2002

Developing a culminating assessment for psychology undergraduate students

Tammy Louise Reichley

DEVELOPING A CULMINATING ASSESSMENT FOR PSYCHOLOGY UNDERGRADUATE STUDENTS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Psychology:

Industrial / Organizational

---

by

Tammy Louise Reichley

June 2002

DEVELOPING A CULMINATING ASSESSMENT FOR PSYCHOLOGY UNDERGRADUATE STUDENTS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

by

Tammy Louise Reichley

June 2002

Approved by:

[redacted]

Dr. Janet Kottke

May 28, 2002
Date

[redacted]

Dr. Janelle Gilbert

[redacted]

Dr. Cynthia Crawford

ABSTRACT

There is an expectation of universities to demonstrate the value added from academic programs. The purpose of this project was to create a comprehensive exam for a California State University to measure student learning in Psychology within a multi-matrix method Outcomes Assessment process. There were two parts of this study: (a) the development of a reliable and valid comprehensive exam appropriate for the target audience, and (b) the evaluation of student performance on the exam as related to ability, content objectives, and the curriculum of the major.

A one hundred item exam was piloted, revised, and re-administered to more than 500 students. Item statistics and reliability and validity coefficients were determined for the exam. Five hypotheses on examinee performance were supported as indicators of the increase in learning from the Psychology program. The content objectives assessed in the exam pertained to the application of real studies and theories that can be generalized into life beyond Psychology classes. As a multi-method framework, the OA program for CSUSB is promising as a measure of net gains in student knowledge.

ACKNOWLEDGMENTS

This project has been a challenging, yet a valuable experience for me. There are a lot of people who were constantly by my side during this more than two year effort. First of all, I want to thank my family, who has been incredibly supportive of me with their love and encouragement throughout my pursuit of higher education. Special thanks to my thesis advisor, Dr. Jan Kottke, whose humor, patience, and encouragement are greatly respected and appreciated and made this project and others during my undergraduate studies memorable learning experiences. I also want to thank a few of the very special people in my life, the true friends, Charlene Reck, who actually read this paper, Lori Macias, Christina Rich, Stacey Smith, Ahu Gokmen, and Carrie Rodgers, who have been so helpful, supportive, encouraging, and most significantly, understanding in the completion of this endeavor. I want to acknowledge my managers, Nancy Sarti and Kathy O'Halloran for being very supportive and flexible with me to finalize this project. Earnest appreciation is also extended to David Morris for his knowledge and insights in the world of statistics that helped me to grasp the statistical concepts

that made the completion of these analyses possible. Finally, I want to sincerely thank my husband, Kendall Reichley, who deserves an award for his unconditional support, patience, nobility, and good humor. He endured late nights of word-by-word reviews and demonstrated an unselfish willingness to learn more than he ever wanted to about assessment of outcomes. His calm presence made it possible for me to go through very trying times and still remain optimistic that I could do it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

LITERATURE REVIEW

The Problem: Measuring College Graduate Outcomes

Many businesses claim they are disappointed that students are graduating from college without the requisite skills to succeed in a professional environment (Magill, 1998). Lacking basic communication and computational skills (Vandament, 1987), it has become very evident that there is a large number of students graduating from institutions of higher education who are poorly qualified and poorly educated, resulting in a diminished confidence in higher education (Resnik & Goulden 1987). Not only are businesses concerned with the quality of their applicants, the public is interested in how their tax dollars are being spent, legislative representatives are pressured to satisfy the public interests, and universities must meet criteria to receive accreditation. Additionally, committees like the Education Commission of the States of 1986 and the National Governor's Association of 1986 have outlined requirements and requested proof of effectiveness to merit funding to schools (Banta & Moffet, 1987). As a result, universities are becoming acutely aware of the need to monitor the

academic growth of students and to provide evidence of the outcomes of students' educational experiences (Halpern, 1988).

Outcomes Assessment

If universities must rely on contributions from businesses, tax money from the public, legislative decisions, and approval from accrediting agencies, there must be some measure of educational accountability or tangible proof to demonstrate the value added from any academic program (Magill, 1998; Banta & Moffet, 1987; Popham, 1981). Outcomes assessments provide schools with the opportunity to demonstrate what the net gains in student knowledge are (Halpern, 1987; Astin, 1987). Tests provide evidence of a school's effectiveness (Popham, 1981) and schools can use outcomes assessment results to provide the quality assurances potential employers and contributors insist on, as well as evidence of a quality education that potential students seek (Halpern, 1987).

By measuring outcomes, the university can track trends and impacts of the university experience on groups of students and then make meaningful comparisons to the university's values and priorities (Halpern, 1987).

Measurement of the effectiveness of learning also allows the value of a good curriculum with good instruction to be demonstrated for a demanding public as well as provide useful input to improving instruction for a demanding student body.

With many pressures to implement "outcomes assessment" programs what does it really mean to the universities? Tom Angelo defined outcomes assessment:

> "Assessment is an ongoing process aimed at understanding and improving student learning. It involves making our expectations explicit and public; setting appropriate criteria and high standards for learning quality; systematically gathering, analyzing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance. When it is embedded effectively within larger institutional systems, assessment can help us focus our collective attention, examine our assumptions, and create a shared academic culture dedicated to assuring and improving the quality of higher education." (California State University Outcomes Assessment Meeting CA State University, Bakersfield, 1999).

