INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

COGNITIVE LEVELS OF MULTIPLE-CHOICE ITEMS ON TEACHER-MADE TESTS IN NURSING EDUCATION

by

KATHLEEN CROSS

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2000

MAJOR: EDUCATIONAL EVALUATION AND RESEARCH (EER)

Approved by:

Advisor Date

Acknowledgements

In completing this study, I have been very fortunate to have had assistance from many sources. I was particularly fortunate to have received small research grants from the Wayne State University Graduate School and from Kappa Iota Chapter of Sigma Theta Tau International.

I have been blessed with a very fine doctoral committe chaired by Donald R. Marcotte, Ph.D. Marilyn H. Oermann, Ph.D., and Shlomo Sawilowsky, Ph.D., have served for more than three years as I have worked on this project, often making progress very slowly. Arthur C. Brown, Ph.D., was part of the original committee until his untimely death in December 1997. I would like to thank Roger DeMont, Ph.D., for being kind enough to join the committee in Dr. Brown's place.

I truly could not have completed this project without the support and work of my colleagues at the College of Nursing and Health at Madonna University. Dean Mary S. Wawrzynski, Ph.D., willingly lent her name (and hand signed 150 letters) in order to help me obtain exams from nursing programs around the country. In addition, she has been a supportive presence throughout my many years of graduate study. Two of my colleagues, Martha A. Donagrandi, M.S.N., and Deborah Dunn, M.S.N., G.N.P., were the raters for the pilot study who helped refine the system for rating cognitive levels. Both of them also rated test items for the main study. I am truly thankful to be part of a faculty where many other colleagues willingly acted as raters for my study. The following persons each volunteered and spent many hours rating exam items: Mildred Braunstein, Ph.D., Patricia Carlson, M.A., M.S.N., Margaret Comstock, M.S., Kathleen Walsh Esper, M.S.N.,

Marilyn K. Harton, M.S., Fran Jurcak, M.S.N., Nancy Kostin, M.S.N., Maureen Gallagher Leen, Ph.D., Gail Lis, M.S.N., Karen Marold, M.S.N., Mary Mitsch, M.S.N., Joycelyn Montney, M.S., and Sandra Wahtera, Ph. D..

Last, but not least, I would like to recognize the efforts of my husband, James W. Cross, who has not only been a first-class computer consultant, but has listened to my frustrations and triumphs throughout the long time period.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

"Will this be on the test?" Spoken or not, this question exerts a profound influence on students' learning. Students draw inferences as to what the faculty considers important from examinations. Performance on classroom tests may have far-reaching consequences for students' academic and professional careers. In spite of this, little attention has been given to the tests students take as part of their academic programs.

There is little question that the style of testing influences student learning (Balch, 1964; Black, 1980; Crooks, 1988; Stiggins, 1988; Linn, 1990; Talbot, 1994). Students quickly learn what type of questions will be asked (Black, 1980) and often adjust their mode of study accordingly. Students' primary goal is often to perform well on examinations and other evaluation measures. This is sometimes perceived as actually conflicting with obtaining a deep understanding of the content. The reality is that many students focus on the demands of the evaluation system instead of mastering essential content (Crooks, p. 445).

Stiggins (1988) identified the "central issue in school assessment [to be] the quality and appropriate use of teacher-directed assessments of student achievement used every day in classrooms" (p. 363). These assessments, including but not limited to teacher-made tests, are the ones "that count" (p. 368). In spite of the recognized importance of teachers' classroom measurements, most research has focused on standardized tests and little is known about the preponderance of school assessment used by teachers in the classroom (Stiggins, Conklin, & Bridgeford, 1986, p. 6). Although there is some variation between educational levels and subject matter, most teachers use and prefer their own

assessments (Stiggins, 1988; Stiggins, et al.,1986; MacCuish, 1986; McMorris & Boothroyd, 1993). Some teachers also use, sometimes uncritically, commercial products that are included with instructional materials (Stiggins, 1988, p. 364).

Although tests are used for instructional and diagnostic purposes as well as for assigning grades, Oescher and Kirby (1990) found that secondary teachers perceived assignment of grades to be the primary purpose for testing. Classroom test scores provide the basis for a variety of educational decisions (McMorris & Boothroyd, 1992; Oescher & Kirby, 1990). For nursing students, success or failure in completing the program, and entering the profession, may depend heavily on written tests (Flynn & Reese, 1988, p. 61). Individuals who achieve passing grades on written examinations in nursing programs, and later on the standardized licensure examination (NCLEX-RN), are allowed to practice nursing.

In developing courses and curricula, nursing faculty formulate learning objectives that include higher cognitive levels. When course objectives call for the use and development of higher level thinking skills, the means for evaluating the achievement of these goals must reflect the same higher cognitive levels. Unfortunately, available studies suggest that many teachers are unable to recognize and develop higher-level test items (Carter, 1984). In Oescher and Kirby's (1990) study, secondary teachers reported that one-fourth of test items were written at the application, analysis, synthesis or evaluation levels. Researchers' examination of these tests revealed less than 8% of items at these levels with virtually none at the synthesis or evaluation levels. This suggests that teachers may not accurately judge the cognitive levels of their own test items.

Critical thinking and higher levels of teaching and learning have been emphasized in nursing education for several years. Indeed, critical thinking is one of the standard outcomes that must be demonstrated to attain and maintain National League for Nursing (NLN) accreditation for nursing programs. Manuel and Sorenson (1995) surveyed agencies employing baccalaureate nurses in Massachusetts and recommended that curricula be redesigned to emphasize critical thinking and independent decision-making incrementally throughout nursing programs. When higher levels of thinking are to be emphasized in teaching, evaluation measures must be appropriate for measuring these levels (Oermann & Gaberson, 1998).

Almost all registered nurses are prepared in either associate degree nursing (ADN) programs or baccalaureate degree (BSN) programs. Graduates of both types of programs take the same licensure examination which demonstrates a minimum level of competence. The associate degree nurse is generally considered to be a technical or bedside nurse, while those with baccalaureate degrees are more likely to advance to other nursing roles. Employers surveyed by Manuel and Sorenson (1995) perceived that nurses with baccalaureate degrees had "better problem solving ability and have a broader based education and a more global perspective" (p.251). BSN graduates were assigned to leadership positions significantly more often, although they were not given more complex patient care assignments than nurses without the baccalaureate degree. Since baccalaureate nursing programs require greater depth in the sciences and a broader general education, it would be expected that these nursing programs will aim for higher levels of learning and, consequently, use evaluation instruments that reflect measurement of these same levels.

Purposes

The purposes of this study were:

1) to examine the current status of teacher-made tests used in nursing education, including the cognitive levels at which test questions are written;

2) to determine if there were incremental differences between cognitive levels of test items used in beginning nursing courses as compared to those used in courses taught later in nursing programs;

3) to determine whether there were significant differences in the cognitive levels of test items used in associate degree nursing (ADN) programs as compared to baccalaureate degree (BSN) programs.

Research Questions

1) What are prevailing practices for teacher-made exams related to appearance, format, and numbers and types of items?

2) Is there a relationship between numbers of students in a course and the use of essay and/or short answer items on exams?

3) What proportion of items on teacher-made tests in nursing are written at each cognitive level?

4) Are there systematic differences in cognitive levels of test items between those written for nursing courses early in programs as compared to those later in the program sequence?

5) Are there systematic differences in cognitive levels of test items used on teacher-made nursing examinations between ADN and BSN programs?

6) Are there differences in cognitive levels of test items related to the length of the exam (total number of items)?

7) To what extent can faculty who are unfamiliar with the source of the exam reliably classify items for cognitive levels?

8) Does inter-rater reliability vary with increasing or decreasing cognitive levels of items?

9) Does inter-rater reliability vary with the number of items (length of the exam)?

Assumptions

In conducting this study, the following assumptions were made:

1) Nursing educators aim to teach, and have students learn at, higher cognitive levels. Therefore, written tests in nursing courses should include items that test at these higher cognitive levels.

2) It was assumed that students have been exposed to the essential content included in the examination through regular classroom or laboratory sessions, assigned readings, or by other instructional methods.

3) Since the literature suggests that teachers at all levels prepare their own tests (Gullickson, 1984; MacCuish, 1985), it was assumed that the tests submitted for the purposes of this study were written by the faculty for their courses. When other sources, such as published or computerized test banks were used, the selection of items from these sources represented deliberate choices on the part of the faculty.

Limitations

1) The course objectives were not examined for congruency with the cognitive level of the test questions. Similarly, there was no knowledge of how content was taught

in the classroom or through other methods, which teaching methods were used, or the clinical experiences the students have completed.

2) Because nursing faculty, given a choice of any examination, would be likely to submit examinations they believe to be their best, the final (or last) examination of the course was requested. Final exams are typically longer, more comprehensive, and more heavily weighted for grading than other exams.

3) Examinations submitted voluntarily for analysis may be of higher quality than most of the exams being actually administered in schools of nursing. Voluntary submitters of exams are likely to be those who feel the most confident about their test-writing skills.

Definitions of terms

Nursing program: An associate degree in nursing (ADN) or bachelor of science in nursing (BSN) program accredited by the National League for Nursing Accrediting Commission (NLNAC) and listed in Directory of Accredited Nursing Programs (1997) whose graduates are eligible to take the professional nurse licensure exam (NCLEX-RN). Nursing diploma programs are not included because they are currently few in number. Institutions that house both ADN and BSN programs and those that do not offer initial professional preparation will also be excluded.

Examination/ exam/ test: A complete written measurement instrument authored by nursing faculty and taken by individual students to determine the level of meeting course objectives. Exams designed to be completed at home are excluded.

Appearance of test: a group of six format or style considerations relative to the test document. This includes overall legibility, directions to the test-taker, indications of point

values, consecutive numbering of pages and items, presence of typographical or usage errors, and inclusions of non-questions.

Item: A question or statement that calls for a response from the test taker. Each required response is considered an item, regardless of the author's manner of numbering.

Multiple-choice item: An item consisting of a stem followed by three or more response options (distractors) from which the test-taker is to select the correct/best one.

Matching item: One of a group of items that require the test-taker to relate an entry from one group (or column) to a corresponding response from a second group (or column). Each required response is considered an "item" regardless of the the author's manner of numbering.

Mathematical items: Are numerical items that require the student to calculate an answer regardless of the format of the item. Mathematical items may be presented with multiple response options (in multiple-choice format) or require the test-taker to enter a number in response (short answer format).

True/false item: Requires the test-taker to indicate whether a given statement is true or false. Items written in a multiple-choice format with two options ("true" vs. "false") are included.

Short answer item: An item that calls for the test-taker to formulate a written response of one sentence or less. These may consist of filling in a blank or writing a short free-form response. Items that require more than one sentence in response are classified as essay items.

Essay item: An item that requires the test-taker to formulate a response longer than one sentence. Each major response is considered an item. When test-takers are

required to answer only some of the items (e.g.- answer three out of four), the larger number is considered the number of items.

Non-question: An item that asks for a response from the test-taker, but is unrelated to the subject matter for the exam.

Cognitive level: One of three levels of classification for exam items based on the type of knowledge or understanding required to correctly respond to the item. (For more complete description of each level, see Appendix C.)

Mean cognitive index: The mean of the cognitive levels of all multiple-choice items included on an exam. The possible range is from 1.000 to 3.000.

Chapter 2

Literature Review

Many experts have described characteristics for valid and reliable teacher-made tests and have made recommendations related to teachers' preparation in evaluation and measurement. Although standardized tests have received a great deal of attention by developers and researchers, few studies have examined the tests teachers devise for their own classes. Studies of testing knowledge and practices of elementary and secondary teachers comprise most of the literature. The testing expertise and practices of post-secondary faculty, including nursing faculty, have been virtually ignored. Have faculty been taught test construction and analysis techniques? If so, do their tests reflect this knowledge and skill? Are faculty testing for simple recall of information, or are higher cognitive levels being evaluated?

Criteria for Test Construction

Criteria for judging the quality of teacher-made tests include general considerations (criteria for the overall test) and considerations related to individual test items. Principles for construction of teacher-made tests have been discussed in any of a number of measurements texts, papers, and articles (Frisbie,1983; Carter, 1986; Carey, 1988; Ellsworth, Dunnell, & Duell, 1990; Reilly & Oermann, 1990; Oermann & Gaberson, 1998).

In one of the few articles addressing teacher-made tests in higher education, Anderson (1987) affirms that teachers' tests "define instructional purposes and instructors' expectations, profoundly influence what students study, and help instructors to gain perspective on their courses" (p.40). College professors, having limited knowledge about

test development and limited time for test-writing, often do not write effective tests, often not focusing on the most important concepts and frequently failing to challenge students intellectually (pp. 40-41). Anderson suggested that review of an instructor's tests should be part of evaluating college teaching and recommended that faculty should be called upon to defend the quality of the tests they administer.

