Occasional Paper No. 8

TEST DESIGN: A VIEW FROM PRACTICE

Lee S. Shulman

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

June 1978

The work reported herein is sponsored by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Teaching Division of the National Institute of Education, United States Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-76-0073)

Institute for Research on Teaching

The **Institute for Research on Teaching** was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.

Abstract

Teachers tend to dislike, mistrust, and disbelieve tests and test design. The author hypothesizes that this is because the kinds of tests used are inconsistent with, and in many cases irrelevant to, the realities of teaching. And the realities of teaching are quite different from the logic of instruction. Another hypothesis is related to research on teacher planning. Various studies have shown that teachers do not focus on outcomes, goals, or objectives, but on activities and content. Tests do not reflect this. Test designers must focus on the realities of teaching if tests are to be improved.

Test Design: A View from Practice[1]

Lee S. Shulman[2]

I have spent the last 10 years studying two types of expertise -- in medicine and in teaching -- and attempting to understand the relationship between them. One of the things practitioners in these two fields share is a constant involvement with tests. Yet, in studying how, why, and with what degree of trust, confidence, and commitment practitioners in medicine and teaching employ tests, some striking and, at first glimpse, paradoxical contrasts are found.

Studies in Medicine and Teaching

We began in the late 1960s to study in depth the cognitive processes of experienced peer-nominated internists in order to understand how they perform their medical work. The many studies conducted and the array of findings are reported in Medical Problem Solving: The Analysis of Clinical Reasoning (Elstein, Shulman, & Sprafka, 1978). We tried to identify the characteristics of expert performance -- what an expert does. We wished to discover how to build both curriculum and evaluation in a way that corresponds to how experts actually perform rather than to the content of traditional medical lore regarding how physicians ought to perform.

---

[1]Paper delivered at the Winter Invitational Conference on Measurement and Methodology, UCLA Center for the Study of Evaluation, January 1978. A modified form of this paper will be published in the proceedings of that conference.

[2]Lee S. Shulman is director of the Institute for Research on Teaching and professor of educational psychology and medical education.

As we studied physicians, dealing with cases across different domains of internal medicine, we found that they exhibited non-generalizability of performance from one domain to another. We had assumed, without thinking in terms of formal domain specification and domain referencing, that internal medicine was a unitary domain. Just as items are generated to sample any other field, we felt we could generate cases to sample expertise in internal medicine. It did not much matter which cases we selected as long as each one, to use the terms we used in our research, was a high-fidelity representation of problem solving in that domain. We expected significant positive correlations among performance measures across cases. Instead, what we found was that for all practical purposes, performance in one case domain provided no basis for predicting performance by the same physician in another case domain. That finding had several consequences. It led us to collaborate and communicate with other research teams which had been studying similar processes, and we discovered that they were reporting similar findings. But, like small children in their early years of sexual development, each one thought it was something that only happened to them.

For example, Christine McGuire (McGuire & Babbott, 1967) and her group at the University of Illinois Medical School have for many years been developing patient-management problems in medical reasoning, medical diagnosis, and treatment. During this time they have been struggling with the problem that performances of physicians and medical students do not correlate highly across cases. For years, they treated this as a measurement problem, a problem of unreliability. Something had to be fundamentally wrong with their measurement procedures; had they written good simulations, they believed, the cases would have had the same

inter-case characteristics as items on a test have with one another.

Another example was the American Board of Internal Medicine, which, after many years, terminated use of its oral examination. This was an examination conducted by taking candidate physicians to the bedsides of patients and actually conducting oral examinations across the bed. The candidates were observed as they examined the patient, and were themselves examined on their understanding of the case -- a process only slightly less mortifying for the patient than for the candidate.