There are numerous internal benefits to institutions that can result from the implementation of such rigorous assessment practices. An outcomes assessment program measures student learning which provides data to influence curriculum development, department objectives, and faculty involvement. If developed properly, outcomes assessments

are based on the curriculum of the school, providing the necessary link between curriculum and assessment (Willis, 1994). Student outcomes assessment results can be used to correct curriculum weaknesses and to confirm curriculum strengths. The results from the student assessments can provide a basis for altering the sequencing in the curriculum, adding or deleting specific courses, creating higher standards, and offering better advisement to students.

Measuring the students' performance with outcomes assessment will also allow the faculty to determine if students have mastered the materials expected as well as get feedback on student gains from their university experiences. That feedback can provide the opportunity for a university or department to fine-tune department objectives and practices. Teaching, in turn, can be based on department objectives resulting in a more focused approach to instruction (Lien, 1971; Krueger & Heisserer, 1987).

It seems with all of the opportunities available to improve the curriculum and ensure that students are learning, that faculty would be the strongest supporters of

the outcomes assessment, but this hasn't been the case. Most faculty are unfamiliar with the assessment language, uncertain of its purposes, and fear it threatens the status quo (Halpern, et al., 1993). In addition to the fear that assessment practices will add to their current responsibilities, faculty members have expressed concerns that they will experience a loss of autonomy in the classroom if assessment is too structured (Willis, 1994). On the contrary, faculty should still have the primary freedom to develop their own teaching methods, select materials, use examples, and present theories to cover their topics, but arguably they should agree upon the same objectives for like courses (Curry & Hager, 1987). To alleviate this concern, Halpern (1987) suggests that teachers participate in curriculum design and assessment policies. If assessments are developed internally, the whole process can draw on the expertise of teachers rather than threaten them (Hargreaves, 1989 as cited in Willis, 1994). By giving them an opportunity to engage in curriculum design, define important learning objectives, and practice assessment development, teachers may voluntarily alter their teaching approaches to be

consistent with a system they have helped develop (Pennycuick, 1990 as cited in Willis, 1998). This faculty involvement is a step toward raising standards because educational quality begins in the classrooms; mandatory assessments alone won't remedy the problems of the educational system (Eisner, 1993, as cited in Willis, 1994).

Vandament (1987) asserts that colleges that have implemented successful assessment programs report positive gains for faculty. By basing educational quality on how much students learn (Krueger & Heisserer, 1987) faculty energy can be redirected to focus on instructional progress for students, resulting in a renewed enthusiasm for teaching and interest in student growth (Vandament, 1987).

Some faculty members have been reluctant to embrace assessment of outcomes because of the existence of factors beyond their control. Curry and Hager (1987) described the implementation of outcomes assessment at Trenton State College where the faculty had been opposed to assessing student learning in general education. They argued that because of the large number of transfer students and the number of elective courses offered they would have

difficulty in measuring learning. After conducting a pilot with an external test, they gained insight into their program, but increased their list of reservations. With no incentive for students to spend the time needed to take the test and no effect on grade point average or graduation, they assumed students wouldn't put forth the effort to do well. In confirmation, students didn't perform very well. The raw score differences between Seniors and Freshmen were much smaller than they had expected. However, they discovered that the differences between native students and transfer students were not significant (Curry & Hager, 1987).

The Trenton State faculty also feared that personnel decisions would be based on the results. After being assured that the assessment was solely an evaluation of the program, not of individual teachers or students, their fears were allayed. With more objectivity, the faculty began to understand the developmental benefits of such a process and started to contribute to its evolution. To evaluate their program, they recognized that they needed more than one method of assessment to measure student gains, and that they needed to agree on educational goals

and match them to courses. They created a means to assess their goals, administered the assessment and analyzed the results. Then they evaluated the instrument, revised the curriculum and created a cycle to repeat the process to ensure constant improvements. Their discovery at this point was that students did not demonstrate the skills that the faculty claimed to be stressing, and they also didn't show the weaknesses they expected. They came to recognize that the tool they had resisted provided a valuable means of improving their program.

Aware of the faculty (and student) fear of punitive actions related to results, UTK created a program with the Tennessee Higher Education Commission to receive bonuses in reward for evaluative efforts and good results (Banta & Moffet, 1987). The assessment results are never used in a punitive way or as a means to restrict funding or place blame, so the assessments are done judiciously in an attempt to attain qualitative information with a focus on continuous improvement.

Halpern (1987, 1988) described three constructive uses of the outcomes assessment data. Aggregate data is often analyzed to measure program effectiveness in meeting goals,

implement enhancements to the curriculum and to improve program services. Budget decisions have also been aided by conclusions from aggregate data in addition to being used to demonstrate accountability to the demanding external and internal constituencies that affect the budget. On the other hand, student level data has been used as a gateway to enforce minimum competency requirements. These three uses only address the institution level perspective, when individual student gains could also be derived from the data (Krueger & Heisserer, 1987). Without the threat of negative consequences, students could be given valuable feedback on their academic progress (Halpern, et al., 1993). One could argue that academic progress could be encouraged with foreknowledge of an assessment. By declaring expectations for achievement, the school may motivate and direct student learning by making students aware of assessment practices. Astin (1987) stated that concrete assessment procedures make clear what skills are expected of students to be developed and demonstrated, and that knowledge of these expectations enhances the learning process.