The nursing education literature contains similiar advice related to test construction and analysis with some adjustment for addressing the special concerns of nursing faculty. Guidelines for test construction and analysis of test results have been discussed by Stanton (1983), Van Ort and Hazzard (1985), Cassidy (1987), Flynn and Reese (1988), Chenevey (1988), Farley (1989, 1990), Reilly and Oermann (1990), Gaberson (1996), and Oermann and Gaberson (1998). Some sources stress construction of higher level questions or matching the level of test questions to the level of the course objectives (Demetrulias & McCubbin, 1982; Frisbie, 1983; Reilly & Oermann, 1990; Joachim, 1992). Due to recruitment of increasing numbers of nursing students from diverse cultural groups, additional concerns relate to avoiding language and cultural bias in nursing exams (Klisch, 1994).

There are no published studies of actual teacher-written exams in schools of nursing. Joachim (1992), a Canadian nurse educator, described a project in which a multiple-choice final examination was constructed using a computerized database. In beginning this process and studying multiple-choice examinations and item analysis, she and her colleagues found that course blueprints and tables of specifications were essential in order to provide consistent planning, teaching, and testing measures. The development of the testing system required careful advance planning and increased computer skills

among the faculty. A standardized method for preparing exams was developed at Joachim's institution which was judged beneficial for both students and faculty. The writing of test questions became an ongoing activity, not a rush to prepare questions under the pressure of an imminent deadline which Anderson (1987) suggested was the norm in higher education.

Teachers' Knowledge of Test Construction

In 1992, Marso and Pigge extensively reviewed the literature addressing teachers' skills and knowledge related to the development and use of teacher-made tests. They compared their findings to those of Mayo who had done a similar study in 1967, a quarter of a century previously, in which teachers' testing knowledge had been found inadequate. Although more teachers had completed a measurement course by 1992, 13% of Marso and Pigge's sample of practicing (K-12) teachers reported having no formal measurement training in their educational programs. About one-third of teachers (33%) reported completion of one course in measurement; measurement content was embedded in another course for another 6%. Most teachers also reported having received no school-sponsored inservice training or assistance in development and use of tests.

In a sample of 41 seventh and eighth grade teachers, Boothroyd, McMorris, and Pruzek (1992) found that 49% had completed at least one undergraduate or graduate measurement course. The majority of those who had taken these courses recalled that much content focused on standardized testing; few of this group (15%) recalled actually constructing classroom tests or critiquing them. The teachers also completed two tests: (1) a 65-item multiple choice test to assess knowledge of measurement concepts specific to classroom testing and (2) another in which teachers were asked to identify flaws in 32

junior high science or math multiple-choice or completion test items. Teachers' scores varied widely but averaged 53% on both tests. The authors concluded that "teachers' knowledge of measurement is not sufficient. . . .[and that these] deficiencies in measurement knowledge probably result at least partially from inadequate training given that 51% never completed a single measurement course" (p. 7).

Elementary and secondary school faculty must meet certification requirements that do not necessarily include coursework in test construction. In 1988, Stiggins reported that most teacher training programs did not require a course in educational measurement and some did not even offer one (p.364). Once out in the field, few school districts offer any direct technical assistance to teachers in dealing with daily classroom assessment. Districts that have evaluation and testing experts and offices are usually concerned with standardized testing (Stiggins, 1988, p.365) . Since a large proportion of K-12 faculty have been found to have inadequate backgrounds in measurement, are faculty in institutions of higher learning any different? At the college level, faculty are selected for their expertise in subject matter. Most have not been formally prepared as teachers. College faculty are not prepared to design, develop, and evaluate tests (MacCuish, 1986, p.7). The evidence suggests that the quality of tests being given in classrooms is mixed but probably of lower quality than is desirable.

Do teachers perceive their own testing knowledge and practices as adequate? Gullickson (1984) surveyed 391 third, seventh, and tenth grade teachers representative of grades and curricula in a rural, midwestern state. The instrument contained 44 items to be rated on a five point Likert scale. Gullickson found that teachers perceived themselves to have an adequate knowledge of testing. College courses were <u>not</u> perceived to be

relevant to classroom testing needs, teachers indicating that they learned how to test through on-the-job experience. Gullickson also found that teachers enjoyed relative independence in deciding how frequently to test and what to test. They also constructed tests reflecting their own instructional emphases. The vast majority of teachers indicated that they prepared and scored their own tests without assistance, although a fairly large number would have liked some assistance. Time constraints for development and analysis of tests were perceived differently, with many teachers indicating that tests could be used more effectively if sufficient time were available. However, few teachers delegate these tasks to others. Gullickson concluded that, in spite of teachers' relative comfort with their testing knowledge and their heavy use of tests, that teachers were possibly too comfortable and lacked sufficient sophistication in classroom evaluation and measurement. The perceived irrelevance of college courses to teachers' everyday testing needs suggested that measurement experts and college faculty have a limited understanding of teachers' needs. This creates a tension which may contribute to variation in the content of measurement courses for teachers.

Studies of Actual Teacher-made Tests

Marso and Pigge (1989) compared teachers' testing proficiencies as perceived by the teachers themselves, their principals, and supervisors. Teachers rated their own testing proficiencies higher than did principals or supervisors. Principals and supervisors had high agreement regarding relative skill level of the 313 teachers in the sample. Marso and Pigge, who also examined 175 teacher-made tests submitted by these teachers, found a marked discrepancy between the perceived quality of teachers' test writing proficiencies as rated by the teachers, principals, and supervisors and the quality displayed on their actual

tests (1989, p.20). They concluded that "principals', supervisors', and teachers' ratings . . . may not be very accurate guides for determining teachers' inservice training needs" (p.23) and that increased attention needs to be paid to item writing skills.

In the same research study, Marso and Pigge analyzed 455 test exercises within the 175 teacher-made tests. A test exercise was defined as "a group of items of a similar item type on a test" (p. 9). The researchers examined eight item types (e.g., completion, multiple-choice, essay), ten test format construction error types (e.g., incomplete directions, nonconsecutive numbering) and 66 item construction error types (e.g., incomplete stems, implausible alternatives). The researchers recorded one item construction error per exercise rather than each occurrence of the error, and one test format error for each type of format error on an entire test. Therefore, an item construction error could occur once or several times in the same exercise and be tallied as "1". Likewise, a test format error could occur once or several times in the same test and still be counted as "1". The investigators felt that this facilitated comparisons regardless of the number of items. Marso and Pigge found a mean frequency of 1.6 test format errors per test. In examining the various item types, the most error-prone test item type was matching items with 6.4 errors per exercise. Other error rates, listed in order of frequency of item construction errors, were completion, essay, true/false, multiple-choice, short response, problems, and interpretive exercises.

Multiple-choice items were the most popular item type and were fairly closely followed by matching and short response. Essay items comprised only 1% of the total items reviewed.

The cognitive levels of test items were rated by two judges and revealed that, except for math and science tests, 90-100% of items were at the lowest cognitive level (knowledge or recall). The findings suggest that "neither beginning nor [experienced] teachers . . .display high levels of proficiency in [item-writing] skills on their tests" (p.25).

In one of the earliest studies of teacher-made tests, Black (1980) studied 142 teacher-made science exams collected in 1971 in Nigeria. Sixty-four percent of the schools contacted submitted examinations, an excellent response rate considering the voluntary nature of the study. The examinations comprised five forms (levels or grades) at the secondary level. Black used a three-judge panel that included himself to classify items according to Bloom's taxonomy. The judges had an inter-rater reliability of .89 using an analysis of variance estimate for 295 questions. The cognitive level of questions on tests was not found to vary with teacher' qualifications. "The better-trained teacher" did not include more higher-level questions than "less well-trained" colleagues (p. 305). Black did find differences in cognitive levels of quesions between subjects and forms (grades). Knowledge level questions comprised 93.44% of biology test items in forms 3 - 5; 80.83% of chemistry test items for the same grades; and 62.21% of physics test items.

One of the most detailed studies of teacher-made tests was done by Billeh (1974) in Beirut, Lebanon. Billeh collected 33 exams from 18 schools, each covering one (three to five-hour) science unit from grade seven or ten. The biology, chemistry, or physics class sessions were tape recorded. Nonverbal activities were also observed and recorded. Other instructional materials and assignments were collected for study. Each teacher constructed a one-hour examination covering the unit under study. The teachers were not informed of the true purpose of the study.

Three judges independently rated each test item using Bloom's taxonomic levels. Discrepancies between judges were discussed and agreement reached as to the cognitive level measured by each test item. Billeh found that 71.83% of the test questions were at the knowledge (lowest) level and, contrary to Black's findings, no significant differences were found among biology, chemistry, and physics courses. Approximately 21% of the questions were at the comprehension level and 7% at the application level. There were no examples of questions above the application level. There was also no significant difference between the level of questions asked of seventh-graders versus tenth-graders. Billeh found no correlation between professional in-service training and the cognitive level of test items, which is in concert with Black's findings. He did find that part-time teachers in his sample emphasized slightly more lower-level items. An interesting finding was that teachers with more teaching experience actually used more lower-level items. Although Billeh's sample of 33 tests was not large, the concurrent recordings of classroom activity and examination of other instructional materials, as well as the concurrence of three judges, increases the credibility of Billeh's findings. Would his findings be validated by study of other educational levels, geographic areas, or later years?

In the United States, a federal court order for desegregation mandated scrutiny of all tests, including teacher-made tests, in the Cleveland school district (Chambers, 1982; Fleming & Chambers, 1983). A sample of 342 tests containing 8819 items was randomly selected from grades 1 - 12. Tests included those given in language arts, mathemathics, science, social studies, industrial arts, and French. Overall test documents were examined for considerations such as numbering, clarity of directions, legibility, and use of universal, non-discriminatory language. Many tests lacked directions. Illegible copies were not

unusual. Nearly one test in five contained errors in mechanics and technical conventions and a large number did not indicate the point value assigned to test questions. This is consistent with Marso and Pigge's findings that most tests had multiple test format errors.

Chambers and Fleming evaluated each test question using criteria specific to its item type. Numbers of items of each type were tallied. As would be expected, mathematics tests included high percentages of problems to solve. Short-answer or completion items were the most popular item type for tests in most other subjects. High-school English test items, expected to include the largest numbers of items requiring extended writing, included few essay questions. Overall, only about 1% of items reviewed were of the essay type. Fleming and Chambers expressed concern about the paucity of essay questions, suggesting that "serious questions about instructional priorities need to be raised" (p.37). It was hypothesized that teachers may perceive a lack of skill in developing such questions and/ or be concerned about the amount of time required to score tests. Fleming and Chambers also examined the questions in terms of six behavioral categories--knowledge of terms, knowledge of facts, knowledge of rules and principles, skill in using processes and procedures, ability to make translations, and ability to make applications. The first three of these are equivalent to Bloom's (1956) first taxonomic level (knowledge), the fourth and fifth to the second (comprehension), and the last to the third level (application). Overall, only 3% of test questions were at the application level, with the highest number of these being in mathematics. Eighty per cent of the overall test items called for knowledge of terms, facts, or rules and principles, with a higher proportion (94%) of lower-level questions being used at the junior high level than in

elementary grades (69%) or high school (69%). These figures are fairly consistent with finding from the earlier studies in Lebanon and Nigeria.

Oescher and Kirby (1990) examined 35 questionnaires and 34 mathematics and science tests containing more than 1400 items from one four-year high school in a mixed suburban-rural school district. The tests examined by Oescher and Kirby contained a range of 14 to 103 items per test and exhibited a variety of format problems. Teachers reported that one-fourth of items were written at Bloom's application, analysis, synthesis, or evaluation levels, but the researchers' analyses placed less than 8% of items at these levels, with virtually none at the synthesis or evaluation levels. Oescher and Kirby concluded that teachers tended to inaccurately classify higher-order items.

Harpster's (1999) doctoral dissertation studied levels of test questions written by public high school mathematics teachers in Montana. He found that nearly 60% of teachers wrote test question that assessed lower-level thinking skills when asked to write a question that assessed higher-level thinking skills. He also found no significant relationship between the teachers' assessment of higher-level thinking skills and professional development, college measurement courses, or continuing education.

McMorris and Boothroyd (1993) studied 82 tests (from 41 teachers) used in 7th and 8th grade mathematics and science classes in New York state. The teachers authored their own tests and more than half (54%) used some type of plan, although not usually a formal written blueprint, to construct a test. Since the sample tests were selected and submitted by the teachers, it is likely that these were the tests judged of highest quality by the submitting teachers. McMorris and Boothroyd suggest that they "likely provide an upper bound for typical practice" (p.336). Unfortunately, these sample tests showed fairly

high levels of flawed items, although it was concluded that "the tests were neither terrible nor worthy of accolades" (p.336). Teacher interviews indicated that many teachers did appropriately select item formats based on content and cognitive level to be tested. There was, however, no systematic study of the cognitive level of the questions.

