A Validity Study of an Adaptive Test of Reading Comprehension

Lutz F. Hornke and Michael P. Sauter
University of Düsseldorf

Adaptive means that a test adapts to the testee´s proficiency level in the proper "can-do" sense. A fair number of items are placed at the testee´s disposal; and solely by means of tactical rules, the testees self-select their own individual subset of items. To achieve this, their previous responses are used to help in making item-to-item decisions. In addition, restrictions on test time are imposed to insure unidimensional interpretations.

In the literature many variants of adaptive schemes are described and discussed (see Hornke, 1976, 1977, 1979a, 1979b, 1979c; Hornke, Sauter, Suessmilch, & Burghoff, 1979; Lord, 1971; Weiss, 1974; 1975; Weiss & Betz, 1973; Wood, 1973). Generally speaking, the idea utilized is that of branching from item to item or between groups of items utilized: The item someone is branched to is made contingent on his/her response(s) to earlier item(s). Thus, whenever a testee answers an item correctly, he/she is presented with more a difficult one on the assumption that his/her proficiency level at this intermediate stage is somewhat higher than that displayed in the item just mastered. The contrary holds for incorrect responses. The complexity and variety of branching rules is not limited (see Hornke, 1976; Weiss, 1978). The more flexible the branching technology, the more adaptive the decision process will be, and this yields very reliable information about a testee´s proficiency and his/her can-do potential.

The term branching technology is used here intentionally because many adaptive testing projects already use computers. According to highly sophisticated estimation procedures based on probabilistic mathematical response models (see Fischer, 1978; Lord & Novick, 1969), items are deliberately retrieved from a larger pool. These approaches use item parameters to estimate a person´s probable standing. After several cycles of item administration and parameter estimation, a person parameter emerges that confidently reflects an individual´s proficiency level. Since items and persons are calibrated on the same scale, by looking at those items (i.e., behaviors), the parameters of which lie in the vicinity of the person parameter, interpretations are readily available.

Computer terminals and micro-computers are quite costly, however, so that paper-and-pencil versions deserve some attention. The basis of their measurement is somewhat less stringent compared with flexible computer-assisted tests; but when properly designed, they should allow equivalent or even better measurement precision than conventional tests (see Hornke, 1979b, Hornke et al., 1979).

The test booklet may look the same as that for conventional tests; the difference is that the testee is asked to use a special pen for marking his/her answer. He/she has to pass this lightly over a bracketed field next to the chosen answer. Chemicals then react and render visible the number of an item to be attempted next. By following these numbers as they appear, a testee is branched through the item set (see Hornke, 1979a; Sauter, 1978, 1979; Sauter & Hornke, 1979). The testee is intended to be guided to just that subset of items that tells something about his/her can-do level, while leaving out all the other boring or otherwise frustrating items. Since a testee zig-zags through a pyramidal item arrangement, he/she will finally end in a score category, a self-evaluating feature of this tactical test design.

Thus, with branching tactics, flexible, fair, self-scoring, and interpretable tests are at hand. Since any mathematical response model or pyramidal pencil-and-paper test rests, respectively, on the quality of the items and the model or arrangement more successful assessment is guaranteed as long as quality levels are maintained. Even conventional tests, however, require some degree of item validity and reliability, unless any interpretation is better than random guesswork. Whether and how adaptive tests will and should be used is still an open research question.

Adaptive Test Designs

Individualization is a concept that meets approval on many different sides. To some extent, assessing an individual in his/her own right solely by what he/she is doing seems fair. Saving time by asking nonsuperfluous questions capitalizes more on the economy and less on the psychology of testing, though in that area, too, something might be gained. Reduction of the stress induced by testing, maintenance of motivation, and lack of boredom are but a few psychological effects. So far very little is known about these side effects and the benefits of individualized testing; these seem to be areas of potential that await further evaluative research.

At present, individualized testing is thought to have positive or at least non-negative effects on testees. To understand the entire range of adaptive programs better, three possible adaptive designs are considered below.