The evaluation of "outcomes" is more than just knowledge acquisition, however (Erwin, 1991). An exit exam can not explain the totality of a student's experience in a university. Erwin points out that the impact of college on students includes intellectual, emotional, cultural and social development, while simultaneously providing a greater exposure to society and individuals. All of these factors contribute to the college experience and influence the graduating senior entering the world. Based on this premise, outcomes assessment practices are designed to incorporate a complete program of systematic evaluation of cognitive, affective, and motivational dimensions (Ewell, 1987). It is recommended to incorporate a range of assessment methods (Sheehan, 1994), which might include student exit interviews, alumni surveys, portfolios, records of achievement, course requirements, professional development exercises, peer reviews, curriculum evaluation, and comprehensive examinations.

Ewell (1987) listed recommendations for implementation of an outcomes assessment program. He suggested that schools should use existing information from the registrar's office, create a visible center for assessment

activities, make a statement about the importance of the program, experiment with pilot programs, critically evaluate existing model programs, and learn from the experience. In agreement with Halpern (1987, 1998), Ewell also recommends that in order for the results to be taken seriously, they need to be used in identifiable ways like improving the curriculum and planning the budget.

Many institutions are developing their own assessments. Locally developed tests are emerging as a preferred approach to be able to provide a better match between test and curriculum (Ewell, 1987; Banta & Moffet, 1987; Curry & Hagar, 1987). Because of the availability of item level data, local tests allow the reflection of the curriculum, detailed analyses of student performance, and the identification of areas of strength and weakness (Astin, 1987). Local test development also allows the flexibility to decide on the use of multiple choice, essay, oral or problem solving approaches as needed to best cover the content. However, because of the absence of normative data from an internal test, it has been recommended to combine efforts with comparable institutions to allow the comparison of scores across schools, cooperate in

development efforts, share a pool of items and save money (Halpern, 1988, Halpern et al., 1993).

Merits of a Comprehensive Examination

One of the most telltale measures of student academic learning is arguably the comprehensive exam. Comprehensive exams can be an extensive assessment of knowledge and application of concepts as well as an opportunity for students to gain a comprehensive grasp of their academic major field and integrate their learning (Loughead, 1997), not just a superficial test of facts requiring rote regurgitation. In this sense, test results can provide evidence of student learning to the university, and they may also serve to reinforce learning within the students (Krueger & Heisserer, 1987; Astin, 1987).

Tests reinforce learning. Many individual college courses are completed, compartmentalized, forgotten, and never revisited by the student, let alone integrated into a meaningful whole (Anderson, et al., 1984). A culminating assessment provides the impetus for a student to revisit material previously learned and to become familiar with it again. If a student goes back to study material from a previous course in preparation for a culminating exam,

Hattie & Jaegar (1998) postulated that there will be an increase in student learning. With the connection of methods and facts a reconceptualization of the information occurs (Hattie & Jaegar 1998). Additionally there is the acquisition of new information not absorbed the first time because the student has developed a greater knowledge structure in which to integrate the information (Mayer, 1989, as cited in Goldstein, 1993). Accordingly, Anderson, Krauskopf, Rogers, and Neal (1984) found in a search of the literature and practice, that comprehensive exams in use are intended to "provide an opportunity for students to organize their thinking and integrate what has been learned."

Brown (1983) states that the primary purpose of achievement tests is "to increase, facilitate, and motivate student learning." Students are compelled to study the materials being tested, their attention is directed to more essential material, and the test reveals if students have learned and retained the material presented in the classroom (Brown, 1983; Ebel, 1980). Research with undergraduate students by Halpin and Halpin (1982) support Ebel's (1980) assertion that examinations provide powerful

incentives to study. Participants in one condition were told there would be a test on the material presented and in the other condition there would be no test, in which case they were told to study just for the sake of learning. Later, both groups were given performance tests. The students in the test condition performed significantly better than students in the no test condition on a performance test. Students in the test condition rated their effort as higher, reported their achievement level as higher, and they reported a feeling of mastering the material. The authors also reported that studying for a test also appeared to affect retention because there was a significant study effect with the test taking condition that resulted in the retention of learning. Students in the no test condition only reported to have *liked* the class more.

Tests directly affect students' attention to the material. Researchers used tests to motivate procrastinators (Tuckman, 1998). Tuckman found that students who only outlined chapters but were not spot-tested on them scored significantly lower on the final exam than students who studied for periodic spot quizzes. An

additional study found that retesting led to performance increases in a retest option (Juhler, Rech, From, & Brogan, 1998). Although the versions of the test were slightly different, some might argue that the test provided a practice effect. The authors offer though, that the option to retest allowed the learners to get feedback on their performance and revisit the material with an incentive to study it again to improve their understanding rather than just release the information that had been attained. They claim that the improved test scores also provided an immediate reward for the students' efforts to learn the material better.