Do college faculty have higher levels of skills? A pilot study done at the University of Central Florida in 1985 surveyed the faculty on their use of statistics. Thirty-one per cent (N=137) of the faculty responded to the survey, and although not generalizable, 91% of these faculty members stated that they designed and developed all the tests and measuring devices they used to evaluate student performance. The remaining faculty developed part of their examinations, but also employed some departmental, professional, or standardized tests (MacCuish, 1985). Most college faculty used the course objectives to develop tests, but did not use a taxonomy for test development. Most faculty did not do item analyses, obtain and use estimates of reliabilty and validity, calculate standard deviations or standard errors, or revise tests based on item discrimination power. Is the faculty at the University of Central Florida unique in this regard? No research studies have yet indicated whether these findings are typical of college faculty or unique to the study sample.

The literature related to teacher-made tests has focused mostly on elementary and secondary testing, most heavily in secondary mathematics and sciences. Systematic study of teacher-made tests in colleges and universities is absent. Lower cognitive levels are over-represented in tests questions included on teacher-made tests created by elementary and secondary faculty in a variety of geographic locations. The lack of systematic study of teacher-made tests in higher education raises a question as to the quality of these tests.

In nursing education, "critical thinking" has received great emphasis in recent years. Complex issues encountered in practice situations require the practicing nurse to make reasoned and informed judgments (Oermann & Gaberson, 1998). In order to be able to make appropriate decisions, higher level thinking is required. The goals of nursing programs must be to prepare graduates able to make sound decisions in a variety of situations. It is imperative that teaching strategies and evaluation measures in nursing education support higher level thought processes.

Chapter 3

Methodology

This study is descriptive and exploratory and examines the current status of teacher-made tests in nursing education. Final examinations were collected from associate degree and baccalaureate degree nursing programs and systematically examined. Current testing practices among nursing faculty are described. Comparisons of examinations given in each type of program and in beginning- and final-semester nursing courses are delineated. These comparisons include number and proportion of item types employed and ratings of cognitive levels of multiple-choice items.

Sample

The target population consisted of National League for Nursing Accrediting Commission (NLNAC)-accredited associate degree and baccalaureate degree schools of nursing in the United States as listed in Directory of Accredited Nursing Programs 1997. Using a computerized random number generator, a stratified random sample of 75 associate degree (ADN) and 75 baccalaureate degree (BSN) programs was drawn from all 50 states (but excluding United States possessions and territories). The resulting sample contained approximately proportional representation of program types, regions of the nation, and sizes of programs.

Considerable reluctance to respond to a request for copies of actual exams was anticipated. A 30-40% response rate (with two tests from each responding school) was expected to yield 90-120 final examinations (approximately evenly distributed between ADN and BSN programs and between first-semester and last-semester courses).

Hypotheses

1) There will be a difference in the use of essay items related to the mean number of students enrolled in a course.

2) Items on teacher-made tests in nursing education will include a predominance of the lowest cognitive level, fewer of the second level, and fewest written at the highest level.

3) There will be significant differences in the mean cognitive level of multiple-choice items when comparisons are made between first-semester and final-semester courses.

4) There will be significant differences in the mean cognitive levels of multiple-choice items when teacher-made tests are compared between ADN and BSN programs.

5) There will be significant differences in mean cognitive levels of test items related to the length of the exam (total number of items).

6) Inter-rater agreement in judging cognitive levels of exam items will be related to the mean cognitive level of items included on the exam.

7) Inter-rater reliability in judging cognitive levels of exams items will be related to the length of the exam (total number of items).

Procedures

After obtaining institutional review approval, the dean or director of each randomly-selected school was sent a packet containing two letters, an information sheet containing essential institutional review material, a one-page demographic instrument, and a stamped return envelope addressed to the researcher's residence. The first letter was a letter of introduction from the Dean of the College of Nursing and Health at Madonna

University in Livonia, Michigan (where the investigator is employed as faculty) which was included to verify the legitimacy of the request and of the researcher's credentials. This was done because it was believed that many program administrators and faculty would be reluctant to release exams without some such assurance and/or having means to check on the integrity or credentials of the person making such a request.

The second letter, from the researcher, explained the study and asked for a copy of each of two final exams given in the program. One of these exams was requested to be from a nursing course taught during the first clinical semester of the nursing major. The other was to be from a nursing course usually taught during the last semester of the program, whether or not all students take the course in the last semester. The dean or director was asked to pass the request on to appropriate faculty members. Deans, directors, and faculty were assured that security of test questions would be maintained and that specific schools or faculty members would not be identified. (See Appendix A for copies of documents related to the conduct of the study.)

Those submitting exams were asked for demographic information related to the nursing program, i.e., state in which the program is located, type(s) of program(s) offered, whether public or private, and whether the institution has sponsored continuing education offerings for faculty related to evaluation and measurement. Additional information included the name of the course, its position in the program, the approximate number of students enrolled in the course, and the highest academic degree of the primary faculty member or course coordinator. The demographic instrument was relatively short in order to maximize response. Each school was assigned a code number in order to identify pairs of exams and guide the follow-up mailing. Responders were asked to omit

the names of schools and faculty members. Any remaining identifying information was removed (or obliterated) on receipt.

Test documents were stored in a locked cabinet outside an educational institution when not being examined by the researcher and the expert raters. Documents will be destroyed after the study is completed.

Three weeks after the initial mailing, a follow-up letter was sent to schools that had not responded.

Analysis of the exams

Examinations arrived with a corresponding completed demographic instrument. Information related to each document was entered on the Individual Exam Record Sheet (Appendix B) and in a Microsoft Excel spreadsheet.

Each test document was examined by the researcher for appearance and format (i.e., legibility, directions to test-taker, inclusion of scoring criteria, consecutive numbering of pages and items), and the number of questions of each item type. This data was recorded on the Individual Examination Data Sheet prior to giving the exams to the raters.

Pilot study. A pilot study was conducted during the summer of 1999 using ten examinations containing a total of 849 items. Most of these examinations were extra exams submitted or were from excluded types of programs. Two experienced nursing educators rated items of all types using a four-level rating system. Inter-rater reliability was not adequate. After further study of these ratings, the classification system was revised to three levels and included more complete descriptions and clarifications of the items properly classified with each level.

The initial perusal of the examinations submitted for the study also revealed that

multiple-choice items were by far the most common item type--91.9% of all items submitted. Because of the difficulty of rating a variety of item types, many of which were small in number, it was decided to limit the study of cognitive levels to multiple-choice items. The prevalence of other item types is included in the descriptive data.

Main study. One hundred ten exams containing multiple-choice items were examined in the main study. The cognitive level of each multiple-choice item (1, 2, or 3) was judged independently by two nursing educators using the criteria outlined by the researcher (and detailed in Appendix C). Fourteen volunteer judges (twelve in addition to the original two) were randomly paired for rating purposes. (One additional judge was added when it became apparent that ratings would not be completed in a timely manner by the existing judges.)

Directions for rating items were given verbally and face-to-face to all judges either individually or in pairs to afford an opportunity to ask questions about the rating process. Written instructions were also supplied. Judges did not know with whom they were paired for rating purposes and did not know the source of the exams. Exams were assigned randomly to raters and usually given to raters in groups of four. Both exams of a paired set were given to the same pair of raters. Single exams were distributed based on the number of items the rater already had (i.e., if both exams of a pair were quite long, a single exam was given instead of a second pair). Each rater recorded her ratings on the Judges' Data Sheet (Appendix B). After the first judge had rated a group of exams, the exams were given to the "partner" judge. The first judge was then given the set recently completed by the other member of the rating pair. This process was then repeated so that all 110 exams could be rated. Thirteen judges rated four "batches" of exams. The

remaining two judges rated two "batches" each. As the ratings occurred over a five-month period (February to July 2000), some exams were rated when raters had little experience in the process and others near the end of the task.

For difficult-to-rate items, judges were given the option of marking an item with two levels ("split-level") in preference to leaving the space blank.

Procedure for Determining Inter-Rater Reliability of Cognitive Levels. After both judges completed rating each document, the researcher determined the cognitive level of each multiple-choice item and the mean cognitive index of multiple-choice test items in the complete document. The following procedures were used:

1. When both judges assigned an individual item the same numerical rating (e.g., 1, 2, or 3), that item was classified as the level indicated by both judges. The number of items so agreed upon were noted in relation to the total number of rated items included on the test and called the "initial agreement" and recorded on the Individual Exam Record Sheet.

2. If one judge used a split-level rating and the other assigned a specific number which included one of the split-level numbers, the item was classified as the level given by the rater using the single number.

3. If both judges used the same split-level combination, the researcher selected the final rating from one of those two levels.

3. If one judge used a split-level rating and the other judge used the single number that was not included in the split-level rating, the researcher determined the final level.

4. After resolving all the ratings that included split-levels from one or both judges, the resulting level of agreement was calculated and entered on the Individual Exam

Record Sheet as the "secondary agreement."

5. When each judge assigned an item rating using one number and those numbers were different, but contiguous (i.e., 1 vs. 2 or 2 vs. 3), the researcher examined the item and rated the item using one of the two numbers assigned by the judges. If the rating numbers were not contiguous (i.e., 1 vs. 3), the judge selected the rating deemed most appropriate using any number from 1 to 3.

6. After completing these procedures, each individual multiple-choice test item had a rating of 1, 2, or 3 which was the final determination of cognitive level for that item. The number of items of each level were then entered on the Individual Exam Record Sheet.

Once the final determination of cognitive level for all items included in a test document were determined, the mean cognitive index was calculated. This is the mean of all the ratings for that document. The potential range is 1.00 to 3.00. This number is used for further analyses.

Data Analysis

After all the examinations were rated and studied individually, aggregated data was analyzed. These include descriptive statistics relating to characteristics of the exam documents themselves as well as those of the responding programs. Appearance and format was also described. Additional descriptive data relates to the prevalence of item types with comparisons between program types and beginning and ending courses.

ANOVA was used to determine whether there were for differences in numbers of students enrolled in courses in relation to the item types employed in testing.

Pearson's product moment correlation coefficient was used to determine whether the level of inter-rater agreement was correlated with length of exam (number of items) or the mean cognitive index of the exam.

Paired t tests and/or ANOVA compared the mean cognitive indexes (MCIs) for 49 pairs of first- and final-semester exams. Independent sample t tests were used to determine whether there were differences between multiple-choice items on first semester ADN and first semester BSN final exams and between those on final semester ADN and final semester BSN exams.

Chapter 4

Results

One hundred thirty exams were received from 66 nursing programs in 31 states. Of this total, five were from excluded program types, five were exams that were sent in addition to the requested pairs, two exams (single submissions from different schools) were sent without demographic information, and three were group exams (one ADN first semester essay, one final semester ADN with 38 multiple-choice and 6 short answer items, and one BSN final semester essay). Of the remaining 115 exams, 110 contained multiple-choice items and were examined in the main study. Some of the excluded exams were used for the pilot study. The remaining five exams, from five baccalaureate programs in five different states, consisted entirely of short answer and/or essay items. Each of these arrived as part of a paired set. Although cognitive levels of essay items were not determined, data related to these exams are included in some comparisons.

Sample

Response. The overall response rate was 44% (50.7% from associate degree programs and 37.3% from baccalaureate programs). Table 1 (p. 30) details the response and the sources of exams.

Of 76 ADN schools (including one that was replaced in the sample), 34 (44.7%) schools sent paired exams and single exams were received from 3 (3.9%). One of these schools sent an extra exam. One school (1.3%) was found to be from an institution that offered both ADN and BSN courses of study and was excluded. Responses were received from 6 (7.9%) additional schools that did not send exams. Of the latter, faculty at three schools were unwilling to share exams, two schools had closed or were closing (including

one replaced in the sample), and one faculty member telephoned to discuss the study but did not send exams. The remaining 32 (42.1%) of the schools did not respond at all (including one for which the mailing was returned by the postal service).

Table 1

Sources of Exams Received

	North Atlantic	Midwest	South	West	Total
Requests to ADN schools	15	21	27	12	75
ADN schools responding with exams	5	10	17	6	38
Number of ADN exams	9	19	33	13	74
Requests to BSN schools	23	22	22	8	75
BSN schools responding with exams	9	10	6	3	28
Number of BSN exams	16	20	12	8	56
Total exams by region	25	39	45	21	130

Of 78 BSN programs, 23 (29.5%) sent paired exams and 3 (3.8%) sent single exams. Two of these schools sent a total of four extra exams. Two additional schools (2.6%) sent exams but were excluded because they did not have an undergraduate generic

program (i.e.- offered only courses for registered nurses who already held associate degrees or diplomas). Responses without exams were received from 11 (14.1%) of the schools. Of this group, six had faculty unwilling to share exams. Two schools which offered only degree completion program for registered nurses stated they did not use exams. These two programs and one new program that did not yet have printed exams were replaced in the sample. Extraneous material was mailed from one school. A completed demographic form without exams was also received. The remaining 39 (50%) programs did not respond at all to the mailing, including one school for which the follow-up mailing was returned by the postal service.

