in the top group but by none in the low group, resulting in values of 1.0. The discrimination indices for some the same-worded pausal units differ for the reasons already discussed.

4 Reliability

The Johnson (1970) and Meyer (1975) scoring procedures treat the recall protocol units/propositions as individual items to arrive at total scores that are then used to make research, evaluation, and instructional decisions. In order to ascertain if the total scores are meaningful and appropriate, internal consistency estimates are needed to supplement inter-rater agreement indices.

Cronbach's alpha and Guttman's split-half reliability estimates were obtained on the recall protocol, resulting in values of .913 and .890 respectively. With regard to the split-half estimate, the items were split every fifth pausal unit beginning with item #1. The next iteration began with item #3, again proceeding with every fifth pausal unit, etc.

In this fashion, no adjacent, and relatively few intrasentential, pausal units clustered in either half. The analyses also revealed that there were no items that would increase alpha if deleted more than a minimal .002. These internal consistency indices indicate that the responses to the items are consistent, and they provide additional information as to the quality of the instrument as used in the present context.

Nevertheless, the issue of item independence in integrative language measure needs to be investigated further with respect to inflated internal consistency and item statistic estimates. Such analyses may be inappropriate, while other reliability indices such as test-retest or alternative forms may be more suitable (Bachman, 1990).

5 Item Weighting

Up until this point, classical item analyses have been applied to the dichotomously scored recall protocol. Although these procedures typically use dichotomous data, the recall protocol is often weighted (Johnson, 1970) and scored accordingly. Some psychometricians (Ebel and Frisbie, 1991; Ghiselli et al., 1981; Nunnally, 1978), however, question the effectiveness of weighting on scoring. In order to investigate the possible influences of weighting on reliability and on scores from the recall protocol, comparisons between unweighted and weighted recall protocol items were undertaken.

The Cronbach's alpha obtained from the weighted protocols is .905, virtually the same as the .913 obtained using only dichotomous scores. The subjects' total dichotomous test scores were then correlated with their total weighted test scores, the correlation being .988. This high correlation indicates that there is essentially no difference in subjects' relative total scores whether the recalls are scored dichotomously or are weighted.

In summary, this evidence indicates that researchers and classroom teachers can forego the weighting system and simply score the protocols dichotomously. Dichotomous scoring will save both researchers and teachers the time and effort currently being expended on the process of weighting propositions. Educators can use the recall protocol without having to spend inordinate time preparing the scoring template of the test.

There are two caveats regarding the weighting, the first pertaining to the scoring system used. The present analysis employed the Johnson system. Researchers who utilize another scoring system, such as that by Meyer, will obviously obtain a different set of propositions that may influence results. This remains to be seen, however, as Bernhardt (1991) found very high correlations between total test scores obtained from the two methods, leading her to conclude that "there is enough overlap in the scores to argue that both systems are tapping the same behavior" (p. 216).

The second caveat is that researchers may have substantive reasons for grouping items together for analysis. Both Lund (1991) and Bernhardt (1991) report that the number of propositions recalled differs according to the linguistic level of the proposition. Lund, who used a modified Meyer weighting system, reported that the propositions higher up in the

hierarchy are comprehended by more subjects. Bernhardt, using the Johnson system, reported that subjects have the most difficulty with level two propositions.

An ANOVA was performed on the present data by comparing the mean p-values across the four levels of importance. The results were significant, $F(3,42)=3.62$, $p<.05$. Post hoc Scheffe' comparisons reveal the only difference, however, to be between levels two and three. Subjects tended to recall the most information from level three, and the least from level two. These results correspond somewhat to what Lund and Bernhardt have reported, namely that subjects recall more of the material pertinent to the main ideas of the text and less of the detail. Further research is needed to elaborate the textual factors that influence subjects' recall.