The major reason the board abandoned oral examinations was the incidence of unreliability of ratings among the judges. The board found that the identity of the oral examiner accounted for a higher proportion of the variance in outcome than did any other single aspect of the examination or candidate. It was only through later research conducted on the computer-based examination for internal medicine by the American Board of Internal Medicine that a group from the Oregon Research Institute led by Paul Hoffman and Robyn Dawes (Hoffman, 1974) tried to disentangle problems of tester unreliability from problems of case or domain specificity. In all those years of oral examination, the examiner had been confounded with the case. A different examiner conducted the examination for each case, since everyone assumed that the domain was singular and all cases were sampled from it. Clearly then, differences among judges could be attributed only to differences in judgment.

When Hoffman and Dawes (Hoffman, 1974) disentangled these two factors, they found that judges given the opportunity to observe a candidate on the same case had an acceptably high inter-rater reliability of judgment. The performers themselves -- the experienced board-eligible internists -- varied dramatically and significantly from case to case and were the reason for the apparently inconsistent judgments.

These three examples, from McGuire, from Dawes and Hoffman, and from our own research, illustrate how we can totally misconstrue the nature of the domain of expertise we are attempting to assess merely by making unwarranted assumptions about the domain or universe, rather that doing systematic studies of how that expertise is, in fact, put to use. I believe that more must be known about the topography of expertise in different areas -- how that expertise is arranged and organized. Expertise in different domains is likely to be cognitively organized in different ways, therefore requiring very different strategies of domain specification, universe designation, and hence, test design.[3]

An example of the curricular consequence of misconstruing a domain through misunderstanding the development of the expertise can be found in the classical medical school curricula inspired by Abraham Flexner, one of the most influential persons in the history of twentieth century medical education.[4]

In the "Flexner curriculum" medical students refrained from clinical work until they had spent two full years studying basic biological sciences. It was asserted that students needed a broad, solid foundational base in all of the relevant biological sciences before they could profit from clinical work.

---

[3]As a footnote to Glaser's (Note 1) observations on studying the development of expertise, a group that has been closely associated with ours (Barrows, Feightner, Neufeld, & Norman, Note 2) at the McMaster School of Medicine in Hamilton, Ontario, has been studying novices, more experienced medical students, and experts on the same cases. The group has been studying the students longitudinally as they go through medical school to obtain data on the development of medical expertise.

[4]Ironically, Flexner was neither a physician nor a Ph.D. He was a former school principal whose Flexner Report (1910) revolutionized the teaching of American medicine.

As we began to see the results of our own research on medical expertise, and the domain specificity of that expertise, we began not only to re-examine our assumptions regarding evaluation of performance in medicine, but also our assumptions about how curriculum should be organized to produce that expertise. If the expertise was domain specific, then perhaps students would not have to know all of biochemistry, physiology, anatomy, microbiology, or pharmacology before doing clinical work. Rather, instruction could be vertically organized and coordinated to permit the beginning of clinical work much earlier. For any *particular* domain of clinical work, there is only a limited area of prerequisite understanding. We designed curricula at Michigan State that reflected this idea, where medical students began doing clinical work in the first *week* of their first term in medical school. To the amazement of our own faculty, the students did very well.

Implications for Test Design

Certain parallels can be drawn between our medical studies and research by Berliner (Note 3), Wiley (Note 4), and others on the relationships among time expended on instructional tasks, content covered, and achievement. In describing this research, Rosenshine (1976) observed that when pupils have fallen (in terms of standardized achievement tests) three years below grade level, instructional time somehow must be found to make up for that difference. In one sense, it requires three years of time to make up for three years of cumulative deficit. Through better quality instruction, of course, it may be possible to compress the calendar time, but the pupils must somehow traverse the content that has to be covered in order to reach grade level. Where will that instructional time come from, if not via

special summer programs? It can only be taken away from the teaching of science, mathematics, social studies, art, or music.

How does our earlier work in medicine correspond to the problem of teaching children who have fallen behind? The question to ask might be: What if the expertise represented in the contents of third- and fourth-grade performance tests, which we simply assume is prerequisite to fifth-grade performance, is not, to use one of Glaser's (Note 1) concepts, really the way the second-grade novice becomes a fifth-grade expert? Indeed, perhaps that sequence of content is merely a curricular convention -- the sequence which instruction has always followed -- so is assumed to be the best way. I am suggesting that this is one of the kinds of implicit assumptions about the relationships between types or levels of achievement that very often condition test design and ought to be subjected to scrutiny. One way to accomplish this is through the developmental studies of expertise that Glaser (Note 1) has described.