The final sample of exams containing multiple-choice items consisted of 110 exams from 61 programs in 29 states. Of these, 67 (32 pairs and 3 singles) came from 35 associate degree programs in 22 states and 43 (17 pairs and 9 singles) came from 26 baccalaureate programs in 18 states.

As can be seen in Table 1, exams from all four areas of the United States were represented for both types of schools. BSN programs from the South (27.3%) responded proportionately less than those from the Midwest (45.4%), North Atlantic (39.1%), or West (37.5%). ADN program response was greatest in the South (63.0%), followed by the West (50.0%), Midwest (47.6%) and the North Atlantic (33.3%).

Public vs. private institutions. All but three of the ADN programs were from public institutions. Two ADN programs were in private institutions and background information was not provided for one program. Baccalaureate programs were almost evenly divided between public and private schools; 14 (53.8%) were public and 12 (46.2%) were private institutions. Ten of the baccalaureate programs were housed in

institutions that also offered graduate degrees. Eight of these institutions were public.

Dates and designation of exams. Of the 110 exams containing multiple-choice items, 87 were labeled as "final" exams. Final exams were requested, but faculty were asked to submit the last exam in a course if there was no final. One exam was labeled as a mid-term and 17 had an exam number (e.g., Exam #3) or referred to unit or module numbers. This information was unclear for five exams. Of the five essay/short answer exams, all were labeled as final exams.

Most exams were received in April and May of 1999, with a few exams arriving during subsequent months. Of the 110 exams , 37 (33.6%)were dated in 1999, 21 (19.1%) from the fall 1998 semester, 21 (19.1%) from the 1997-1998 academic year, 4 (3.6%) were dated prior to fall 1997, and the remaining 27 (24.5%) had no date indicated on the exam.

Of the five essay/short answer exams, one was from the 1997-1998 school year, one was from the Fall 1998 semester, two were from Spring 1999, and one did not have a date indicated.

Types of Courses

Course titles and content showed some variation. First-semester ADN courses (N=34) were identified by a number only (n=11) or had "introduction," "fundamentals," "foundations," or "nursing process" in the title (n=20). Three courses had "adult" in the title. First-semester BSN courses (N=24) were likewise identified by a course number (n=4) or had "introduction," "fundamentals," "foundations," "nursing process," "basic concepts" or "basic skills" in the title (n=13). The remaining courses (n=7) related to "adults" (2), pharmacology (1), health assessment (1), and other topics (3) .

Final-semester courses showed more variation among programs. Among the 33 final semester exams from ADN programs, eleven were identified by a course number only. Ten had titles that included "advanced medical-surgical nursing," "high acuity," "critical," or "complex." Only two had titles indicating leadership and management content. In fact, when the actual subject matter of the exams was studied, it was found that 22 (66.67%) of these exams appeared to be from entirely clinically-based courses. The remaining 11 (33.33%) of the exams included leadership and management content or covered issues related to beginning practice. Nine of these 11 had leadership and management content integrated with adult medical-surgical nursing or pediatric nursing.

Of the 19 final-semester exams from BSN programs, four were from courses identified only by numbers, and eleven had "leadership" and/or "management" in the title. "Community health" was in three titles and "adult health" in only one. When examining content of test items, eleven (57.9%) exams appeared to be entirely related to leadership and management and/or professional nursing issues. In three exams, leadership and management was integrated with another area (nursing research, community health, or adult health). Only five (26.3%) of the exams covered material that was essentially clinical in nature (three were community health and two were adult health). All of the exams that were entirely composed of essay and short answer items came from the final semester of BSN programs. All of these covered leadership and management topics.

Faculty Preparation

Faculty submitting exams were asked to indicate the highest earned degree of the primary faculty member or course coordinator. Only two of the exams from associate degree programs had a doctorally-prepared faculty member in this role. The remaining 65

had masters' degrees, mostly a master's degree in nursing. For baccalaureate programs, 18 exams came from courses with a course coordinator who had a Ph. D. with one more indicating Ph.D. candidacy. Three additional courses had faculty coordinators with Ed.D. or J.D. degrees. Just over half (51.2%) of BSN courses were coordinated by doctorally-prepared faculty. The remaining course coordinators had masters' degrees, mostly in nursing.

Availability of Continuing Education on Evaluation and Measurement

Data were collected as to whether or not continuing education for faculty related to evaluation and measurement had been sponsored by the institution in the past five years. It is not known whether the authors of the exams in the study attended these sessions when they were offered or if they attended sessions elsewhere. Slightly more than half of the responding schools had sponsored continuing education related to this topic. It is not known if this is typical and what influence, if any, it may have had on the response rate and the study findings.

Table 2

Faculty Continuing Education on Evaluation and Measurement

Program type	Yes	No	Don't know	Total
ADN	19 (54.3%)	14 (40.0%)	2 (5.7%)	35
BSN	13 (50.0%)	11 (42.3%)	2 (7.7%)	26
RN-> BSN or joint ADN/ BSN	3 (60.0%)	0 (0.0%)	2 (40.0%)	5
Total	35 (53.0%)	25 (37.9%)	6 (9.1%)	66*

*Includes all programs that submitted exams whether or not included in study sample.

Appearance and format of exams

Each of the 110 exams containing multiple-choice items was examined for each of seven indicators of appearance and format. The first area noted was the presence of absence of directions to the test-taker. Directions ranged from minimal (e.g., "Choose the single best answer") to a separate cover page containing very complete instructions. Sixty-three (57.2%) of the exams included directions. Two more exams contained directions for one part of an exam, but no overall directions. Inclusion of directions did not appear to differ appreciably between programs or position of course in the program.

Point values for exam items were indicated on only six (5.5%) of the 110 exams, all from BSN programs. An additional ten exams showed point values for certain items.

Although all the exams submitted were at least fair in neatness and legibility, a number of problems were noted. Poor duplication quality was noted on five, variable font sizes on four, small hard-to-read print on 13, and problematic crowding of items on six. At least nine exams had multiple-choice items where one or more distractors were on a different page than the stem and/or the other distractors.

Eighty-nine (80.9%) of the exams had consecutively numbered pages. Problems with page sequencing included two exams (from separate institutions) that were missing alternate pages. Apparently only one side of a two-sided document was copied. Another exam was missing two or three pages (in the middle) that apparently contained twelve items. Two more exams were missing one page (and five items) each. Yet another exam contained a blank numbered page in the middle but apparently was not missing any items. When items were missing, they were deducted from the item total and were not considered in the analysis.

All but one exam had consecutively numbered items, although a few with multiple item types restarted numbering with a new section of the exam or when a different item type was used. As defined for this study, a required response is equal to one item. In some cases (with short answer and matching items), one item number required multiple responses. In other cases, one essay response was assigned to multiple item numbers. Item types and numbers were recalculated according to the number of responses required by the test taker.

Problems relating to item sequence were not uncommon. Two exams omitted one item each (i.e., skipped a number). One of these exams with an omitted number also had two items numbered in reverse order (i.e - #47 preceded #46). Two exams included the same item(s) twice; there was one repeated item on one exam and four repeats on the other. Since finding these errors was not a primary focus of this study, there may have been other repeated items that were not detected by the raters or the investigator. Three tests had an incomplete multiple-choice test item (one had no distractors, one had no discernible question, and one was missing a required diagram). Item numbers were adjusted appropriately when the actual number of complete items differed from the test numbering system.

Seeking out typographical and other usage errors was not a primary goal of this study. Raters were asked to mark obvious errors, but not to focus on proof-reading. The researcher also found many errors when analyzing the exams. The number of errors was counted for each exam and then divided by the total number of items on the exam in order to provide a basis for comparing exams of different lengths. See Table 3 (p.37) for figures on occurrence of errors. It can be seen that, while there were many exams for which no

errors were noted, error-free exams were not the norm and a few exams were rife with errors.

Table 3

Typographical and Usage Errors on Exams

Error Proportion	No. of Exams	% of Exams
0	23	20.9
up to 0.015	24	21.8
0.015-0.045	42	38.2
> 0.045	21	19.1

Non-questions were excluded from the analysis. These are items that are not related to the subject matter. They are essentially free questions and faculty probably include them either to give additional points to students or to have an even number of items on the exam. Thirteen exams each included one such item. Only one of these came from a final semester BSN exam. The other twelve were evenly distributed between ADN first-, ADN final-, and BSN first-semester exams.

Types of test items used

General. Multiple-choice items were by far the most commonly used item type and comprised 91.9% of all items submitted. The popularity of the multiple-choice item held true across all levels and program types. (See Table 4, p. 38) Mathematical problems and matching items each consisted of 2.4% of items, followed by true/false items at 1.3% and short answer at 1.1%. Essay items comprised 0.8% of items submitted.

Table 4

Exam item types

Item types ->	Multiple choice	Short answer	Matching	True/ False	Math problem	Essay	Total
Main Sample (N=110)	9,116 (92.6%)	93 (0.9%)	228 (2.8%)	136 (1.4%)	247 (2.5%)	28 (0.3%)	9,848 (100%)
Essay Exams (N=5)	0	5 (11.1%)	0	0	0	40 (88.9%)	45 (100%)
Excluded Exams (N=15)	871 (89.6%)	26 (2.7%)	37 (3.8%)	4 (0.4%)	17 (1.7%)	17 (1.7%)	972 (99.9%)
All exams (N=130)	9,987 (91.9%)	124 (1.1%)	265 (2.4%)	140 (1.3%)	264 (2.4%)	85 (0.8%)	10,865 (100%)

Comparison by program and semester.. Comparison of item types between programs and position in program also revealed some differences. Associate degree programs relied more heavily on multiple-choice items for testing than baccalaureate programs both at the beginning level (93.9% for ADN and 89.5% for BSN) and in the final semester (96.4% for ADN and 83.6% for BSN). Mathematical problems consisted of 2.9% of items on both first- and final-semester ADN exams but 1.8% on beginning BSN and 1.7% on ending BSN exams. Table 5 (p. 39) and Figure 1 (p.40) show item types by program and semester for the main sample and five individual essay exams.

Short answer, matching, and true-false items were relatively more common in BSN exams at both levels, showing use of a greater variety of item types.

Essay questions were totally absent from ADN program exams with one

exception. This was a first-semester exam from a course related to ethical, legal, and general professional issues. It was to be worked on initially by individual students and then completed by pairs. Because this study's definition of an exam specifies use by individual students, it was excluded from the main analysis. One BSN first-semester exam included one essay item. Overall, essay items intended for individual response comprised 0.8% of all items submitted, but 4.3% of items for BSN final-semester exams.