Tests motivate performance. Additional support for improved student motivation due to a test condition may be described by McClelland's concept of achievement motivation. As defined by McClelland (1951, 1961, as cited in Lawler, 1994), achievement motivation is a desire to perform in terms of a standard of excellence and a need to be successful in a competitive situation. If that tenet is applied to undergraduates, by presenting a challenging task like a culminating assessment, it is expected that students high in achievement motivation will put forth effort to

perform the task. Conversely, with no culminating assessment in place to evaluate their performance, students are less likely to show gains in performance. Research by Pintrich and Schrauben (1992, as cited in Tuckman, 1998) states that "(a) the value of an outcome to a student affects that student's motivation and (b) motivation itself leads to cognitive engagement, with such engagement manifesting itself in the use or application of various learning strategies," leading to improved outcomes.

To motivate students to benefit from a test and direct their learning efforts, they need to know, from the beginning of their coursework, what skills, and abilities are expected to be demonstrated (Loughead, 1997; Astin, 1987; Halpern, et al., 1993). The instruction should support those objectives by teaching how to acquire the necessary competencies to succeed and then the students will have a clear path to educational success. Sebatame (1998) stated the benefits of assessment to students quite clearly:

> "First, assessment directs teachers' and students' attention to particular topics and skills. Second, responding to questions or testing requires active participation on the part of the students, to process the material being assessed. Third, assessment provides practice for students on material, which

> helps to consolidate learning. And finally, assessment can provide feedback that clarifies understanding and corrects misconceptions."

Tests provide closure. Moreover, an additional benefit of comprehensive examinations to students is the feeling that they have *earned* their degree; completion represents a rite of passage or closure (Loughead, 1997; Anderson, et al., 1984). The results provide clarity and quantifiable evidence that they have learned something; a final confirmation of knowledge before entering the professional world providing reinforcement and validation for the effort put forth by a student (Bloom, Madaus, & Hastings, 1981).

In spite of all the benefits to students from testing practices, there still remains the argument that tests are not a valid measure of student abilities. Some opponents of student assessment argue that testing is not a valid representation of what is taught. Tests have been considered "inadequate measures of complex learning procedures and outcomes" (Willis, 1994). They have been criticized for not tapping into higher cognitive processes like how a student organizes information and for testing only the recall of declarative knowledge. Still, others harbor a distrust of an "assessment culture" that has been

more concerned with means than with ends (Broadfoot, 1992, as cited in Willis, 1994). Psychometricians have earned a bad reputation by focusing on validity as a statistical property of the test rather than a verification that the test is representing a construct and is being utilized properly (Messick, 1995; Dwyer, 1998).

By combining the psychometric qualities of "how to test" with the qualitative concerns of educators, a valid test can be created for a fair assessment of student learning. Observing the principles of test theories, steps can be followed to plan a test wisely and to write items that tap into higher levels of cognitive functioning using the Taxonomy of Educational Objectives described by Bloom (Krathwohl, Bloom, & Masia, 1956).

Modern Test Theories

Test theory is the "practical procedures for converting psychological observations into numerical form," (Weiss & Davison, 1981). Most people have participated in some sort of psychological measure, whether it be aptitude tests in school, attitude surveys from manufacturers, or ability tests when applying for a job. The data from these measures are then statistically analyzed through various

methods. Whatever the method, the approach will be based in some form of test theory. Two of the predominant modern test theories are described here: Classical Test Theory (CTT) and Item Response Theory (IRT). A summary of each will follow as well as an analysis of the properties of each.

Classical Test Theory

With a focus on test level information, Classical Test Theory (CTT) considers a pool of examinees and empirically analyzes their collective success rate on a dichotomously scored item (Fan, 1998). Also called classical reliability theory or true score theory, its major use is to estimate the strength of relationship between observed scores on a test and true ability. The goals are to maximize the reliability of a total test score as representing true ability and minimize random and systematic error (Suen, 1990).

Item analyses in CTT provide statistics on item difficulty, item discrimination, and internal consistency. Based on responses from the sample population each statistic provides a piece of information about how items are contributing to the strength of a test. The item

difficulty index is the success rate, or average percentage of the sample who got the correct answer; a higher percentage of respondents getting the item correct indicates an easier item. To discriminate between high and low ability examinees in the sample, a Pearson product moment correlation coefficient is run between the scores on the items and scores on the total test (Crocker & Algina, 1986; Suen, 1990; Fan, 1998). Reliability estimates of internal consistency are established from the covariances between items on the exam to indicate if the items appear to be measuring the same trait.

It is a relatively simple model for test scoring, test development, and item analysis, which makes it easy to apply in many testing situations, and the statistics are simple and easily understood by examinees (Ndalichako & Rogers, 1997; Hambleton & Jones, 1993, as cited in Weiss & Davison, 1981). However, its simplicity lies in its relatively weak theoretical assumptions, making it considered useful for test construction but has been questioned as a methodology (Weiss & Davison, 1981).