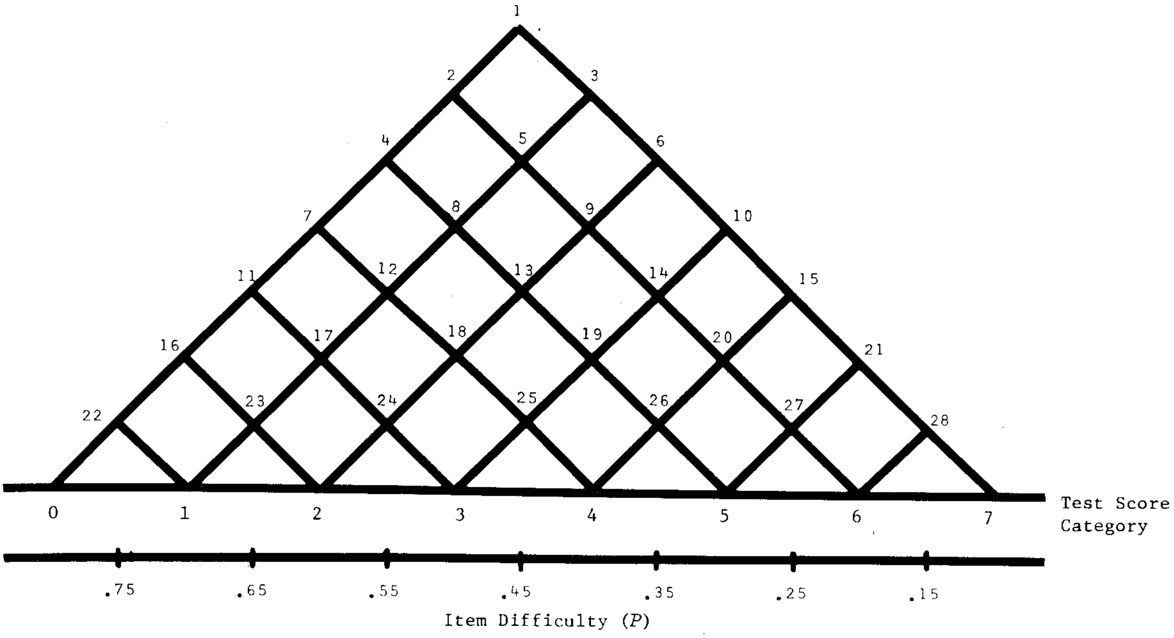
Curtailed item sampling. This approach, a naive type which has some intuitive appeal, resembles the examination models used in classrooms. A teacher asks a student several questions, with content and complexity varying according to the answers given. After a specified period of time the teacher stops and evaluates the student. In comparing several oral examinations, considerable variation would easily be found in the number as well as in the difficulty of questions: This is a genuinely adaptive approach. Thus, two students may earn the same grade but may have been asked different questions as far as number and/or complexity was concerned. Variation in the number seems fair because students who are asked more have a chance to demonstrate their true behavior level; whereas with others, final evaluations are quite obvious after only a few questions.

Computer-assisted testing. Curtailing the numbers of questions, i.e., restricting the sampling of items from a behavioral domain, is a reasonable decision. For adaptive tests this would mean evaluating a testee's distance from a set of criterion levels. Testing is stopped when, for a fixed number of items, a testee is irrevocably located on either side of the decision point. This may be achieved fairly soon. When there are 16 items and the criterion is set at 50%, testing should stop after 8 correct responses. However, this could occur when all the first 8 responses are correct. A testee who made an error on the first 2 items has to be tested for at least 8 more items, yielding a total of 10. Varying numbers of items will occur when several students are tested, one typical aspect of flexible adaptive tests. The example of an oral examination given above dealt with two possible adaptation criteria: (1) the number of questions before a terminal decision can be made and (2) the quality of questions needed to make a procedural decision. A very flexible adaptive testing program will have to consider both criteria; this may be possible with computer-assisted testing (see Weiss, 1975, 1978).

Paper-and-pencil pyramidal tests. Since large-scale adaptive testing by means of computers is hampered by costs, other means have been invented and used to achieve a branching test system, even with group testing; and a pyramidal test design for use with paper-and-pencil devices has emerged. According to its feasibility and overall value, it lies somewhere between curtailed sampling and computer-assisted testing. By pyramidal is meant an item arrangement that is structured like a network. For a certain population the item locations on some dimensions are known.