6 Sato Caution Index

While the total test score may provide important information regarding the relative standing of subjects with respect to each other, this unit of measurement does not reveal information about an individual's pattern of item responses. Test developers, researchers, and teachers may be interested in knowing not only which items examinees found easy or difficult, but whether the pattern of response yields useful information about the students or the test itself (Blixt and Dinero, 1985; Harnisch, 1983; Tatsuoka and Tatsuoka, 1983).

Sato (1975) has derived a summary statistic for dichotomously scored items that indicates the extent of departure from a perfect Guttman scale, called the disparity coefficient. The disparity coefficient indicates the extent of disparity between students' response pattern and items' response pattern. Sato considers disparity coefficient values of .4 acceptable, while values above .6 indicate caution. Blixt and Dinero (1985) recommend .50 as the value above which caution is warranted. A high disparity coefficient can signal that either the items are heterogeneous and not functioning well together, or that the group of examinees performs inconsistently across the items.

The disparity coefficient obtained on the reading recall protocol is .31. This value in conjunction with the high Cronbach's alpha indicate that the pausal units of the recall protocol form a relatively homogeneous set of measures. The computer program also yields a caution index for each student and each item similar in concept to item and person fit indices in item response theory. Results of the recall protocol analysis indicate that all subjects and items were below the recommended .50 caution index level.

Some new refinements of the Sato procedure (D'Costa, 1993; D'Costa and Deville, forthcoming) indicate that the caution indices may not always uncover erratic test takers or inconsistent items because the typical indices combine errors made, e.g., those within a particular examinee's ability level and those beyond his/her level. Separating these two types of errors into distinct indices (Appendix B) reveals that the "omnibus" Sato might be below the recommended cutoff, while the separate W and B indices are above the cutoff. Simulation and other studies are in progress to determine how the W and B indices function in

relation to different data parameters and in relation to the more conventional item indices.

Because these procedures are still under investigation, only limited discussion and substantive examples will be offered here. The W index for item #17 is 1.00, well above the Sato value of .28. The caution, however, indicated by the W index appears to correspond with the low point biserial and discrimination index for the item (.12 and .00, respectively), as might be expected for an item exhibiting little variability. The high W index means that the inconsistent (i.e., inconsistent in a deterministic scaling or Guttman approach) response pattern to the item were such that all "errors" were of the type where subjects answered the item correctly although the item was not within their ability level. Similarly, the high B index of 1.00 for item #5 indicates that the errors made on this item were of the kind that were not beyond the ability level of the subjects, i.e., perhaps careless errors. This diagnostic information about items and test takers can be helpful to test developers and practitioners when examining test performance. In addition, whenever summated item scores are used on such integrative measures, one might consider whether to include (score) such items and/or examinee responses. (A program to calculate the Sato, W, and B indices can be made available by the first author upon request).

V Conclusion

The results from the Sato Caution analysis corroborate those from the classical test analyses, all indicating that item analyses of dichotomously scored recall protocol pausal units can be applied and can yield interpretable results. Because the purpose of this study is to illustrate the application of modified scoring and item analysis to the reading recall protocol, results from but one text were presented in detail. The authors have examined data from other applications of the reading recall protocol, all using the Johnson scoring system, and obtained favorable item statistics and reliabilities in the range of .70-.90. In addition, since results reported in this study were generated from the reading of a single German text and scored using the Johnson pausal unit analysis system, these results need to be replicated using other texts, other linguistic or text units, and other weighting systems.

The intention in this paper is not to promote the recall protocol as a superior integrative measurement instrument. Rather, the straightforward procedures presented in this study should be employed whenever integrative language tests using summated "item" scores are utilized to make theoretical, instructional, or evaluative decisions, thus ensuring that such scores are the result of meaningful and appropriate items.

References

Bachman, L.F. 1990: Fundamental considerations in language testing. Oxford: Oxford University Press.

Berkemeyer, V.C. 1991: The effect of anaphora on the cognitive processing and comprehension of readers of German at various levels of baseline German language ability. Unpublished doctoral dissertation, The Ohio State University.