Table 5

Items of each type by program and semester

Program	Item Type	First Semester	Final Semester	Total
ADN	Multiple choice	3067 (93.9%)	2911 (96.4%)	5978 (95.1%)
	Short answer	35 (1.1%)	0	35 (0.5%)
	Matching	57 (1.7%)	19 (0.6%)	76 (0.8%)
	True/false	11 (0.3%)	2 (0.1%)	13 (0.2%)
	Mathematical	96 (2.9%)	87 (2.9%)	183 (2.9%)
	Essay	0	0	0
	Total ADN items	3266	3019	6285
BSN	Multiple choice	1849 (89.5%)	1289 (83.6%)	3138 (86.9%)
	Short answer	50 (2.4%)	13 (0.8%)	63 (1.7%)
	Matching	63 (3.0%)	89 (5.8%)	152 (4.2%)
	True/false	66 (3.2%)	57 (3.7%)	123 (3.4%)
	Mathematical	38 (1.8%)	26 (1.7%)	64 (1.8%)
	Essay	1 (0.0%)	67 (4.3%)	68 (1.9%)
	Total BSN items	2067	1541	3608

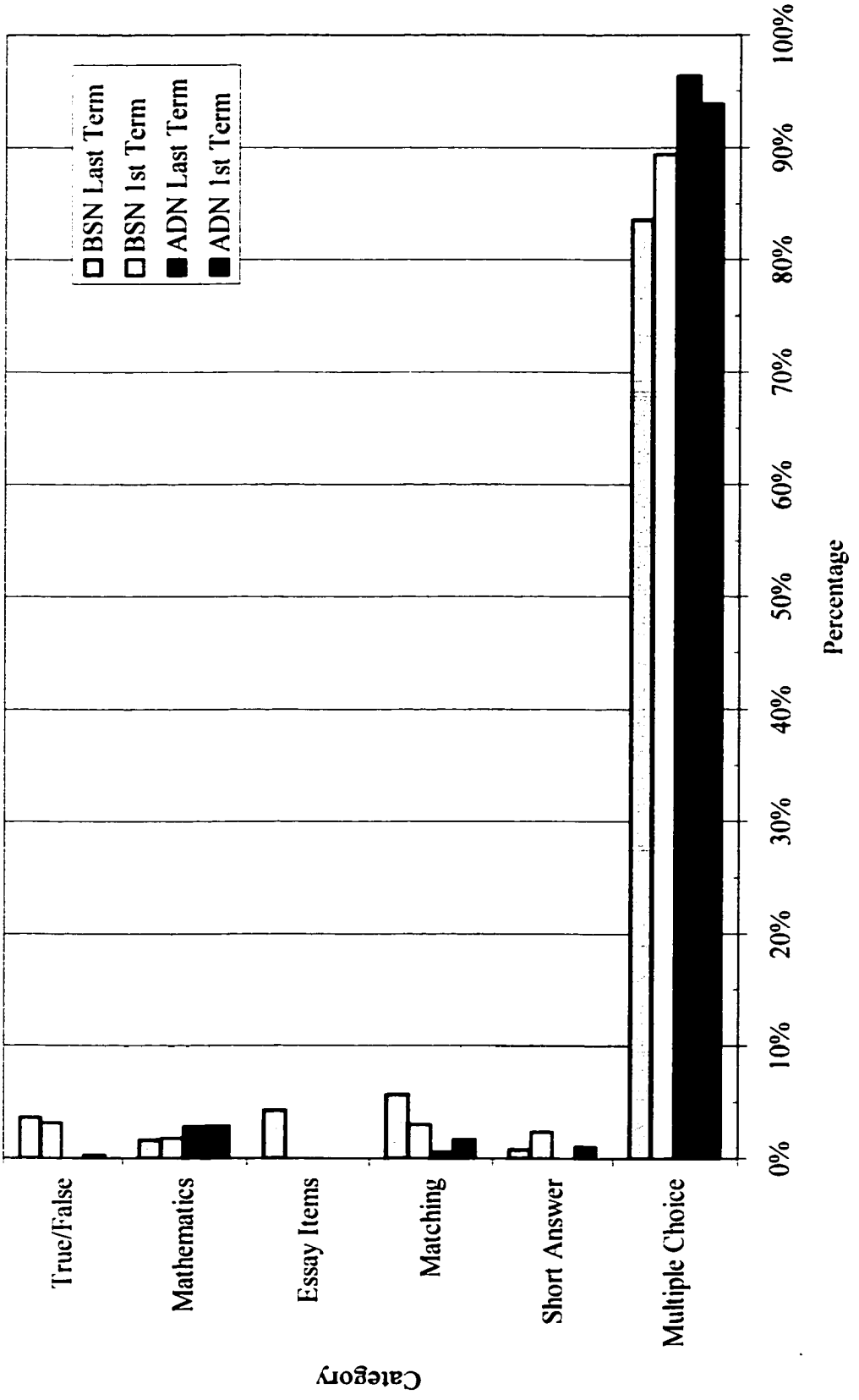

Figure 1: Exam item types by program & semester

By number of students. Exams were divided into three groups according to the types of items used. The groups were: (a) exams containing only multiple-choice items , (b) exams containing multiple-choice as well as any other item types except essay items , (c) exams containing essay items, with or without other item types. A one-way ANOVA was computed comparing the numbers of students enrolled in the courses represented by each of these types of exams.

Table 6

One-way ANOVA: Numbers of Students and Types of Exams

Source	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5365.380	2	2682.690	3.921	.023*
Within Groups	75942.243	111	684.164		
Total	81307.623	113			

There was a significant difference in the numbers of students taking the three types of exams. Tukey's HSD was used to determine the nature of the differences between exam types. Exams containing essay items were used in courses with fewer students (M=27.83, SD=11.78) than exams with all multiple-choice items (M=51.17, SD= 26.36) or exams containing a mixture of item types without essay items (M=49.66, SD= 27.63). There was no significant difference in student enrollment for the latter two types of exams.

Students enrollment ranged from 8 to 150 for the courses represented by these exams with a mean of 47.75 (SD=26.82). The mean number of students taking ADN exams was 51.36 (SD=27.49) and 42.60 (SD= 25.24) for BSN exams. This did not represent a statistically significant difference at the alpha = 0.05 level ($t[112] = 1.732$, $p = .086$).

Cognitive Levels of Exam Items

Each judge assigned a numerical rating of 1, 2, or 3 to each multiple-choice test item based on the rating criteria detailed in Appendix C. The first level is substantially similar to Bloom's knowledge level and/or Ebel's terminology and factual information levels. The second is comparable to comprehension in Bloom's scheme and explanation in Ebel's. The third level comprises all higher levels.

Proportions of items at each level. Table 6 and Chart 2 (p.44) indicate that, except for final semester ADN exams, more than half of multiple-choice items are written at the lowest cognitive level and only a little more than 5% are at the highest level. There were also similar percentages of second and third level items for all but final-semester ADN exams. Final-semester ADN exams had proportionately fewer first level items and more second and third level items. Cognitive levels are also reflected in the mean cognitive index for each exam. Comparisons of mean cognitive indexes between semesters and programs are described in a later section.

The mean cognitive index (MCI). After the final rating of each item was determined, the mean cognitive index (MCI) for all the multiple-choice items on a given test was determined by calculating the mean of all ratings for individual items included on that exam. The possible range is from 1.00 to 3.00. The actual MCIs ranged from 1.063 to 2.000.

Table 7

Proportions of Multiple-Choice Test Items at Each Cognitive Level

Program/ Semester	N	# MC items	Level 1	Level 2	Level 3
ADN first	34	3067	1673 (54.5%)	1230 (40.1%)	164 (5.3%)
ADN final	33	2911	1292 (44.4%)	1383 (47.5%)	236 (8.1%)
BSN first	24	1849	981 (53.1%)	766 (41.4%)	102 (5.5%)
BSN final	19	1289	734 (56.9%)	490 (38.0%)	65 (5.0%)
Total	110	9116	4680 (51.3%)	3869 (42.4%)	567 (6.2%)

Inter-rater agreement. Each exam was rated by two judges. All judges were volunteers and teaching colleagues of the researcher. Three were doctorally-prepared and the remaining twelve had masters' degrees, many with other advanced certifications. All were full-time faculty with a mean of 15.5 (range 6 - 28) years of experience teaching nursing. The judges were paired randomly and did not know who else was rating the same test items. Due to the number of responses and time required for analysis, an additional volunteer rater was added as the rating period progressed.

Of the 9116 items a total of 18 (<0.2%) were left blank by raters. Nine of these were on one exam by one member of the first rating pair. For this particular case, the level of agreement was based on the lower number of items, but the mean cognitive index (MCI) was calculated by using the rating of the judge who completed those items. Nine

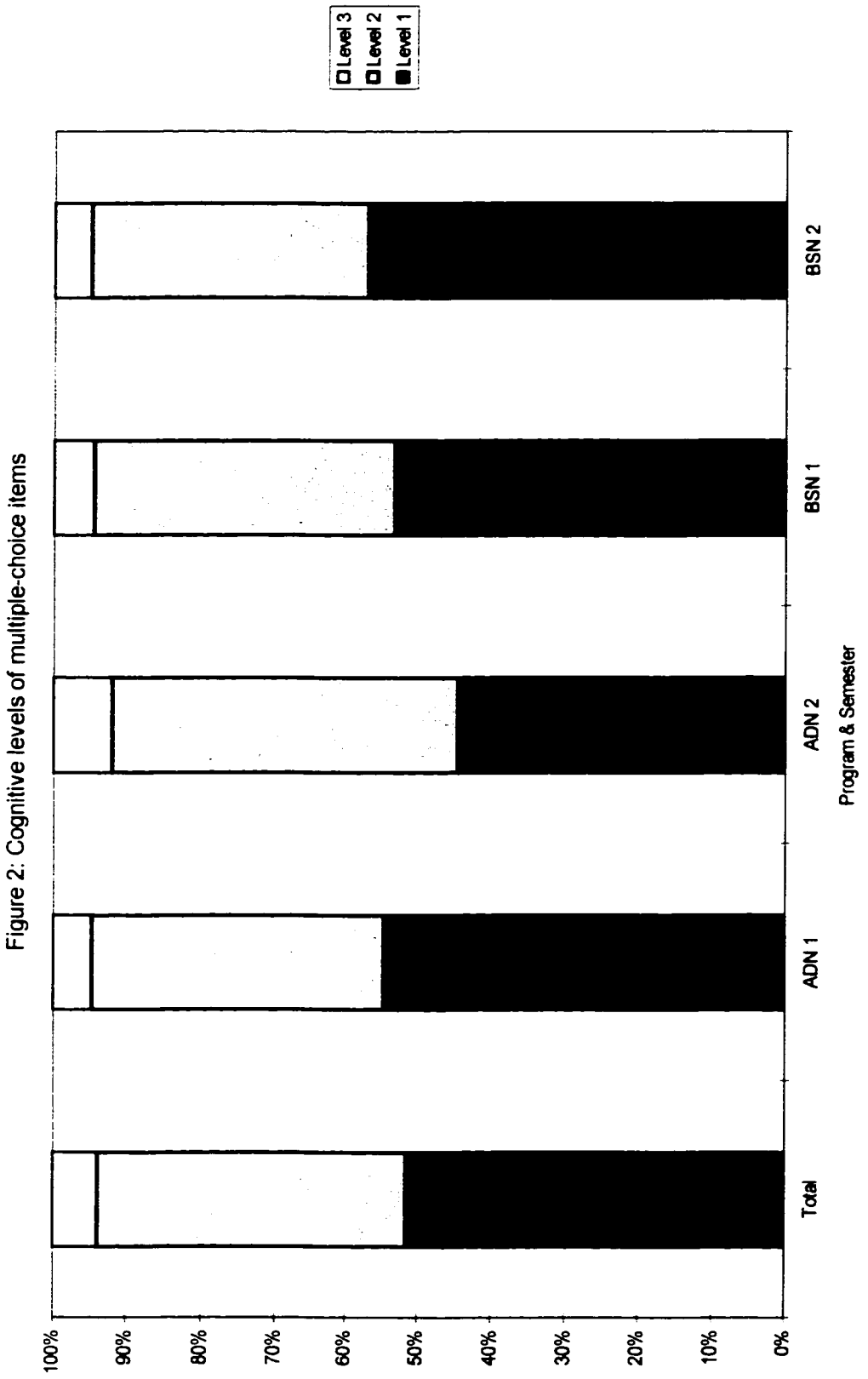

Figure 2: Cognitive levels of multiple-choice items

other items were left unrated, but by different raters and distributed over several different exams. In these cases, the rating of the judge completing the item was used to compute the MCI and the rater leaving the blank was assumed to be in agreement with the rating when calculating the raters' secondary agreement (i.e., the agreement after split-level ratings are resolved).

Table 8

Inter-rater agreement of cognitive levels

Pair #	Exams	Items	*Agreement
1	14	1313†	61.2%
2	16	1274	62.8%
3	16	1307	64.2%
4	15	1300	54.6%
5	17	1314	58.1%
6	16	1309	64.9%
7	7	628	56.5%
8	9	681	65.9%
Total	110	9116	M= 61.0%

* = Raters' secondary agreement
† = Agreement calculated on 1304 items

Influence of the third rater. The investigator acted as a third judge to resolve items on which the rating pairs disagreed. After all exams were analyzed, the investigator compared the final determination of each MCI with the MCIs that would have been assigned by each individual rater. Of the 110 exams containing multiple-choice items, the MCIs of 89 (80.9%) were between the individual raters' MCIs (i.e., ≥ the lower rating

and ≤ the higher rating). The third judge's ratings resulted in a higher MCI than either judge for 7 (6.4%) and a lower MCI than either judge for 14 (12.7%).

Relationship between inter-rater agreement and MCI. A Pearson product moment correlation coefficient was calculated to determine whether there was a relationship between the MCI and the level of agreement of the raters. These were found to have a significant negative correlation ($r(108) = -.660$, $p = <.001$). The more items of higher cognitive level, the poorer the agreement between raters.

Relationship between inter-rater agreement and exam length. A Pearson correlation coefficient was also calculated to determine whether the length of the exam (or number of items to rate) was related to the level of inter-rater agreement. In this case, there was no significant correlation ($r(108) = .103$, $p = .283$). There also was no significant correlation between the mean cognitive index (MCI) of the exams and the number of items ($r(108) = .146$, $p = .129$).