In order to design such a test, items are deliberately selected to form a desired hierarchical item order (see Figure 1). At the top the testee gets the starting item (Item 1), which has to be answered by each candidate. When a cor-

Figure 1
Model of a Pyramidal Item Order



rect answer is given, testees are branched to the right. Consequently, a more difficult item has to be attempted. The contrary holds for an incorrect response. Thus, contingent on their responses, testees are individually branched through the item arrangement and will finally end in a test score category that tells something about the behavioral level attained.

To contrast this approach with curtailed and computer-assisted testing, it becomes quite obvious (1) that there are available far more items than a given testee has to attempt, (2) that testees find their individual paths through the item network and come (or ought to come) close to the upper bounds of their proficiency level, (3) that testing ends after a preset number of items has been attempted, and (4) that the final item leads directly to a test score category, i.e., no further scoring is necessary because the test is essentially self-scoring. The dominant design feature is to adapt the quality of the items, and not their number, to any testee.

The pyramidal test is a fixed strategy as far as item number and arrangement are concerned, but a testee works more or less flexibly on items that are assumed to suit his/her proficiency level more and more. The technical problems with the pencil-and-paper format and group testing were undertaken by means of chemicals. The list of adaptive test designs here is far from complete; many other versions have been described (e.g., see Hornke, 1976; Weiss, 1976, 1978). The report above was meant to examine closely various construction characteristics, i.e., flexibility in item number, item difficulty, or both.

Construction of an Adaptive Pyramidal Test

The studies of Sauter (1978) and of Hornke et al. (1979), looked closely at the adaptive test format and especially at the pyramidal item order in use. It was the aim of both Sauter (1978) and Hornke et al. (1979) to construct and to evaluate such a test design; nevertheless, the choice of the linguistic item material was not accidental. The pyramidal item order requires question forms that can be evaluated objectively, e.g., multiple-choice items or items with a blank. Moreover, it should be possible to rank these items according to their empirical, as well as according to their content, difficulty, which should reflect a higher level of linguistic competence. In addition, the choice of the item material was influenced by the fact that it was not possible, or necessary for this purpose, to construct and to evaluate new items. It was therefore inevitable to seek proven items in existing tests.

One test that approximately meets the above prerequisites is the Cologne Placement Test (see Bonheim & Kreifelts, 1979), which is a traditional placement test for students at the beginning of their first semester in the course "English as a Foreign Language." It consists of four subtests: Vocabulary, Grammar and Usage, Reading Comprehension, and Style and Verbal Logic. According to the needs of a pyramidal item order, reading comprehension items seemed to fit best.

In fact, however, it is not very easy to show what reading comprehension questions actually do test. Definitions are usually tautological: "Reading comprehension tests the ability to read and to understand a particular language." This definition, however, covers a multitude of aptitudes that have only been described very incompletely up to now. Some language and test experts (see Harris, 1969; Heaton, 1975; Lado, 1967; Pynsent, 1972) have tried to discover a few of the factors involved and to put them into a hierarchical order with regard to their level of difficulty and complexity. Obviously, at a more basic level reading comprehension requires the understanding of the meaning of words or word groups in the context in which they appear as well as the recognition of structural clues and the comprehension of structural patterns. These aspects of language are usually dealt with in tests of vocabulary and grammar—that is, the testee has to show his/her ability to ascertain the verbal meaning of a straightforward sentence or phrase. On an advanced level, reading comprehension involves higher mental abilities, such as how to comprehend paragraphs and to select the main ideas, how to draw conclusions from the text, and how to make inferences and to read between the lines. The level of reading comprehension that is actually tested depends to a certain extent on the item type that is used. For example:

Example 1

He asked me to ...... him two thick slices of beef.
(A) carve (B) slash (C) peel (D) split (E) shave
(Jackson, 1976, p. 171)

It is obvious that this question form does not put too great a demand on the testee´s reading comprehension abilities and can rather be looked upon as a vocabulary item. The testee has only to know that "carve" is the appropriate word for meat. He/she can answer this item correctly just on the basis of his/her knowledge of vocabulary. To a limited degree this item type can also test grammatical knowledge by offering choices/words that all seem to fit according to their meaning; but, in fact, only one fits for syntactical reasons. With this item type it is therefore very difficult to say to what extent reading comprehension is involved (cf. Jackson, 1976).