A potentially dangerous vertical monopoly, one that I am increasingly uncomfortable with, has developed in the education industry: The same companies are producing both the standard curriculum materials and the standardized tests. Therefore, what constitutes average expected performance at a given grade level in a subject area by some remarkable coincidence corresponds to what the curriculum makers have chosen to define as the content of that grade level. I wonder if we are confronting a problem that is a consequence of how the process of education has become organized both politically and entrepreneurially. I trust that the work of Porter, Schmidt, Floden, Freeman and their colleagues (Note 5) at the Institute for Research on Teaching may help us understand more fully the links between curriculum content and testing.

Diagnostic Expertise and Test Design

At the Institute for Research on Teaching (IRT), we study the *expert* as a basis for sensibly defining domains for our work. We are doing studies similar to those we conducted on medical diagnosis. Diagnostic work in the areas of reading and learning disabilities is being done in an attempt to understand how the diagnostic process works in these areas (Vinsonhaler, Wagner,& Elstein, Note 6). Our computer simulations of diagnostic work closely correspond to the characteristics of expertise that Glaser (Note 1) has described. We find expertise composed of two components -- clinical strategy and clinical memory. Subdividing clinical memory into its components as we do in the simulations, our formulations are also very similar to Glaser's.

How do the experts feel about tests and test design? Referring again to our studies of medical expertise, it is easy to recognize how starkly contrasting test use and attitudes about tests are between the fields of medicine and teaching. The Center for the Study of Evaluation (CSE) has conducted surveys to study teacher use of tests, observing that when teachers are asked about tests, they tend to dislike, mistrust, and disbelieve them.

The IRT sponsored a set of studies in San Jose by Greta Morine-Dershimer (Note 7) and Bruce Joyce (Note 8; Note 9) which involved intensive study of 10 teachers as part of a Teacher Corps experience. Morine-Dershimer and Joyce observed the reactions of the teachers when a set of domain-referenced diagnostic tests that the state had mandated was returned for each of the classroom teachers' use. Performance of each pupil was keyed to each objective and, if pupils were low, the print-out specified what kind of curriculum materials could be used to remediate the deficiency. The investigators waited until two weeks after the tests had come back to interview the teachers because they wanted to study

how teachers' conceptions of their pupils had changed since the beginning of the year, especially after this marvelous new set of information had arrived. It turned out, however, that not a single one of the 10 teachers had looked at those test results. They simply did not find them useful. They were convinced that they already knew more about their students than any one of those tests could possibly detect. Most of the teachers did not believe the tests were of any value at all.

Such observations are striking when compared to the situation in medicine, where a major problem is overcommitment to tests. Probably the single factor that is contributing most to increasing the cost of medical care is that clinicians typically order more tests than they need for a diagnosis. A study by Oskamp (1965) showed that if clinicians are provided with a few tests results, asked for their diagnostic assessment and treatment plan, and then given further increments of test information, the clinicians' diagnostic accuracy and the quality of their treatment plan, after a fairly small number of tests, reaches asymptote. They are doing as well as they ever will long before they stop ordering tests. Although diagnostic accuracy levels off, it can be seen that a clinician's *confidence* increases as s/he orders more tests. This procedure has been replicated by Slovic (personal communication, Note 10) with horserace handicappers, who, given five or six pieces of information, are about as accurate as they can be in predicting the outcome of horseraces, but continue asking for more information because they seem to feel better that way. More recently, and most disturbingly, Sisson, Schoomaker and Ross (1976) have found that additional increments of information can reduce the quality of medical decisions. More information is not without its own dangers.