Comparison of Mean Cognitive Indexes for Exams

First-semester vs. final-semester exams A paired samples *t* test was calculated to compare the mean cognitive index for each first-semester nursing course to the mean cognitive index for the corresponding final-semester course for the 49 pairs of exams. The mean MCI for the first semester exams was 1.503 (SD= .165) and the mean MCI for the final semester exams was 1.58 (SD= .207). This represented a significant difference in MCI between first- and final-semester tests ($t[48] = -2.272$, $p = .028$).

Results from paired exams were also examined separately for ADN and BSN programs. The mean MCI for first semester ADN exams was 1.515 (SD=.148) and the mean MCI for corresponding final semester exams was 1.636 (SD= .183). This also

showed a significant increase from first to final semester (t[31]= -3.039, p= .005). When examining results from BSN programs separately, the mean MCI for first semester exams was 1.481 (SD= .195) and the mean MCI for corresponding final semester exams was 1.477 (SD= .216). There was no significant difference between first- and final-semester exams in BSN programs (t[16]=.053, p= .959).

Comparisons between programs. When examining single exams in addition to the paired exams in comparing results between programs, the results of independent sample t tests showed no significant difference between mean MCIs for ADN and BSN first semester exams. The mean MCI for 34 first semester ADN exams was 1.513 (SD = .144) compared with a mean MCI of 1.493 (SD = .234) for 24 first semester BSN exams (t[56]= .448, p = .656).

Comparisons of multiple-choice items on final semester exams for the two types of programs showed a mean MCI of 1.633 (SD= .181) for 33 final semester ADN exams and a mean MCI of 1.480 (SD = .223) for 19 final semester BSN exams. An independent samples t test revealed a higher MCI for multiple-choice items on ADN final-semester exams as compared to final-semester BSN exams (t[50]= 2.683, p = .010). Six of the 19 final semester BSN exams included essay items. Essay items contributed 50% to the score on one exam, 8.7% on another, and 5% on a third. Although the total number of points on the remaining three exams was unclear, the test documents indicated the value of essay items as 10, 18, and 40 points. The mean MCI of the remaining 13 exams (which were more similar in types of items to the corresponding ADN exams) was 1.510 (SD= .218). When these thirteen final-semester BSN exams were compared with the final-semester ADN exams, the difference was not statistically significant (t[44]= 1.988, p= .057).

Chapter 5

Discussion

The findings of this study have implications for nursing faculty in preparation of exams for their own courses. The descriptive data provide information about prevailing testing practices not previously found in the literature. Findings relative to format and appearance, use of various item types, and cognitive levels of test items are fairly consistent with previous studies done in other settings.

Issues related to format and mechanics

Since teacher-made nursing examinations have not previously been systematically studied, data related to test construction in higher education or in nursing programs has not been available. Consistent with Marso and Pigge's (1989) and Chambers and Fleming's (1983) findings, format and appearance deficiencies in the exams were fairly common. Although the majority of the exams included written directions to the test-taker, these were often minimal. There was a nearly universal lack of written indication of relative values of items on the documents. It is possible that this information was given to students verbally or at another time. Most students would also probably assume that items are valued equally on a 50 or 100-item exam consisting of entirely multiple-choice items. However, many exams which lacked point value information had multiple item types and/or an "odd" number of items (e.g. - 63, 73, 58, 44). Faculty would be well-advised to pay increased attention to supplying written instructions and written indication of point values of items.

Errors in assembling the test document were also fairly common. Although consistently perfect test documents may not be a totally achievable goal, missing pages,

missing or incomplete items, and items that "straddle" pages are avoidable. All of these problems potentially contribute to reduced validity of the exam. These deficiencies may stem from faculty time pressures to write exams under the pressure of a deadline as suggested by Anderson (1987). However, additional time is required to double-check for these problems before administering the exam to students and may require reading by a second faculty member.

Although all the exams received were neatly and legibly printed, the legibility of some was compromised by small font sizes and crowding of items. Older students, who are becoming a larger proportion of nursing students, may have particular difficulty with small fonts. A few exams (some obviously with multiple authors) had different print sizes and styles for different sections of the exam. This could constitute a distraction for some students.

Typographical and other usage errors are probably underestimated in this study. Although seeking out errors was not a focus of the study, many errors seemed to "jump off" the pages at the raters. Some errors were relatively minor; others changed the sense of the item and/or made interpretation difficult. Proof-reading one's own exams before administration is essential. Although consistently perfect exams may not be realistic, faculty need to take the extra time to check for errors and perhaps have a second faculty member proofread. When errors alter the sense of the item or make interpretation difficult, the issue of how to handle the scoring of that item will inevitably arise.

Evaluation and Measurement Inservice

Literature from other areas suggest that most teachers have inadequate preparation related to evaluation and measurement. Although accrediting bodies in nursing have been

increasingly stressing evaluation strategies, the typical background of nursing faculty in this area is not known. Continuing education on this topic had been offered to faculty at just over half of the schools in the sample in the past five years. It is not known whether this is typical. It is possible that schools which have sponsored such inservice sessions place a greater value on evaluation. These schools may have been more likely to respond for a request for exams and may have produced exams of higher quality. Data was not requested as to whether the authors of the specific exams submitted for the study actually attended the inservice session(s) offered at the school, if those faculty attended similar offerings elsewhere, or what formal coursework the faculty may have had in this area.

Types of test items

The preponderance of multiple-choice items on teacher-made nursing exams is very clear. It is unknown whether exams from other fields of study in higher education are similar in this regard. One possible reason for heavy use of multiple-choice items may be that the nursing licensure exam (NCLEX-RN), taken by new graduates of both types of programs, uses multiple-choice items in a computerized testing format.

Although multiple-choice is the predominant item type, associate degree programs in this sample used them more heavily and baccalaureate programs used a greater variety of item types. Mathematical problems (often presented in a multiple-choice format) were included in ADN exams (2.9% at in both first- and final-semester exams) and BSN exams (1.8% in first-semester and 1.7% in the final semester). The majority of mathematical problems involved calculation of medication dosages, an issue related to safe professional practice and a perennial concern of nursing faculty .

Less than 1% of all items submitted were of the essay variety. This is similar to findings of Fleming and Chambers (1983) and Marso and Pigge (1989). Essay items are low in number partially because fewer can be included on an exam. More time per item is required both for test taking and for grading. Time for grading also increases with the number of students, which may be an important consideration for busy nursing faculty. In addition, the grading of essay items can be more difficult to defend with students than when grading other item types. Having well-defined grading criteria decreases, but probably does not totally eliminate, this problem.

Essay items were used almost exclusively in final-semester BSN courses. Essay items can be written at a variety of levels and are usually intended to evaluate higher-level thinking skills. Of the twenty-four BSN final-semester exams, nearly half (n=11) included essay and short answer items. (Five were entirely essay and short answer and six more included some essay and short answer items as well as multiple-choice items.)

Difficulty of items and cognitive level

There is some tendency to compare difficulty of test items to cognitive level. These are two different concepts. The "difficulty index" is the proportion or percentage of students in the group who answer the item correctly (Carey, 1988, p. 250). Factual information and definitions, which would be included in the lowest cognitive level, can be very "difficult" if students have not been exposed to the information. It is more accurate to think of the cognitive level in relation to the types of thinking and/ or prerequisite knowledge required. Does the item demand knowing, understanding, or using? Do students need to recall, interpret, or utilize?

Inter-rater agreement of cognitive levels

The most serious problem in interpreting findings of this study is related to the relatively low level of agreement between the raters. All the raters were faculty in the same baccalaureate nursing program. Their professional backgrounds were somewhat varied, but all were given the same verbal and written instruction about the rating process and allowed to ask questions. Judges were asked to look at the item as it was written and not to make assumptions about what was or was not taught in class.

The ratings were done over a five-month period. Several factors made rating items a difficult and time-consuming task. Oescher and Kirby (1990) concluded that secondary teachers tended to overestimate the cognitive levels of their own test items. In this study, inter-rater agreement decreased as cognitive level increased, suggesting that faculty also have difficulty classifying higher-order items written by others. Part of this difficulty lies in not knowing what was presented in class or the textbook. A complex scenario that was thoroughly discussed in class and subsequently appears on an exam may be answered by students' recalling information, even though the item may appear to require higher level thinking. On the other hand, a recall-level item may follow description of a complex-appearing scenario.

Another possible confounding factor is the varied backgrounds of the raters even though all were full-time faculty in the same institution. Although no clear patterns were evident, the raters' familiarity with the material on the test may have influenced ratings. Each rater had exams containing topics similar to those she taught as well as others with which she was less familiar.

Several of the raters informally commented that rating items was difficult and that many items were puzzled over before reaching a decision. Some raters expressed concern that ratings done on one batch of exams were not perfectly consistent with their own ratings on another batch. All indicated that they referred to the written criteria fairly often and tried to rate them in accordance with the criteria.

Cognitive levels of test items

Although several tests had a large number of higher-level multiple-choice test items, they are probably fewer than most faculty believe. Over half of all multiple-choice items on first-semester ADN and both first- and final-semester BSN exams were judged to be of the lowest cognitive level.

First-semester courses tend to be introductory in nature and, of necessity, include learning terminology and factual information. Although BSN students usually have had more general education course work, students in both programs need to learn much of the same material.

The difference between the mean cognitive indexes for final-semester ADN exams as compared to final-semester BSN exams was not anticipated and is likely to be of concern to faculty in BSN programs. It must be remembered that only multiple-choice items were considered in calculating this index. The aggregate of BSN final semester exams used fewer multiple-choice items and a wider variety of item types. Depending on the number and kinds of items used, the complexity of the overall exam may be increased or decreased. For example, the exam with the lowest MCI of the 110 (1.063) was a leadership and management exam from the final semester of a BSN program. The MCI was calculated from ratings of 32 multiple-choice items with no stated point values. This

same exam also contained seven essay items worth a total of 15 points. The overall complexity of an exam with multiple item types may not be well-represented by a rating of only the multiple-choice items. In addition, written exams, from any type of program, usually represent only part of the student evaluation process. Exams may be favored for the evaluation of factual information with other techniques being employed for evaluation of higher level thinking skills.

There are other limitations in interpreting this information. First, the majority of the exams did not indicate relative values for scoring items. Where scoring can be determined at all, essay questions generally carry larger values for grading. True/false and matching questions are often assigned lower values. Second, the exams could make a major or a relatively minor contribution to the final course grade. Third, other means of student evaluation such as papers and projects may make major contributions to course grades. Fourth, there is no information as to how many items need to be answered correctly in order to achieve a passing score. Tests that require 70% correct as a passing score will be different from those that require 80% for passing. In addition, some faculty may grade based on the class average or distribution of scores.

It was interesting to note that some schools do not use written exams for student evaluation. This appears to be more common for baccalaureate degree completion programs (i.e., bachelor's degree programs designed for ADN and diploma graduates who already have nursing licenses) and for final-semester courses. This was a reason given for not participating in the study as well as the reason for receiving some unpaired exams.

An interesting sidelight is that the same items often appeared on exams from widely varied locations. No count was kept of this, but when an exam item seemed very

familiar, previously analyzed exams were re-examined and a similar, and often identical, item was frequently found. This suggests a common source of test items, most likely published test banks provided with texts or NCLEX review materials.

Limitations in the ability to generalize

In examining results, it must be noted that exams from schools who voluntarily sent exams for study may differ in important ways from exams from other institutions. Submitters of exams are likely believe that their exams are of better-than-average quality. Faculty who are less confident of the quality of exams are less likely to have submitted them. Caution is, therefore, required in applying these findings.

Suggested Guidelines for Faculty

Although judging the cognitive levels of one's own test items is fraught with difficulty, faculty need to be conscious of cognitive levels when tests are being written. If the aggregate cognitive levels of items submitted for this study are prevailing practice, first-semester exams contain just over half (54%) of first level items, with most of the remainder (40.6%) at the second level. This leaves a small proportion (5.4%) of higher-level items. Faculty need to consider whether the overall cognitive levels of their tests are congruent with the objectives or desired outcomes for their courses and programs. Most faculty probably believe their own tests contain more than the average number of higher-level items. Most faculty probably need to consider adding items of higher cognitive levels to their exams and replacing lower-level items with those of a higher level.

It would be somewhat logical to expect to find fewer first level items and more higher level items on multiple-choice exams given later in programs. The ADN final-semester figures of 44.4% first level, 47.5% second level, and 8.1% third level does show

this difference. Faculty who teach at any level of any program need to be aware of graduates' performance on program outcome measures such as performance on the licensing examination and feedback from employers. If these are acceptable, perhaps there is no need to make major changes. If outcome measures show need for improvement, close examination of teacher-made exams should be undertaken as part of the study process.