Item types that do not lay too much stress on the knowledge of particular words are more usual, and items consisting of a short reading extract of only a few sentences that ask the testee to interpret it in some way seem more appropriate.

Example 2

Parents can give their children enormous help so long as they don´t talk too much, give the game away, or block the children´s thought. "Come along, dear, we´re going to play with this lovely clay, let´s see what we can make with it. I think we can make a lovely elephant, come along, what about the trunk dear..." That poor child will have made a mental note that whatever he takes up as a career it won´t be sculpture.

Why is this child called "poor"?

(a) He is not allowed to work out his own ideas.
(b) He will never wish to become a sculptor.
(c) He has begun to dislike playing with clay.

(d) He is being taught skills for which he is too young.
(Sauter & Hornke, 1979, p. 165)

Example 2 shows clearly that it tests not only the testee's knowledge of syntactical structures and vocabulary but primarily his/her ability to interpret the text in some way, for the correct answer is not just a paraphrase of the item stem. This item type seems to be capable of testing what Carroll (1968) calls "complexity of information processing--at what level of complexity can the individual process linguistically-coded information?" (p.53)

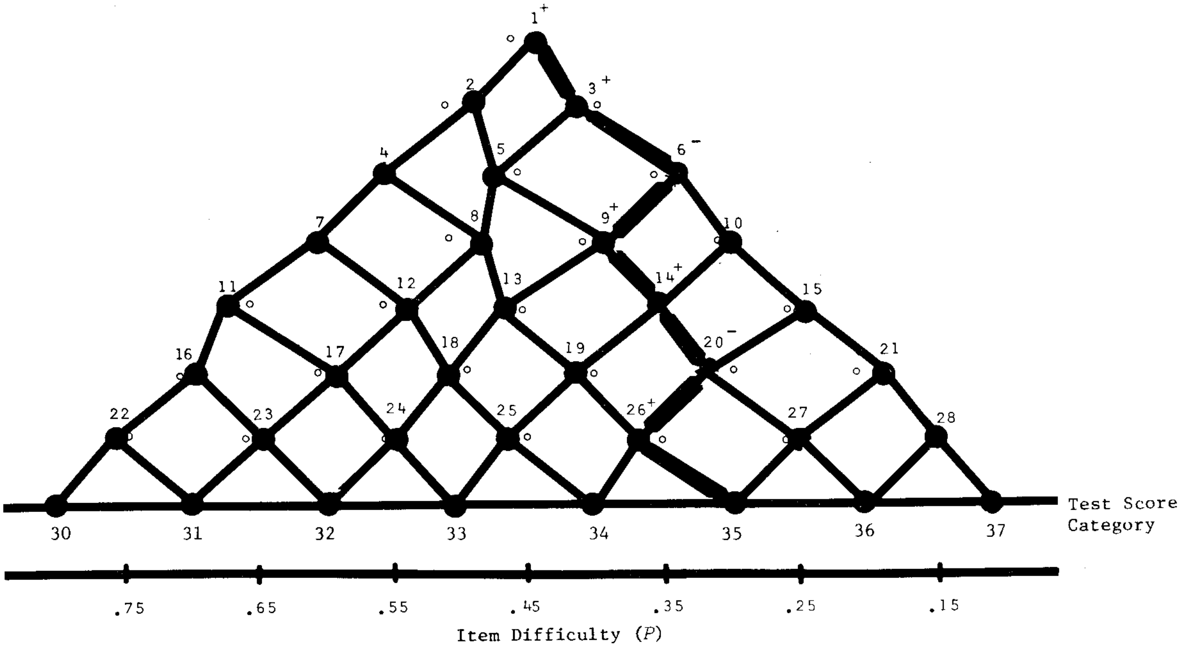
This should be the linguistic dimension that reading comprehension items test, at least in the adaptive test. In practice, however, it is very difficult to find items that represent this dimension even approximately. Even factor-analytic studies can give little help. Thus, it is inevitable that what were regarded as reading comprehension items in the above-mentioned sense do, in fact, correspond rather to a lower level of reading comprehension. The problem with any test construction is that this can cause some confusion, especially in the pyramidal item order by branching testees to incorrect items with regard to their own level of reading comprehension ability.