Bernhardt, E.B. 1983: Three approaches to reading comprehension in German. Modern Language Journal 67, 111-15.

---. 1991: Reading development in a second language: theoretical, empirical, and classroom perspectives. Norwood, NJ: Ablex Publishing Corporation.

--- and Deville, C.W. 1991: Testing in foreign language programs and testing programs in foreign language departments: reflections and recommendations. In Teschner, R.V., editor, Issues in language program direction: assessing foreign language proficiency of undergraduates, Boston, MA: Heinle & Heinle Publishers, Inc.

Blixt, S.L. and Dinero, T.E. 1985: An initial look at the validity of diagnoses based on Sato's caution index. Educational and Psychological Measurement 45, 55-61.

Brown, G. and Yule, G. 1983: Discourse analysis. Cambridge: Cambridge University Press.

Carrell, P. 1983: Three components of background knowledge in reading comprehension. Language Learning 33, 183-207.

--- 1984a: Evidence of a formal schema in second language comprehension. Language Learning 34, 87-113.

--- 1984b: The effects of rhetorical organization on ESL readers. TESOL Quarterly 18, 441-70.

Chen, T. and Davison, M.L. 1993: A multidimensional scaling, paired comparisons approach to assessing unidimensionality in the Rasch model. Paper presented at the Seventh International Objective Measurement Workshop, Atlanta, April 1993.

Connor, U. 1984: Recall of texts: differences between first and second language learners. TESOL Quarterly 18, 239-56.

Cziko, G. 1982: Improving the psychometric, criterion-referenced and practical qualities of integrative language tests. TESOL Quarterly 16, 367-79.

Cziko, G. and Nien-Hsuan, J. 1984: The construction and analysis of short scales of language proficiency: classical psychometric, latent trait, and nonparametric approaches. Paper presented at the 68th Annual Meeting of the American Educational Research Association, New Orleans, April 1984.

D'Costa, A. 1993: Extending the Sato caution index to define the within and beyond ability caution indexes. Paper presented at the National Council for Measurement in Education Conference, Atlanta, April 1993.

D'Costa, A. and Deville, C.W. Forthcoming: Interpreting the W, B, and Sato caution indexes.

Ebel, R. and Frisbie, D.A. 1991: Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ghiselli, E., Campbell, J., and Zedeck, S. 1981: Measurement theory for the behavioral sciences. New York: W.H. Freeman and Company.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. 1991: Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Harnisch, D. 1983: Item response patterns: applications for educational practice. Journal of Educational Measurement 20, 191-206.

Henning, G. 1989: Meanings and implications of the principle of local independence. Language Testing 6, 95-108.

Johnson, R. 1970: Recall of prose as a function of the structural importance of linguistic units. Journal of Verbal Learning and Verbal Behavior 9, 12-20.

Lund, R.J. 1991: A comparison of second language listening and reading comprehension. Modern Language Journal 75, 196-204.

Meyer, B. 1975: The organization of prose and its effects on memory. Amsterdam: North-Holland.

Nunnally, J.C. 1978: Psychometric theory. New York: McGraw-Hill.

Ramsay, J.O. 1977: Maximum likelihood estimation in multidimensional scaling. Psychometrika 42, 241-66.

Ramsay, J.O. 1991: MULTISCALE: Computer program and manual, extended version. Montreal: McGill University.

Roznowski, M., Tucker, L.R. and Humphreys, L.G. 1991: Three approaches to determining the dimensionality of binary items. Applied Psychological Measurement 15, 109-27.

Tatsuoka, K. and Tatsuoka, M. 1983: Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement 20, 221-30.

Wainer, H. and Lewis, C. 1990: Toward a psychometrics for testlets. Journal of Educational Measurement 27, 1-14.