When exams consist of several item types, the cognitive levels of all item types need to be considered. True/false items, by their very nature, are classified at the lowest cognitive level. Matching and short answer items are often at the lowest cognitive level, although it is possible to write them at higher levels. Adding lower level items to an exam will not raise the overall level of the exam. Well-written essay items, even with the drawbacks of increased time and difficulty of scoring, offer an alternative for evaluating higher level thinking.

Recommendations for Further Study

The most serious problem in interpreting this study is the relatively low level of inter-rater agreement of cognitive levels. It is very difficult to classify cognitive levels of test items whether one is writing items or rating those written by others. Continued refinement of definitions of cognitive levels may or may not make for better inter-rater agreement in a future study. Other approaches to the systematic study of teacher-made exams may need to be formulated.

Overall, exams in nursing school today are of mixed quality. Although none of the exams in this study was terrible, only a few were worthy of high praise. Many of the problems observed in the exams in this sample of teacher-made tests were previously

discussed in literature from other educational levels and from several countries. Any deficiencies, therefore, are not likely to be unique to nursing education. Exams from other areas of study in higher education also need to be examined.

Appendix A

Letters and Documents

Related to Conduct of Study

MADONNA UNIVERSITY

36600 SCHOOLCRAFT ROAD
LIVONIA, MICHIGAN 48150-1173
(734) 432-5300/FAX 432-5393

April 14, 1999

FIELD(1) FIELD(2) FIELD(3)FIELD(4)
FIELD(5)
FIELD(6)
FIELD(7)
FIELD(8)

Dear FIELD(1) FIELD(3):

This letter is a brief one of introduction for Ms. Kathleen Cross, who is a doctoral candidate at Wayne State University in Detroit, Michigan. She is currently working on her dissertation research. Kathleen has been an excellent nursing faculty at Madonna University for the past eleven years.

The enclosed letter from Kathleen asks that you pass on her request for two samples of nursing exams to appropriate faculty in your program. Since exams are sensitive materials and ordinarily closely guarded by Schools of Nursing, I wanted to assure you that Kathleen's intentions are legitimate and that she has been a conscientious, ethical, responsible, and mature nursing faculty for all of the eight years I have known her as Dean of the College of Nursing and Health here at Madonna University.

If you have questions, you may call me at (734) 432-5465. Please note this number listed on page 34 of the AACN Membership Directory for 1997-1998. Thank you, in advance, for your kind consideration of Kathleen's request.

Sincerely,

Mary S. Wawrzynski, PhD, RN
Dean, College of Nursing and Health

MSW:kc

Kathleen Cross
14060 Golfview
Livonia, MI, 48154
Phone: (734)432-5452 (work); (734) 464-7863 (home)
E-mail: kcross@smtp.munet.edu

FIELD(1) FIELD(2) FIELD(3)FIELD(4)
FIELD(5)
FIELD(6)
FIELD(7)
FIELD(8)

Dear FIELD(1) FIELD(3):

My name is Kathleen Cross. I am an Assistant Professor of nursing at Madonna University in Livonia, Michigan and a doctoral candidate in Educational Evaluation and Research at Wayne State University in Detroit, Michigan. My colleagues and I, like nursing faculty everywhere, write several exams each semester. These exams, typically authored by faculty for their own courses, have a strong influence on what students study and learn. My dissertation research relates to teacher-made tests given in nursing schools--a topic with a dearth of research.

In order to study this, I need exams from a variety of schools of nursing. Your school was randomly selected from a list of NLNAC-accredited schools of nursing. I need your help to obtain a final exam from:

1) a beginning nursing course (preferably the course that would most closely correspond to "fundamentals of nursing") **and**

2) a course usually taken in the final semester before graduation (perhaps one related to "nursing leadership/ management" or "senior nursing").

If there is no "final," the last written test taken by the students for the course should be substituted..

Would you be so kind as to pass on my request to the appropriate faculty members? If you would prefer that I make this request directly to the faculty member(s), will you please give me information as to the best way to contact the correct people? My telephone numbers and e-mail address appear at the top of this page as well as on the information sheet.

I assure you and your faculty that I will maintain the security of the exams submitted to

me and will not use them for any purposes other than to determine the status of teacher-made exams in nursing education today. Results will be reported as aggregates. Individual faculty or schools will not be identified.

My study has been approved by the Behavioral Institutional Review Board and the Graduate School at Wayne State University. I have enclosed a copy of an information sheet for those submitting nursing examinations as well as a short demographic data sheet. Please return the data sheet with the examinations in the envelope provided. If you need an additional return envelope, data sheet or information sheet, I will be happy to send another upon request.

Thank you in advance for your co-operation.

Kathleen Cross, RN, MSN, M.Ed.

Information for Submitters of Nursing Examinations

Introduction/ Purpose:
The purpose of this research study is to determine the cognitive levels of questions on teacher-made tests used in nursing programs. Comparisons of test questions from different types of programs and different levels of the same programs will be made.

Procedure:
1. Please send a copy of the final exam from:
   a) a nursing course taught in the first semester of the nursing major **and**
   b) a nursing course taught during the last semester of the program (even if some students take the course before the final term).
   If the course has no "final," please submit the **last** written test taken by students. .
2. **DO NOT INCLUDE YOUR NAME OR THAT OF YOUR SCHOOL.** (If the name of the school appears on the copy, it will be removed immediately upon receipt.)
3. Complete the **Nursing Exam Data Sheet** which was mailed or faxed to you. Please mail this form with the examinations to:

Kathleen Cross, RN, MSN, MEd.
14060 Golfview
Livonia, MI, 48154-5282

Privacy/ Protection of Subjects:
Your submission of exams will be considered permission to use them only for the purposes of this study as well as permission to destroy them after the study is completed. Exams will be stored in a locked cabinet from time of receipt until they are destroyed following the study. Individual schools or faculty members will not be identified. Results will be reported only as aggregates.

Potential Benefits:
Examinations given in nursing courses have a profound influence on what students study and learn. Faculty put a great deal of time and effort into planning and evaluating instruction and aim to stimulate critical thought. To date, little is known about how nursing educators actually test. It is hoped that this study will increase awareness of the cognitive levels of test questions.

Questions/ concerns: If you have any questions concerning this study or your participation in it at any time, please contact me by mail (at the address above), by telephone at (734)464-7863 (Home) or (734) 432-5452 (Madonna University) or by E-mail at kcross@smtp.munet.edu.

If you are interested in my findings, I will be happy to share them when my study is complete . Please telephone, mail or e-mail requests to the addresses/ numbers above. Your willingness assist in my research is greatly appreciated.

Kathleen Cross
14060 Golfview
Livonia, MI, 48154
Phone: (734)432-5452 (work); (734)464-7863 (home)
E-mail: kcross @ smtp.munet.edu

May 19, 1999

FIELD(1) FIELD(2) FIELD(3)FIELD(4)
FIELD(5)
FIELD(6)
FIELD(7)
FIELD(8)

Dear FIELD(1) FIELD(3):

Recently I sent requests for copies of teacher-made nursing exams to a random sample of nursing programs around the country. Response to my request has been encouraging, although more exams are still needed.

If your program is one that has already responded, I would like to convey my most sincere gratitude for your willingness to share exams as well as having taken the time to send them.

Perhaps you and/ or your faculty plan to send exams, but in the busyness that occurs near the end of the academic year, have not yet done so. I want to let you know that I am still interested in obtaining exams from schools such as yours. As I know that faculty from many programs will be away from campus during the summer months, I would like to encourage you to take a few minutes before finishing your school year to put exams in the return envelope, fill in the Nursing Exam Data Sheet, and drop it all in the mail.

I have enclosed another Nursing Exam Data sheet. If you need another return envelope or information sheet, I will be happy to send another upon request by telephone, e-mail, or regular mail. I want to assure you again that I am holding the exams I receive in strict confidence and that individuals or schools will not be identified.

Thank you for considering my request,

Kathleen Cross, RN, MSN, M.Ed.

Kathleen Cross
14060 Golfview
Livonia, MI, 48154-5282
Phone:(734)464-7863(home); (734)432-5452 (work)
Email: kcross@smtp.munet.edu

February 15, 2000

Dear Colleague:

Thank you for your willingness to serve as an item-rater for my research. This is an important study for nursing educators and I appreciate your assistance.

Enclosed you will find three or four examinations, a copy of the descriptions of each cognitive level, and sheets for recording your ratings of the items. Please enter your initials and date on each rating sheet page. Use a pen or pencil to indicate the level that you judge to best fit the item. **Please use the criteria as outlined** to rate the **multiple-choice questions only.** (I have already marked out the items that are not multiple-choice.) **Do not** be concerned with the intent of the teacher, what you believe was probably presented in class, or taxonomies that you previously studied. **Focus on the item as it is written** and refer to the descriptions as needed. It is not necessary for you to be a content expert or even to know the correct answers. In the event that you are unable to rate an item after reasonable consideration, please enter the two categories you were most considering (e.g. - "1/2" or "2/3"). Please use this option as sparingly as possible.

Please circle obvious typographical, spelling, or grammatical errors as you encounter them, but don't be particularly concerned with proofreading. In the event you find an incomplete item, please make a note in the rating space.

Your ratings will be compared with those of another volunteer judge for inter-rater reliability. The submitters of the exams have been promised anonymity and that the security of their exams will be protected. For this reason, copies cannot be made either of the exams or of individual items from them.

Thanks again for all your help. It is truly a pleasure to work with colleagues who are so cooperative and willing to share in this project.

Appendix B

Tools for Data

Collection and Organization

#_____ **Nursing Exam Data Sheet**

Program data:

1. Please indicate the state in which your program is located _____

2. Indicate program(s) offered. (Check all that apply).
   ADN ____ BSN ____ MSN ____ Ph. D. _____ Other _____

3. Is your institution? Public ____ Private ____

4. Has your institution sponsored continuing education offerings for faculty related to evaluation and measurement within the last five years?
   Yes ____ No ____ Uncertain _____

Exam data:

Test from **first semester** course in **nursing major:**

Title of course _____

Number of students in course _____

Highest degree of primary faculty or course co-ordinator?
(BSN, MSN, PhD, etc.) _____

Test from **last semester** course in **nursing major:**

Title of course _____

Number of students in course _____

Highest degree of primary faculty or course co-ordinator?
(BSN, MSN, PhD, etc.) _____

Individual Exam Record Sheet

Exam # _____ Test is from first semester _____ Last semester _____

Total # of Questions/ items _____

<---------JUDGES' RATINGS---------->

Item type	# on test	# level 1	# level 2	# level 3
Multiple-choice				

Short answer/ completion _____

Matching _____

True/false _____

Mathematical _____

Essay / extended answer _____

Initial agreement: _____

Secondary agreement: _____

Mean cognitive level: _____

Overall test characteristics	**Yes**	**No**
Are directions to the test-taker included?		
Are point values of items indicated?		
Is the copy legible / neat?		
Are pages consecutively numbered?		
Are test items consecutively numbered?		
Are spelling, grammatical, or typographical errors evident?		
Are "non-questions" included? (Number)		

Comments:

Rater #1
Rater #2

Test # _____

Judge's Data Sheet

Please rate each question individually using the criteria given in "Instructions to Judges." Your rating will be compared with that of another judge. If ratings differ, you will be asked to discuss the item with the other judge. Use additional sheets as necessary.

Item #	Cognitive Level		Item #	Cognitive Level
1			21	
2			22	
3			23	
4			24	
5			25	
6			26	
7			27	
8			28	
9			29	
10			30	
11			31	
12			32	
13			33	
14			34	
15			35	
16			36	
17			37	
18			38	
19			39	
20			40	

Appendix C

Definitions of Cognitive Levels

Definitions of cognitive levels

Please use the following descriptions when rating test items and refer to the accompanying examples for clarification of the three levels. Remember to **rate the item based on these descriptions,** not assumptions about the intent of the teacher or what was presented in class. Be alert for items that may look like complex scenarios but only ask the student for recall or comprehension.

Level 1 - Knowledge/ Information

Includes items that ask for:

A. the meaning of a particular word or term. Essentially asks. "What is x?"
Examples:
1. item asks for definition of terms

2. item asks for phrase or sentence in which the term is used most appropriately

3. item asks for selection of response that does NOT fit the definition / term

B. knowledge of a specific fact or descriptive detail; based on recall
1. what is true/ how things actually are in real world

2. may ask "**who, what, where, when, how much?**" (e.g.- single lab values, names, dates)

C. knowledge of principles and generalizations
1. Stem may include words such as **"generally," "usually," "normally," or "often."**

2. May ask student to *recognize* a law or principle (e.g. Boyle's law or Starling's law), general rule, guideline, "rule of thumb."

3. May ask for *recall* of general descriptions or characterizations, including:
   a. which factors are involved in x?
   b. which is normally/ most common/ most usual . . .
   c. interventions or actions that are rule-bound
   d. trends or developments

4. comparison of types or classes recognizing similarities and/or differences (e.g.- types of medications)

Level 2 - Comprehension/ Interpretation

Includes items that test understanding and must be answered on basis of reasoning rather than recall. (Basic "knowledge" plus some interpretation is necessary for response.)