In this Cologne Placement Test (Bonheim & Kremfelts, 1979), reading comprehension items had been administered to an average of 750 students (up to a maximum of nearly 2,000 students) from 1974 until 1978. Since the placement test had been newly assembled at the beginning of each semester, proven as well as newly constructed items were used, and those items that did not turn out to be satisfactory were left out. The item pool finally contained 88 items from which items were systematically borrowed in order to construct the adaptive test. Each of the 88 items had been carefully analyzed to see whether it could be placed at a certain branching point within the pyramidal item order (see Figure 1). However, with the present state of knowledge, these decisions were not easily made, because there were neither guidelines nor previous experience for item selection that could guarantee a successful branching order. Additional problems that had to be solved were those of time limits and the positional effects of the items in the Cologne Placement Test (for a detailed description, see Hornke et al. 1979; Sauter, 1978; Sauter & Hornke, 1979).

Twenty-eight items were borrowed from the item pool in order to form a pyramidal test, which consisted of seven stages and extended to a difficulty level from P (Probability of a Correct Response) = .75 to P = .15. All items were placed on branching positions according to their empirical difficulty and discrimination. Figure 2 compares the ideal item order with the actual order that is based on the available item data. It shows only relatively small deviations from the positions on the ideal model.

The testee begins with a medium-difficult item (P_1 = .45) and is branched to a more difficult (P_3 = .40) or an easier item (P_2= .50), depending on whether he/she answered the preceding item correctly or incorrectly (see Figure 2). In this way, he/she is branched through the item order until he/she finally reaches his/her score group. He/she is given only one item at each stage, which eventually means that he/she has to work on only 7 out of 28 items. This seems to be reasonable, assuming that those items that are easier than the items he/she an-

Figure 2
Pyramidal Order of the Adaptive English Test with the Branching Path of a Hypothetical Testee



swered correctly are probably too easy for him/her. On the other hand, those items that are more difficult are supposedly too difficult for him/her; he/she would most probably answer them incorrectly (see Hornke, 1976, 1977). Thus, only those items are presented to the testee that are most suited for him/her using the pyramidal item order. With the test under consideration, the invisible ink response mode was used in a group setting.

Results of Two Empirical Investigations

Two adaptive reading comprehension tests were investigated--one in a pilot study by Sauter (1978) and the other in a larger validity study by Hornke et al. (1979). Both studies showed that adaptive testing by means of the paper-and-pencil version is quite feasible in group settings. Students had hardly any problems in following branching instructions properly by themselves.

Validity of the Pyramidal Item Order

The design of Sauter's (1978) study asked each student (1) to work through an item set of 28 items in the branching manner and (2) to solve all items left out during the branching in the conventional manner. This yielded two scores per person--an adaptive score and a conventional score, where the first was based on 7 items and the second on 21 residual items. Thus, complete response data were available on all items. This allowed the validity of the pyramidal item order to be investigated in some detail.

The results of an item analysis indicated that all 28 items had become easier than in the original conventional test. However, rank orders between previous and present item difficulties correlated as highly as *r* = .77, indicating that the order as such had largely survived. Of particular interest was the correlation between scores on the 7 adaptive items and the 21 conventional remaining items, which was *r* = .47 for 93 testees. Taking the unreliability of the entire set of items into account, however, a stepped-up correlation of *r* = .64 resulted. Thus, a score based on the 7 optional items had quite a reasonable predictive power to a score based on 21 items.