Tests Are Not Enough

Given such abiding faith in tests among other professionals, why is it that teachers do not use, trust, or like tests very much? It matters not whether the test is domain or norm-referenced,[5] though for a long time we argued that teachers make little use of tests because norm-referenced tests cannot tell teachers what they want to know. That does not seem to be the answer. I would like to suggest several hypotheses to account for the contrast between the professions. I reject what may be an obvious explanation, which is that tests in medicine are far more reliable. They are not. There have been studies on the reliability of laboratory tests in medicine whose results are frightening. Medicine has serious reliability problems. Educators, in contrast, have the most reliable tests in the world. Thus, reliability is not the problem.

My hypotheses are the following: In general, the kinds of tests we use are inconsistent with, and in many cases irrelevant to, *the realities of teaching.* And the realities of teaching are quite different from *the logic of instruction.*

First, it is clear that no physician ever treats a laboratory test as self-standing, as being sufficient by itself. No physician would ever conduct an inquiry about a patient using only tests, expecting the test to provide all the information necessary to make the diagnostic classification or general assessment judgment. Medical practitioners know that the test is only part of the assessment procedure. *Clinical tests cannot be made*

[5]The terms *domain* or *criterion-referenced* tests and *norm-referenced* tests have taken on a variety of meanings. In this context, I define criterion-referenced tests as those in which an individual is assessed relative to a certain standard, whereas norm-referenced tests assess the individual's performance relative to other individuals or to a group average.

clinician-proof! Yet our strategy in educational test design and development has been to attempt to design tests that are *sufficient* portrayers of student performance, not subject to frequent modification or rescaling by the teachers who spend months at a time with the examined pupils, instructing, observing, and interacting with them. Nevertheless, we must remember that the test is but a very small behavior sample, extremely well calibrated but contextually restricted.

The physician uses observation, interview, touching and feeling, as well as testing, and develops an assessment and a plan by aggregating across those sources of information rather than by giving almost total weight to any one source and subordinating the others to it. A similar strategy might work better in the field of educational test design. Rather than attempting to develop tests that assess everything that is relevant to student performance in the classroom (bemoaning the fact that moving higher up in Bloom's (1956) cognitive taxonomy, or from cognition to affect, or from more to less advantaged youngsters, the tests don't do quite as well), why not treat the test as only one part of the assessment procedure and begin working on ways of helping teachers document in a better calibrated manner the other observations which they make so frequently and richly in the classroom? In that way, assessment need not depend solely on tests, but will be better informed by tests.

Another hypothesis for why teachers do not use or trust tests is related to work conducted on teacher planning (Zahorik, 1975; Yinger, Note 11; Clark & Yinger, Note 12; Peterson, Marx, & Clark, 1978). This work, replicated and extended in other areas, has shown that teachers *do not* focus on outcomes. They *do not* focus on objectives. They *do not* focus on goals, though the goals are probably there implicitly as dis-

cussed by Wiley (Note 4) in his suggestion that curriculum affect goals. Teachers focus on *activities* and *content*. Their attention is directed to questions *what will we do* and *what will we cover.*

For years those of us in educational research, especially in evaluation and measurement, have been insisting that teachers learn to think straight educationally. By that we meant they have to learn to think of outcomes stated in terms of behavioral objectives. However, if generations of practitioners do *not* think in such a way, an alternative consideration might be that there is something adaptive in focusing instead on activities and content covered. Teachers appear not to evaluate their day-to-day activity in terms of general assessments of achieved outcomes, but rather attend to variations in *student involvement*. When we ask teachers, "What did you achieve today?", they are inclined to say, "Well, we covered three more pages of math, and the kids were really involved." We then become critical and berate teachers for not thinking in terms of objectives -- which ones they achieved and which not. I believe we have to treat the teachers' observations as data rather than as sources for blame. That is how teachers evaluate what they do. When they plan their instruction, they plan for such things as grouping, pacing, and involvement.