A. Reasoning:
   1. Stem may include words such as **"why" or "because."**

   2. Asks the student to identify causes, effects, reasons, rationale, purposes, functions, factors, or expected consequences.

   3. Response requires an understanding of relationships

   4. Item may call for evidence for or against a statement or procedure
      a. which is/ is not appropriate care/therapy/ nursing measure for x?
      b. which is correct /incorrect?
      c. which has priority/ should receive emphasis?

B. Principles/ generalizations:
   1. Asks student to select appropriate illustration, rewording, explanation of a law, principle, theory, guideline, etc. (More complex than simply recognizing or stating.)

   2. Asks student to recognize which law/principle is being used/illustrated.

C. "Simple" scenarios
   1. Asks for appropriate assessment measures/ actions /interventions in "classic" or "textbook" cases.

   2. Asks student to select the common / most usual nursing diagnosis for a general /unspecified person/ group with . . . .

   3. Asks student to select appropriate nursing intervention/measure/ action for a general/unspecified person/ group with . . .

D. Interpretation of tracings, simple charts, graphs, multiple value lab reports (e.g.- ABGs, CBC)

Level 3 - Application/ Utilization

Include items that require making inferences and judgments based on knowledge and understanding of criteria/principles/ standards.

A. Prediction questions which require reasonable inferences
   1. what will result over time or after certain factors in the situation are changed (more complex than simple cause-and-effect)

   2. most effective or appropriate treatment/ intervention in order to achieve x

B. *Use* of laws and principles (the precise law or principle may or may not be specified) e.g. - "Apply principles of growth and development to plan care for . . ."

C. Complex practice scenarios for which student must select the best/ most appropriate action/ intervention/ response for a specific care circumstance/ situation. (More complex or less well-defined than "classic" or "textbook" cases.)

D. Judgments based on internal and external criteria.
   1. Asks student to identify the extent to which materials, objects, etc. meet criteria. (May include complex legal and ethical issues).

   2. Items requiring student to select nursing diagnosis when student is required to know defining characteristics of the diagnosis and then determine whether the situation meets those criteria.

   3. Asks student to select the best/most appropriate nursing intervention/ measure/ action for a particular care situation when the student is required to analyze those needs.

It may help to consider the three levels as "knowing, understanding, using" or "recall, interpretation, utilizing."

Sources:

Bloom, B. S. (Ed.) (1956) Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I: Cognitive Domain. New York: David McKay Company.

Ebel, R.L. & Frisbie, D. A. (1986). Essentials of Educational Measurement. (4th ed.) Englewood Cliffs, NJ: Prentice-Hall.

Oermann, M. H. & Gaberson, K.B. (1998). Evaluation and Testing in Nursing Education. New York: Springer Publishing Company.

2-4-00

References

Anderson, S. B. (1987). The role of the teacher-made test in higher education. In D. Bray & M. J. Belcher (eds.) Issues in Student Assessment: New Directions for Community Colleges, no. 59. San Francisco: Jossey-Bass.

Balch, J. (1964). The influence of the evaluating instrument on students' learning. American Educational Research Journal, 1, 169-182).

Billeh, V. Y. (1974). An analysis of teacher-made science test items in light of the taxonomic objectives of education. Science Education, 58(3), 313-319.

Black, T. R. (1980). An analysis of levels of thinking in Nigerian science teachers' examinations. Journal of Research in Science Teaching, 17(4), 301-306.

Bloom. B. S. (Ed.) (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I: Cognitive Domain. New York: David McKay Company.

Boothroyd, R. A., McMorris, R. F., & Pruzek, R. M. (1992). What do teachers know about measurement and how did they find out? Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA., April 1992. (ERIC Document Reproduction Service No. ED 351 309).

Carey, L. M. (1988). Measuring and Evaluating School Learning. Boston: Allyn and Bacon.

Carter, K. (1984). Do teachers understand principles for writing tests? Journal of Teacher Education, 35(6), 57-60.

Carter, K. (1986). Test-wiseness for teachers and students. Educational Measurement: Issues and Practice, 5(4), 20-23.

Cassidy, V.R. (1987). Test construction techniques. Journal of Nursing Staff Development, xx (Fall), 154-158.

Chambers, B. (1982). Quality control review of teacher-made tests. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada, April 1983. (ERIC Document Reproduction Service No. ED 227 166).

Chenevey, B. (1988). Constructing multiple -choice examinations: Item writing. The Journal of Continuing Education in Nursing, 19(5), 201-204.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 58(4), 438-481.

Demetrulias, D. A. & McCubbin, L.E. (1982). Constructing test questions for higher level thinking. Nurse Educator, 7 (3), 13-17.

Ebel, R. L. (1965). Measuring Educational Achievement. Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L. & Frisbie, D. A. (1986). Essential of Educational Measurement (4th ed.) Englewood Cliffs, NJ: Prentice-Hall.

Ellsworth, R. A., Dunnell, P. & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? Journal of Educational Research, 83 (5), 289-293.

Farley, J. K. (1989). The multiple-choice test: Developing the test blueprint. Nurse Educator, 14(5), 3 - 5.

Farley, J. K. (1990). Item analysis. Nurse Educator, 15 (1), 8 - 9.

Fleming, M. & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W.E. Hathaway (ed.), Testing in the Schools. New Directions for Testing and Measurement, no. 19. San Francisco: Jossey-Bass.

Flynn, M. K. & Reese, J. L. (1988). Development and evaluation of classroom tests: A practical application. Journal of Nursing Education, 27(2), 61-65.

Frisbie, D. A. (1983). Testing achievement beyond the knowledge level. Journal of Nursing Education, 22(6), 228 -231.

Gaberson, K. B. (1996). Test design: Putting all the pieces together. Nurse Educator, 21(4), 28 - 33.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77(4), 244-248.

Harpster, D. L. (1999). A study of possible factors that influence the construction of teacher-made problems that assess higher-order thinking skills. Dissertation Abstracts International, 60 (04A), 1055.

Joachim, G. (1992). A dynamic test system: Its development and implementation. Computers in Nursing, 10(5), 213- 218.

Klisch, M. L. (1994). Guidelines for reducing bias in nursing examinations. Nurse Educator, 19(2), 35-39.

Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. Teachers College Record, 91(3), 422-436.

MacCuish, D. A. (1986). The course development model in higher education: Improving tests and instruction. (ERIC Document Reproduction Service No.ED 273 169).

Manuel, P. & Sorenson, L. (1995). Changing trends in healthcare: Implications for baccalaureate education, practice and employment. Journal of Nursing Education, 34(6), 248-253.

Marso, R. N. & Pigge, F. L. (1988). An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, April 1988. (ERIC Document Reproduction Service No. ED 298 174).

Marso, R. N. & Pigge, F. L. (1989). The status of classroom teachers' test construction proficiencies: Assessments by teachers, principals, and supervisors validated by analyses of actual teacher-made tests. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, CA, March 1989. (ERIC Document Reproduction Service No. ED 306 283).

Marso, R. N. & Pigge, F. L. (1991). An analysis of teacher-made tests: Item types, cognitive demands, and item construction errors. Contemporary Educational Psychology, 16, 279-286.

Marso, R. N. & Pigge, F. L. (1992). A summary of published research: Classroom teachers' knowledge and skills related to the development and use of teacher-made tests.

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 1992. (ERIC Document Reproduction Service No. ED 346 148).

McMorris, R. F. & Boothroyd, R. A. (1992) Tests that teachers build: An analysis of classroom tests in science and mathematics. (ERIC Document Reproductions Service No. ED 350 348).

McMorris, R. F. & Boothroyd, R. A. (1993) Tests that teachers build: An analysis of classroom tests in science and mathematics. Applied Measurement in Education, 6(4), 321 -342.

National League for Nursing Accrediting Commission (1997). Directory of Accredited Nursing Programs. New York: Author.

Oermann, M. H. & Gaberson, K. B. (1998). Evaluation and Testing in Nursing Education. New York: Springer.

Oescher, J. & Kirby, P. C. (1990). Assessing teacher-made tests in secondary math and science classrooms. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA, April 1990. (ERIC Document Reproduction Service No. ED 322 169).

Reilly, D. E. & Oermann, M. H. (1990). Behavioral objectives: Evaluation in nursing (3rd ed.). New York: National League for Nursing.

Stanton, M. P. H. (1983). Objective test construction: A must for nursing educators. Journal of Nursing Education, 22(8), 338 - 339.

Stiggins, R. J. (1988) Revitalizing classroom assessment: The highest instructional priority. Phi Delta Kappan, (January), 363-368.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. Educational Measurement: Issues and Practice,5, 5-16.

Talbot, G. L. (1994). Revitalizing teacher-made tests: Quality control procedures. (ERIC Document Reproduction Service No. ED 380 406).

Van Ort, S. & Hazzard, M. E. (1985). A guide for evaluation of test items. Nurse Educator, 10(5), 13-15.

ABSTRACT

COGNITIVE LEVELS OF MULTIPLE-CHOICE ITEMS ON TEACHER-MADE TESTS IN NURSING EDUCATION

by

KATHLEEN CROSS

December 2000

Advisor: Dr. Donald R. Marcotte

Major: Educational Evaluation and Research (EER)

Degree: Doctor of Philosophy

This study examined teacher-made tests from nursing programs throughout the United States. Randomly-selected nursing programs were asked to submit two teacher-made final examinations. One hundred thirty examinations were received from 66 associate degree and baccalaureate degree nursing programs in 31 states.

Multiple-choice items were the most common type (91.9%) of item submitted. Other item types in order of popularity were mathematics (2.4%), matching (2.4%), true/ false (1.3%), short answer (1.1%), and essay (0.8%). All essay items were on baccalaureate program tests.

Cognitive levels of multiple-choice test items from 110 test documents were rated by two experienced nursing educators. Inter-rater agreement was 61%. There was a significant negative correlation ($r(108) = -.660$, $p = <.001$) between inter-rater agreement and cognitive level, which supports other research that suggests that faculty have difficulty

correctly identifying higher-level items. When raters disagreed, a third judge (the researcher) determined the final rating level according to pre-established criteria. Once ratings for individual items were determined, a mean rating was calculated for all multiple-choice items on the examination.

Paired sample t tests showed an increase in cognitive level from beginning to final-semester courses for associate degree programs ($t[31] = -3.309$, $p = .005$), but not for baccalaureate programs ($t[16] = .053$, $p = .959$). Independent samples t tests revealed no significant difference in cognitive level between beginning-semester exams between the two types of programs ($t[56] = .448$, $p = .656$). The mean cognitive index for associate degree program final-semester exams, however, was higher than that for final-semester baccalaureate exams when all exams were compared ($t[50] = 2.683$, $p = .010$). Final-semester baccalaureate exams covered different types of courses and utilized a greater variety of item types. When final-semester exams containing multiple-choice items but no essay items were compared between programs (33 associate degree exams and 13 baccalaureate exams), there was no statistically significant difference between programs ($t[44] = 1.988$, $p = .057$).

Overall, over half (51.3%) of all multiple-choice items were judged to be at the lowest cognitive level, with most of the remainder (42.4%) at the second level, and fewest (6.2%) at the third (or highest) level.

Kathleen J. Way Cross

Professional Education:

Doctor of Philosophy, Wayne State University, December 2000
Major: Educational Evaluation and Research

Master of Education, Wayne State University, December 1994
Major: Educational Psychology.

Master of Science in Nursing, Wayne State University, December 1969
Functional area, Teaching; Clinical area, Medical-Surgical.

Bachelor of Science in Nursing, Wayne State University, June 1965.

Professional Experience:

August 1990 - present. Full-time faculty, Madonna University College of Nursing and Health, Livonia, Michigan. Assistant Professor since September 1993.

August 1988 - April 1990. Adjunct clinical faculty, Madonna College Department of Nursing.

June 1976 - August 1990. Assistant Chief, Nursing Service, (part-time, evening shift) Veterans Administration Medical Center, Allen Park, Michigan.

July 1973 - June 1976. Staff nurse, Veterans Administration Medical Center, Allen Park.

January - April 1968. Staff nurse, Eugene Talmadge Memorial Hospital, Augusta, Georgia (teaching hospital for the Medical College of Georgia).

June - September 1967. Staff nurse, Visiting Nurse Association, Detroit (Northern Office).

August 1965 - May 1966. Staff nurse, Veterans Administration Hospital, Ann Arbor, Michigan.

Professional Memberships:

National League for Nursing / Michigan League for Nursing

Sigma Theta Tau (International Nursing Honorary Society), Kappa Iota chapter

Personal: Married, three children, two grandchildren.