Validity of the Adaptive Test

The second study (Hornke et al., 1979) had two main purposes, namely, to investigate the validity of an adaptive test and to look at the details of pyramidal item hierarchies. In order to answer the first question, a multitrait approach was used. According to the underlying theory, reading comprehension items ought to call for processes that are different from vocabulary or grammar exercises. Thus, it was expected that there would be a closer relationship between scores for an adaptive and a conventional reading comprehension test than with scores from both grammar and vocabulary tests. The study used a two-method--Adaptive versus Conventional--by three-trait--Reading Comprehension (RC) × Grammer (G) × Vocabulary (V)--design. Due to financial restrictions, however, it was impossible to investigate adaptive *and* conventional test formats with all three traits. The study thus contrasted adaptive versus conventional reading comprehension only.

It is quite obvious that all three traits should correlate with each other because they are genuine parts of language behavior themselves. However, the results in Table 1 indicate that despite all that they have in common, the three item sets measured quite differentiable aspects that pertain to the hypothesized discriminant relation. This means, too, that the data warrant an interpretation of three different traits, even though intercorrelations were not zero (but they are low enough).

However, reading comprehension scores, assessed either in the adaptive or in the conventional way, did not converge to the extent expected. The resulting correlations were too low for tests designed to measure the same trait. The correlation between RC_1 (adaptive) and RC_1 (conventional remainder) especially contradicted any convergent interpretation, despite the fact that both item sets are virtual subsets of a larger one. Here, a correlation of .6 to .7 would be more suitable to justify any convergence. It still remains an open question whether adaptive branching of items used with reading comprehension tests introduced a source of error or variation that accounted for the low correlations. A comparison of RC_1 (conventional remainder) with the RC_2 (conventional) scores indicates some dissimilarity in the item sets, which appear to be more different than their common label would lead one to expect.

Conclusions

Although adaptive tests are initially intriguing, there are many problems to overcome. The major problem lies in the fact that for foreign language testing, a properly defined construct is necessary. Consequently, all items ought

Table 1
Correlations Between All Tests and Formats Used

Variable	RC$_1$		Conventional (28 Items)		
	Adaptive (7 Items)	Conventional Remainder (21 Items)	RC$_2$	G	V
Convergent					
RC$_1$					
Adaptive	-	.405	.379		
Conventional Remainder	.531	-	.419		
RC$_2$					
Conventional	.218	(.403)	-		
Discriminant					
G (Conventional)	.295	.511	(.068)	(.419)	
V (Conventional)	.355	.431	(.214)	-	

Note. Correlation coefficients in parentheses are based on group means instead of individual data.

to belong to an appropriately defined behavioral domain. This is not always easy to achieve, and there might often be a lack of expert consensus. Instead, empirical studies are needed to substantiate any item's relation to the construct in question.

A quite substantial problem for adaptive tests may be seen in the necessary heierarchical order for a pyramidal arrangement. Any branching decision here implies strongly that the hierarchy is valid and stable across samples of the population. The two studies cited above indicated, however, that this may not be the case. As far as there are changes in item difficulties from one sample to the other, this might not matter very much as long as all item positions stay within the hierarchical order intended. Whenever there are changes or shifts in positions, the pyramid is invalidated locally, and false branching occurs. To circumvent this problem, rigorous item analysis may help to keep this weakness within limits. It has to be questioned, too, whether difficulty indices (i.e., the proportions of answers correct) are good and reasonable criteria for a hierarchical ordering of items. With narrowly defined populations and applications, this might be practicable. However, better estimates of an item's scale and hierarchical position are available and should be used. With these two studies cited, it was not possible to perform item analyses, since data were not available for this purpose.

Taking these two arguments together, it follows immediately that there will be hardly any chance to take a conventional test, to rearrange its item order, and to get an adaptive version. With any test construction, careful item writing and analysis is necessary. This is true for adaptive as well as conventional tests; ad hoc test construction hardly conforms to the careful scrutiny that is called for. It should not be expected that adaptive or conventional tests from this source have any value in decision making at all. In foreign language testing only after a good deal of research and empirical investigation has been carried out will there be adaptive tests for a variety of purposes; but, in fact, they are essential in a program where students' proficiency is expected to vary considerably and where decisions of some kind are to be made.