Barr and Dreeben (1978) have pointed out that in areas like mastery learning, where we prescribe instruction, diagnostic tests are used to help pace that instruction. Teachers in classrooms without mastery learning technology, as reported by Dahloff (1971) and Lundgren (1972) in Sweden, appear to use "steering groups." They attend to particular subgroups of pupils in the classroom to detect cues to help them decide whether to speed up, slow down, reiterate, or change topics. If that is one of the decisions teachers are most concerned about, but our tests are not measuring anything relevant to it, then should we be at all surprised that the practitioner

finds our tests of little use?

One possibility may be that what the standardized tests measure and what teachers are evaluating are really two parallel and somewhat independent systems. Being in a classroom is similar to being on a cross-country train. The high correlation between pretest and post-test achievement scores over an entire year suggests that pupils are on a trajectory and there is little chance that instruction can produce great changes in that trajectory. If such is the case, the day-to-day events that teachers can and must monitor need not be a very good indicant of the shape of that trajectory. The job of the teacher is to deal with life *on* the train and make it the most involving and most meaningful in a day-to-day sense. The teacher's first priority would not be to accelerate dramatically the progress of the train into the station. In the field of test design, we seem to have focused our efforts on providing data most useful from the perspective of those who are scheduling the trains, rather than from the perspective of the people who are conducting them.

Yet, are teachers likely to be totally oblivious to student progress in academic achievement? In the recent California BTES studies of teaching, researchers have sought the best possible process variable to predict achievement. They have concluded that the best proxy for achievement is *academic learning time*, a combination of *content covered* and degree of *engagement* or *involvement*. I find it ironic that the best available predictor of ultimate student performance is a combination of precisely those two indicators that teachers already intuitively use to monitor student progress. Our researchers and test developers have plenty to learn from the wisdom of practitioners.

Conclusion

The principles of test design have typically led to our selecting items for tests in a manner that throws out or rewrites items if they appear unstable. If they seem to fluctuate in important ways from day to day or week to week, we try to develop tests that lessen those fluctuations. But the job of the teacher is to deal with the fluctuations, to make sense of them, and to make the environment in which those fluctuations constantly occur sensible and educative. It is conceivable that in designing tests that are relatively immune to the variations in experience and response that characterize pupils during the course of an instructional experience, we throw out precisely those sorts of test items that the teacher might dearly love to use. We may have here another example of Cronbach's (1957) "two-disciplines" paradox. Cronbach observed that what was error variance for the correlationists was true-score variance for the experimentalists and vice versa. It may be that the test designers' error is the teachers' true-score in the day-to-day workings of the classroom. Unless designers begin to focus on the realities of teaching, the ways in which teachers do their work in classrooms, they are fated to learn 20 years from now, in yet another CSE study, that teachers continue to dislike, distrust, and feel uncomfortable with tests.

Those who would design educational tests must see themselves as inhabiting the interstices between two domains of expertise. Expertise of the first kind is the achievement of pupils who are being educated in the system, which is the subject of Glaser's (Note 1) research. Expertise of the second kind is the expertise of pedagogues -- the teachers who must monitor, make sense of, and guide the development of that first expertise. All of us in our first psychology course were taught

as a proposition that learning is an internal process in the learner, unobservable, which we infer from indices of learner behavior. If learning is an inferred process, then teaching is a profession rooted in the ability to use cues of various kinds to make inferences about that process in order to guide learning. The well-designed test can be an extraordinary tool to inform the expertise of that judge-- the teacher -- in making necessary inferences about the developing expertise of the learner. So far, tests have not fulfilled that promise. I suggest that this might be part of a new agenda for the field of test design.

Reference Notes

1. Glaser, R. The measurement of expertise. Paper presented at the Winter Invitational Conference on Measurement and Methodology, UCLA Center for the Study of Evaluation, Los Angeles, January 1978.

2. Barrows, H.S., Feightner, J.W., Neufeld, V.R., & Norman, G.R. Analysis of the clinical methods of medical students and physicians (A report submitted to the Province of Ontario Department of Health and Physician's Services Inc. Foundation). Hamilton, Ontario, Canada: McMaster University, March 1978.