An astute observer might recognize and question the reliance of the item statistics on the sample population to

which the items were administered. Specifically, observed score is item sample dependent and the item difficulty and discrimination are examinee sample dependent (Weiss & Rogers, 1981; Suen, 1990). This could prove to be problematic in test development if the sample population performed poorly, resulting in items that appear to be more difficult that they really are. Critics of CTT generally object to this circular dependency, especially that reliability estimates are a function of the particular set of items and a particular sample of individuals on which the data have been collected (Weiss & Davison, 1981). The fact that CTT produces item and person level statistics that can vary across examinee or item samples is the most ardent complaint against the approach (Weiss & Davison, 1981). Second to that, because scores are not standardized, it is difficult to compare scores across exams. However, some researchers have proposed practical solutions to these statistical dilemmas to include test equating and other ad hoc empirical procedures (Fan, 1998), and for test construction it is still commonly applied producing scores with good reliabilities (Suen, 1990).

Item Response Theory

Item Response Theory (IRT), on the other hand, is a theory-grounded approach to test development that focuses on item level development and models the probabilistic distribution of scores with an ogive curve. Also known as latent trait theory, IRT attempts to create item statistics that do not change across examinees, and essentially attempts to estimate examinee ability while holding item characteristics constant. Hence, it is touted to produce

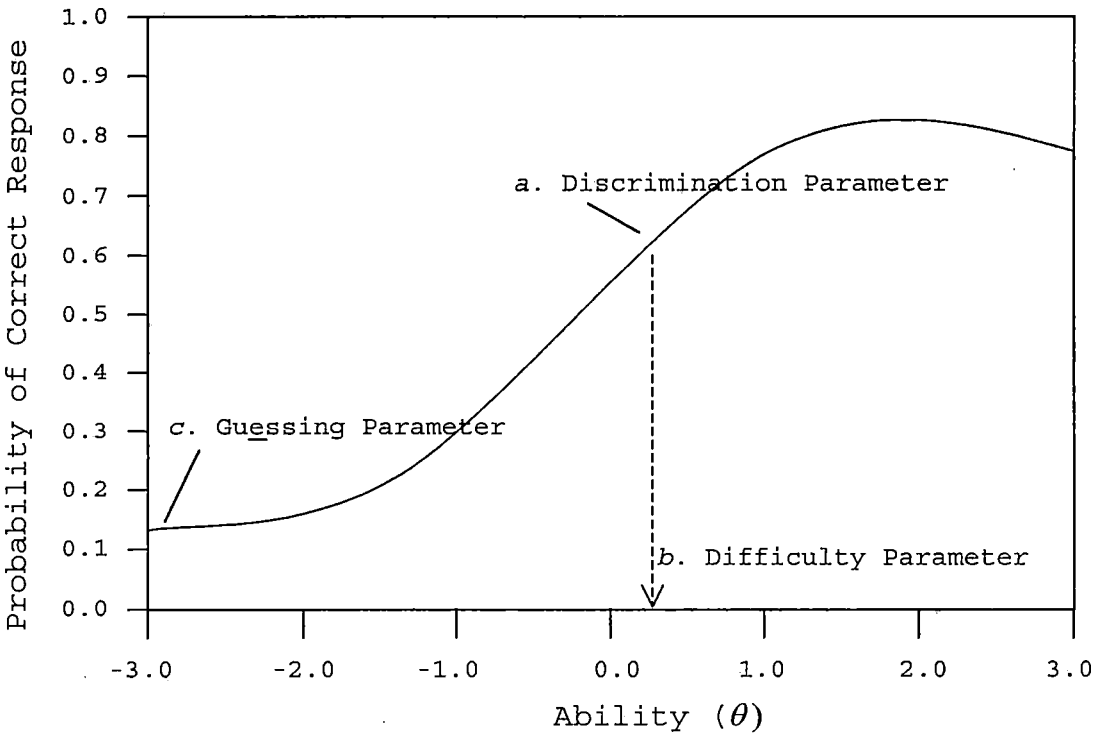


Figure 1.[1] An Ogive Item Characteristic Curve

item statistics independent of examinee samples and person statistics independent of a particular set of items administered. This feature makes it attractive for computer adaptive testing and test equating (Dragow & Hulin, 1990).

Complex, mathematically calculated logistics curves called Item Characteristic Curves (ICC) are used to describe the relationship between the examinee's ability level on the trait being measured by the item and the probability that the examinee will respond to the item correctly (Suen, 1990).

The IRT framework encompasses multiple models. Three models are available for parameter estimation for dichotomously scored items: 3-Parameter model (3-P), 2-Parameter model (2-P), and 1-Parameter model or Rasch model. As seen in Figure 1, The three characteristics of the ogive curve, difficulty parameter (b), discriminability parameter (a), and probability of guessing correctly parameter (c) are calculated using the model with the right fit (Suen 1990).

In simple terms, the relationship between ability (θ) and probability correct is known for each item on a test. If the data fit the 3-P model, estimates of parameters for

guessing, item discrimination, and item difficulty can be made. The 2-P model does not allow for guessing, but can estimate item difficulty and item discrimination, while the simpler Rasch model has constraints on two of three possible item parameters leaving only item difficulty to be estimated. Appropriate model selection is critical because item parameters could be different if estimated under the differing assumptions of the three models. For example, in most multiple choice or true/false tests it is probable that guessing can occur, in this case, the 3-P model would have to be the best fit (Suen, 1990).