REFERENCES

Bonheim, H., & Kreifelts, B. Ein universitätseingangstest für neuphilologen abschtussbericht der arbeitsgruppe sprachtests (as) an der Universität Köln zur Verlage beim BMBW. Köln: Universität Köln, 1979.

Carroll, J. B. The psychology of language testing. In A. Davies (Ed.), Language Testing Symposium. London: Oxford University Press, 1968.

Fischer, G. H. Probabilistic test models and their applications. German Journal of Psychology, 1978, 8, 298-319.

Harris, D. P. Testing English as a second language. New York: McGraw-Hill, 1969.

Heaton, J. B. Writing English language tests. London: Longman, 1975.

Hornke, L. F. Grundlagen und probleme antvortabhängiger testverfahren. Frankfurt: Haag & Herchen, 1976.

Hornke, L. F. Antwortabhängige testverfahren: Ein neuartiger ansatz psychologischen testens. Diagnostica, 1977, 23, 1-14.

Hornke, L. F. Four realisations of pyramidal adaptive testing. Programmed Learning and Educational Technology, 1979, 16, 164-169. (a)

Hornke, L. F. Konstruktion eines adaptiv-antwortabhängigen fragebogens zur erfassung von preufungsangst. Diagnostica, 1979, 25, 208-218. (b)

Hornke, L. F. Testdiagnostische untersuchungsstrategien. In K.-J. Groffman & L. Michel (Eds.), Handbuch der psychologischen diagnostik (Vol. 6, 2nd ed.). Göttingen: Hogrefe, 1979. (c)

Hornke, L. F., Sauter, M. F., Süssmilch, B. H., & Burghoff, U. R. Konvergente und diskriminante validität eines adaptiv-antwortabhängigen Englischtests fur Anglistikstudenten (DFG HO-758-1). Unpublished research report, Universität Düsseldorf, Erziebungswissenschaftliches Institut, 1979.

Jackson, S. H. Reading comprehension questions in tests of English as a foreign language. In Kongressberichte der 7. Jahrestagung der Gesellschaft für Angewandte Linguistik. Tier: GAL e.V., 1976.

Lado, R. Language testing. The construction and use of foreign language tests (5th ed.). London: Longmans, 1967.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, *8*, 147-151.

Michel, L. Allgemeine grundlagen psychometrischer tests. In R. Heiss, K.-J. Groffmann, & L. Michel (Eds.), *Psychologische diagnostik*. *Handbuch der psychologie* (Vol. 6). Göttingen: Hogrefe, 1964.

Popham, W. J. The case for criterion-referenced measurement. *Educational Researcher*, 1978, *7*, 6-10.

Pynsent, R. B. The objective reading comprehension test. In R. B. Pynsent (Ed.), *Objektive tests in Englishchunterricht der schule und universität*. Frankfurt: Athenaeum, 1972.

Sauter, M. P. *Entwicklung und erprobung eines antwortanhängigen testverfahrens zur überprüfung des leserverständnisses in Englisch*. Unpublished doctoral dissertation, Universität Düsseldorf, Erziehungswissenschaftliches Institut, 1978.

Sauter, M. P. Adaptive tests im Fremdsprachenunterricht. In *Kongressbericht der 9. Jahrestagung der Gesellschaft fur Angewandte Linguistik* (Vol. 3). Heidelberg: Julius Groos, 1979.

Sauter, M. P., & Hornke, L. F. Adaptives testen im Englischunterricht. Entwicklung eines flexiblen testverfahrens zur messung von leserverständnis. *Anglistik & Englischunterricht*, 1979, *8*, 151-166.

Weiss, D. J. *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

Weiss, D. J. *Computerized adaptive trait measurement*: *Problems and prospects* (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975. (NTIS No. AD A018675

Weiss, D. J. *Computerized ability testing 1972-1975 (Final Report)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1976. (NTIS No. AD A024516)

Weiss, D. J. *Proceedings of the 1977 Computerized Adaptive Test Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Weiss, D. J., & Betz, N. E. *Ability measurement*: *Conventional or adaptive?* (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)