3. Berliner, D.C., & Rosenshine, B. The acquisition of knowledge in the classroom. In Beginning Teacher Evaluation Study (Technical Report Series). San Francisco, California: Far West Laboratory for Educational Research, February 1976.

4. Wiley, D.E. Policy-responsive evaluation. Paper presented at the Winter Invitational Conference on Measurement and Methodology, UCLA Center for the Study of Evaluation, Los Angeles, January 1978.

5. Porter, A.C., Schmidt, W.H., Floden, R.E., & Freeman, D.J. Impact on what? The importance of content covered (Res. Ser. No. 2). East Lansing: Institute for Research on Teaching, Michigan State University, 1978.

6. Vinsonhaler, J.F., Wagner, C.C., & Elstein, A.S. The Inquiry Theory: An information-processing approach to clinical problem-solving research and application (Res. Ser. No. 1). East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, 1978.

7. Morine-Dershimer, G. Teacher conceptions of pupils. East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, in press.

8. Joyce, B. Teachers' thoughts while teaching. East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, in press.

9. Joyce, B. The teaching styles at South Bay School. East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, in press.

10. *Slovic*, P. Personal communication, 1975.

11. Yinger, R.J. A study of teacher planning: Description and theory development using ethnographic and information processing methods. Unpublished doctoral dissertation, Michigan State University, 1977. (Also available as Res. Ser. No. 18 from the Institute for Research on Teaching, Michigan State University, East Lansing, Michigan.)

12. Clark, C., & Yinger, R.J. Research on teacher thinking (Res. Ser. No. 12). East Lansing, Michigan: Institute for Research on Teaching, Michigan State University, April 1978.

References

Barr, R., & Dreeben, R. Instruction in classrooms. In L.S. Shulman (Ed.), Review of research in education (Vol. 5). Itasca, Ill.: F.E. Peacock Publishers Inc., 1978.

Bloom, B.S. (Ed.). Taxonomy of educational objectives: The classification of educational goals. Handbook 1 - - cognitive domain. New York, David McKay, 1956.

Cronbach, L.J. The two disciplines of scientific psychology. American Psychologist, 1957, 21, 11.

Dahloff, U.S. Ability grouping, content validity, and curriculum process analysis. New York: Teachers College Press, 1971.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. Medical problem solving: The analysis of clinical reasoning. Cambridge: Harvard University Press, 1978.

Flexner, A. Medical education in the United States and Canada (Bulletin No. 4). New York: Carnegie Foundation for the Advancement of Teaching, 1910.

Hoffman, P. Physicians appraise other physicians: Improving the decisions of a medical specialty board. Oregon Research Institute, Research Bulletin, 1974, 14 (4).

Lundgren, U.P. Frame factors and the teaching process: A contribution to curriculum theory and theory on teaching. Stockholm: Almquist & Wiksell, 1972.

McGuire, C., & Babbott, D. Simulation technique in the measurement of problem solving skills. Journal of Educational Measurement, 1967, 4, 1-10.

Oskamp, S. Overconfidence in case study judgments. Journal of Consulting Psychology, 1965, 29, 261-265.

Peterson, P.L., Marx, R.W., & Clark, C.M. Teaching planning, teacher behavior, and student achievement. American Educational Research Journal, 1978, 15(3), 417-432.

Rosenshine, B. Classroom instruction. In N.L. Gage, (Ed.), The psychology of teaching methods: Seventy-fifth yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1976.

Sisson, J. Schoomaker, R., & Ross, J. Clinical decision analyses -- the hazard of using additional data. Journal of the American Medical Association, 1976, 236, 1259-1263.

Wiley, D.E. Policy-responsive evaluation. Paper presented at the Winter Invitational Conference on Measurement and Methodology, UCLA Center for the Study of Evaluation, Los Angeles, January 1978.

Zahorik, J.A. Teachers' planning models. Educational Leadership, 1975, 33(2), 134-139.