The concept that the item characteristics are invariant across samples is accomplished with a linear transformation of various mean sample scores to make the item parameters have the same values. Once these values are the same, it allows differences in ability and random error to show. Conceptually, researchers in IRT assert that difficulty, discriminability and probability of guessing are inherent in the item, examines may perform differently on the item, but the item characteristics don't change (Suen, 1990). In contrast, in CTT, item difficulty and

discriminability indices may vary depending on the sample population to which the exam was administered.

The skeptical observer might ask, "how are the ability and probability correct for the item known in the first place?" This is a very good question, because, to estimate *a*, *b*, and, *c*, one needs to know θ and to estimate θ one needs to know *a*, *b*, and, *c* (Suen, 1990). Interestingly enough, the initial estimate, or item calibration, is based on sample data and the CTT *p*-value (difficulty) and observed score. These two sample dependent statistics are used to estimate up to four sets of parameters for an IRT model. This conversion of less data points into more factors is termed insufficient statistics. To compensate for this apparent shortcoming, a host of complex model fit likelihood estimations are available (Suen, 1990).

Another potential limitation in IRT is the standard error of estimation which is similar to SEM. However, it is a function of θ, so it is ever-changing, resulting in no meaning of a reliability coefficient. Reliability scores are based on a certain ability range best suited for a test. Conceptually, how different is that from being sample dependent?

While IRT has been depicted as alluring in much of the literature because of its multi-faceted functionality, its complexity is a legitimate deterrent. Moreover, in light of what appears to be a common foundation for IRT and CTT for initial item calibration/item statistics, one must decide if the results will be significantly different and if the data and application merit the additional effort of the complexity of IRT. To this end, in extensive empirical studies, Fan (1998) found that the person and item statistics derived from the two measurement frameworks were quite comparable. Likewise, Ndalichako and Rodgers (1997), in their comparative study of five scoring models, found ranking of examinees to be essentially the same and the mean absolute differences revealed good agreement among scores. Both sets of researchers asserted that their findings, plus the simplicity of CTT in test scoring and item analysis support the continued use of CTT for test scoring and item analysis.

Test Development

Test Development Process

Detailed steps in developing a standardized test have been delineated in the literature for others to follow

(Crocker and Algina, 1986; Brown, 1983; Suen, 1990; Drasgow & Hulin, 1990). In brief, the process of developing a test involves (a) planning the test: purpose, content domain, and target audience; (b) determining test specifications: method of assessment, format, and level of difficulty; (c) developing items: constructing a pool of items, pre-testing, selecting, and field testing; (d) analyzing statistical properties: item characteristic indices, reliability, and validity; (e) producing guidelines for administration: normative, pseudo-normative, and interpretive data. Some of these steps are quite critical and complex and are explained in further detail to follow. Specifically, writing items at various levels of difficulty according to an established taxonomy and assessing the statistical properties of a newly constructed exam.

Bloom's Taxonomy

Bloom's taxonomy of educational objectives provides a useful structure for item writing. If items are written utilizing these different levels of cognitive evaluation, an objective test can be developed to tap into deeper levels than merely knowledge recognition (Bloom et al., 1981; Willis, 1994, Suen, 1971).

In further detail of the six levels of the taxonomy, Bloom describes knowledge as the lowest-level category. By the recognition of key words it involves only recall of facts, ideas, or material in a form close to that in which it was originally encountered. Most multiple choice tests are constructed to tap only into this level of cognitive functioning, which has contributed to the discontent with the current testing paradigm (Willis, 1994; Broadfoot, 1992, as cited in Willis, 1994). The next level Bloom describes is comprehension, or use of knowledge. Bloom defines comprehension to be a low level understanding of the meaning or intent of materials or ideas evidenced in translating, interpreting, or extrapolating information without any real integration of information (Bloom et al., 1981; Brown, 1983).

Application is the third level of cognitive evaluation, and it requires the knowledge of abstractions to be understood well enough to demonstrate their use. It is evidenced by performing in new situations by generalizing, organizing, or classifying information (Brown, 1983). The fourth level, analysis, involves the breaking down of elements in a situation and clarifying the

relationships between the pieces. It involves full knowledge, comprehension, and application of the content to be understood (Bloom et al., 1981). Synthesis, the fifth level identified by Bloom, is the combining of elements to create a new pattern or structure (Brown, 1983). This level of evaluation requires divergent thinking and may best be assessed through an open-ended response (Bloom et al., 1981). The highest level, evaluation, is a complex level of cognitive processing. It requires the person to make judgments about an issue, the value of an idea, or the utility of a method, for example, using a set of criteria or standards as a basis (Erwin, 1991). The difficulty in assessing this level of cognitive processing lies in the judgment of the evaluator to decide if the thought process is logical and relevant (Bloom et al., 1983).

If questions are designed to tap into the varying levels of cognitive functioning, the argument that objective tests assess only trivial knowledge can be mitigated (Willis, 1994). Applying Bloom's taxonomy to a comprehensive exam means that students not only will be expected to remember knowledge and facts, but also to demonstrate comprehension and the ability to apply the

knowledge to practical situations. Further, if the exam poses short answer or essay questions, the students would likely be expected to synthesize various types of information into a well-organized set of ideas.