APPENDIX A

Text

Im Volksgarten

"Ich möchte einen blauen Ballon haben!" sagte Anna.
"Da hast du einen blauen Ballon, Anna!" sagte die Mutter.
"Ich möchte ihm die Luft auslassen," sagte sie einfach.
"Willst du ihn nicht diesem armen Mädchen schenken, Anna?!"
"Nein, ich will ihn fliegen lassen!" Sie läßt den Ballon aus, sieht ihm nach, bis er verschwindet.
"Mutter, ich hätte ihn lieber dem armen Mädchen geschenkt!"
"Da hast du einen anderen blauen Ballon, Anna, schenke ihr diesen!" sagte die Mutter. (Paul Altenberg)

Source: Kunst, H. 1977: Texte und Übungen: intermediate readings and exercises. Englewood Cliffs, NJ: Prentice-Hall, Inc.

APPENDIX B

pausal units with p-values, point biserials, discrimination and Sato indices

Pausal Units	P-Value	Pt_{bis}	Discrim	Sato	W	B
1. In the People's Garden (Park)	.29	.19	.28	.31	.74	.25
2. "I want (would like)	.84	.23	.25	.35	.11	.66
3. a blue balloon	.93	.19	.17	.29	.05	.75
4. to have (to get)!"	.11	.39	.45	.08	.49	.04
5. said Anna.	.95	.15	.08	.30	.04	1.00
6. "There (already, then)	.54	.13	.31	.40	.44	.55
7. you have	.82	.29	.33	.31	.12	.59
8. a blue balloon,	.71	.38	.50	.26	.17	.48
9. Anna!"	.13	.30	.27	.17	.57	.06
10. said the (her) mother.	1.00	-	.00	-	-	-
11. "I want (would like)	.50	.40	.65	.25	.30	.37
12. the balloon	.50	.54	.82	.15	.24	.21
13. the air	.48	.43	.73	.21	.29	.24
14. out of (from)	.27	.39	.64	.17	.45	.14
15. to let (release),"	.54	.53	.82	.17	.22	.24
16. said she	.55	.25	.49	.33	.36	.45
17. simply.	.04	.12	.00	.28	1.00	.03
18. "Do[n't] you want (Would[n't] you like)	.73	.67	.92	.07	.06	.15
19. it (the balloon)	.75	.77	.92	.01	.02	.06
20. not	.39	.63	.82	.08	.26	.12
21. this (the) poor girl	.71	.78	.92	.01	.05	.16
22. to give (to present).	.68	.71	.83	.06	.06	.15
23. Anna?!"	.11	.22	.18	.22	.68	.06
24. "No,	.39	.59	.82	.10	.31	.15
25. I want	.66	.15	.24	.40	.23	.51
26. it (the balloon)	.64	.19	.33	.38	.24	.48
27. to let fly (go, fly away)!"	.66	.17	.33	.39	.23	.50
28. She (Anna) lets out (lets go of, releases, lets loose)	.86	.28	.33	.29	.08	.51
29. t e balloon	.82	.34	.41	.27	.09	.51
30. looks after it,	.30	.32	.37	.23	.48	.17
31. until it disappears.	.23	.21	.28	.28	.61	.16
32. "Mother,	.13	.36	.37	.11	.71	.06
33. I would have (should have)	.30	.71	1.00	.07	.11	.14
34. it (the balloon)	.64	.70	1.00	.08	.10	.17
35. rather (better)	.21	.43	.64	.12	.41	.09
36. (to) the poor girl	.55	.76	1.00	.05	.09	.10
37. given (presented)!"	.50	.70	1.00	.07	.13	.11
38. "There (already, then)	.42	.36	.39	.24	.30	.21
39. you have	.64	.52	.67	.18	.16	.33
40. another	.88	.45	.42	.14	.04	.43
41. blue balloon,	.86	.27	.24	.30	.08	.50
42. Anna,	.30	.23	.28	.29	.64	.24
43. give (present)	.39	.50	.73	.15	.35	.18
44. (to) her (the girl)	.38	.42	.64	.19	.38	.19
45. this one!"	.32	.48	.82	.14	.37	.15
46. said the (her) mother.	.70	.57	.92	.15	.12	.25