Table 1.[2] Bloom's Taxonomy of Educational Objectives

Level	Cognitive Behaviors
1. Knowledge	To know specific facts, terms, concepts, principles, or theories.
2. Comprehension	To understand, interpret, compare and contrast, explain.
3. Application	To apply knowledge to new situations, to solve problems.
4. Analysis	To identify the organizational structure of something; to identify parts, relationships, and organizing principles.
5. Synthesis	To create something, to integrate ideas into a solution, to propose an action plan, to formulate a new classification scheme.
6. Evaluation	To judge the quality of something based on its adequacy, value, logic, or use.

[2] Table is cited from California State University Psychology Outcomes Assessment Meeting handout, (1999).

Statistical Properties

To determine if a test and its items are functioning as intended, responses from the sample population are statistically analyzed to describe the distribution of responses, to identify inadequate or redundant items, and to determine the reliability and validity of the instrument (Crocker & Algina, 1986; Drasgow & Hulin, 1990). Item analysis procedures from Classical Test Theory will be described here because even in an IRT approach when constructing an exam, the basic analyses used in CTT are still applied. Item analyses of participant responses to items would include mean and variance, item difficulty or proportion correct, biserial and point biserial correlations or item discrimination, and internal reliability or dimensionality. From these item statistics, decisions can be made about item retention, revision, and removal.

Bloom, et al. (1981) suggested guidelines for determining item retention based on the indices of item difficulty and item discrimination. Items that are at the appropriate levels of difficulty and discriminability are retained in the test, while items that are out of the

acceptable range are either revised or removed (Bloom et al., 1981; Ebel, 1965 as cited in Crocker and Algina, 1986; Drasgow & Hulin, 1990). Additionally, when piloting a new test, the indices of item difficulty and item discrimination can be tracked and used as estimates for predicting performance on future tests with similar students (Bloom et al., 1981).

Item difficulty has been defined as the proportion of test takers in the sample population who answer the item correctly. Bloom's guideline for item retention states that if an item is below .4 the item is considered too difficult, .5 to .6 is the mid-range or appropriate difficulty for a standardized test, and .7 to .9 range items are considered too easy because 70% to 90% of respondents answered the item correctly (Bloom et al., 1981). However, these criteria may be adjusted based on the application of the exam. For example if an exam is intended to discriminate between students for admission to graduate school, then items of higher difficulty would be retained (Drasgow & Hulin, 1990).

Item discrimination is an index of how effectively an item differentiated between those who scored highly on the

overall test and those who scored low (Crocker & Algina, 1986). It is measured with biserial and point biserial correlations to show the relationship between an item and a measure of the overall test score (Crocker & Algina, 1986; Bloom et al., 1981). A point biserial correlation is a Pearson product moment correlation between a continuous variable and a dichotomously scored variable that indicates how well an individual did on the item in relation to his or her total score (Crocker & Algina, 1986; Bobko, 1995). If the total score is high, and the point biserial is low or negative for the item, the item should be examined. A biserial coefficient is a correlation between the latent variable being measured and a continuously distributed criterion such as a test score (Bobko, 1995). Biserial correlations, however, assume a normal distribution and may be systematically higher than point biserial correlations, suggesting greater differences between items than really exist. Additionally, the assumption of the normal distribution of the biserial correlation may prove to be problematic if the sample size is not large enough or an assumption of normality cannot be justified (Kottke, Psychology 644, class notes 1998; Bobko, 1995). For item

retention decisions, the index of discrimination follows a general rule of thumb that anything over .20 is considered worth keeping. Additionally, because the discriminability index is a sample population item total correlation, it is considered a normative reference. Therefore, its use with a criterion-referenced exam is limited to essentially assist with decisions about distractors.

Another major consideration for item retention is the criticality to the domain. Some items are crucial to the domain and can't be deleted without compromising content validity (Suen, 1990) even if item difficulty and discriminability are unattractive. If this is the case, the content should be analyzed and the item revised accordingly. Finally, a second iteration of item writing, revision, data collection, and statistical analysis is generally required to ensure a sound instrument (Drasgow & Hulin, 1990).

To determine the integrity of a test, reliability and validity data must be determined based on analyses of the final version of the test. At a fundamental level, reliability can be looked at as the reliability index (observed score of the individual versus the true score)

and the reliability coefficient (stability of scores over multiple administrations)(Crocker & Algina, 1986; Bobko, 1995). To determine the reliability coefficient methods like test-retest and alternate forms are employed.

Correlating results between responses is considered a strong indicator of consistency in scores (Bobko, 1995). However, the correlations indicate relationships that are usually not perfect. In fact, one must try to identify how much error is associated with the relationships. To address this variance between true ability and observed score, standard error of measurement (SEM) can be calculated to describe the expected variation of each individual's observed score and true score averaged for the group. From the SEM, confidence intervals can be identified to provide a range of expected error in scores (Crocker & Algina, 1986; Bobko, 1995).

At the same time, test developers are usually interested in the internal consistency of items when creating an exam (Bobko, 1995). Most programs of item analysis display estimates of reliability for internal consistency. Split-half reliability can also be calculated to determine internal consistency. However, experts caution

against the use of internal reliability as a single measure of the reliability of an exam. Bobko (1995) explicitly argues that split-half measures of internal consistency are useful when constructing an exam to measure a single construct, but should not be implied as the true reliability of the exam even if averaged into a composite like Cronbach's alpha, because reliability should be maintained as a measure of test stability over time.

Interestingly enough, tests can be reliable over multiple administrations, but may not be valid exams. Historically, test developers have identified three distinct types of validity: content, criterion, and construct. It should be stated though, that this trinitarian concept of validity has been vigorously challenged in the literature as being an artificial dissection of a holistic concept (Landy, 1986; Messick, 1995). Whatever terminology for validity that is chosen to be applied, there are factors that vary across applications that should be considered when determining the validity of an instrument. Therefore, for purposes of discussion, the three generally accepted concepts of validity will be discussed further.

Content validity is a measure of the adequacy of the content domain sampling. However it is difficult to fully define the domain, and, because content can be a qualitative notion, it is often difficult to empirically assess. Crocker and Algina (1986) recommend assessing content validity by matching test items with the pre-defined objectives and measuring inter-rater agreement on the mapping. Measures can include the percentage of items matched to objectives, the index of items by congruence, or the percentage of items not assessed by any item on the test.

Criterion validity refers to the relationship between test scores and a criterion, further, it is the standard to which test performance is referenced (Crocker & Algina, 1986). Criterion-referenced validity can be assessed from the perspective of predicting future success or as a measure of concurrent ability; a single approach should be selected depending on the use of the exam. Criterion validity can be used as a concurrent measure of a person's present standing in relation to a criterion, as opposed to a normative comparison. As a predictive measure, great care must be taken to set the cut-off at the appropriate level

to avoid false positives and false negatives. According to Landy (1995), this method of validation is given far too much credence because it places too much emphasis on predictability and not enough focus on understanding what the inference of the score really means, and as a result has been misapplied in legal settings as a sole factor of validity in court decisions about employment. Determining the criterion may be a difficult task that must be agreed upon by the governing body of the test's development keeping the use of the test in mind. Ironically, when developing a test, normative data are often used to establish an expected target for the criterion.

The third, construct validity, is a method of validation that attempts to ensure the test is a valid measure of the construct. Exams are often administered to differentiate between groups (e.g. high ability and low ability students or qualified applicants), where it is obviously critical that the construct has been represented properly by the test to allow inferences made from test scores to have meaning (Crocker & Algina, 1986). It is possible to create a test that shows differences in observed scores between groups that don't really exist in

true ability as a result of poor test construction. It is hard to detect though, because measuring construct validity potentially involves numerous variables that may not all be accounted for. One method suggested by Bobko (1995) was to identify an hypothesized network of expected relationships or a "nomological net". With this approach, correlations can be conducted for all identified variables to see if the relationships point in the anticipated direction and if the pattern of relationships makes sense. These correlations cannot prove that an instrument is measuring what it purports to, but can certainly increase confidence in the measure.

It is readily apparent that these three concepts of validity cannot be mutually exclusive of one another and must all be taken into account when constructing a test (Landy, 1986). Aware of the common distinction between the three types of validity, Messick (1995) argues that test developers need to measure test validity with a comprehensive view. He asserts that construct validity should be defined as the most comprehensive view of the construct defined, the content represented, the criterion determined, and the consequences or social validity

considered (Messick, 1995). He broadly defines validity as "nothing less than an evaluative summary of both the evidence for and the actual--as well as potential--consequences of score interpretation and use (i.e., construct validity conceived comprehensively)." This unitary concept of validity refers to the meaning of the test and the use of the examinees' scores. In application, when designing a test one must keep in mind the use of test, consider context of the assessment, consequences of results, sampling domain, population, construct representation, content validity, criteria, and predictive and concurrent validity as all being factors into the measurement of validity (Messick, 1995; Camera & Brown, 1995, Landy, 1986). In light of the potential social implications of an outcomes assessment for college undergraduates, the validity of the test must be evaluated with all of these factors in mind.

California State University at San Bernardino Psychology Department Outcomes Assessment Proposal

Not being immune to any of the stirrings in the academic arena, the California State University, San Bernardino (CSUSB) Psychology Department is in the process

of constructing an Assessment of Educational Outcomes tool. The department would like to determine if, "students are learning what they need to know when they leave," Diane Halpern, former chair of the Psychology Department commented. The tool will be implemented to provide feedback to the department and to the students on the breadth and quality of Psychology education provided by CSUSB. The completed assessment tool will be composed of six separate assessments: course requirements, exit interviews, alumni surveys, professional development experiences, student peer reviews, and a culminating assessment.

As Halpern (1998) suggested, CSUSB has been collaborating with other California State Universities (CSU) to identify knowledge, skills, and values unique to Psychology and to operationalize these characteristics to develop goals for Psychology academic programs (CSU Psychology Outcomes Assessment Meeting, California State University, Bakersfield, 1999). Goals common to most Psychology departments have been identified and are now beginning to be applied at various campuses within the CSU system. The program at CSUSB was designed by a faculty committee, who identified twenty-two objectives specific to

its curriculum grouped under six major goals, for the instruction of its students. Please see Appendix A for a complete listing of "Goals and Objectives for Psychology Majors." To assess the information they seek about the value added from the department, the faculty has proposed local development of multiple methods of assessment to be delivered through the efforts of faculty and staff members. They will incorporate multiple methods of assessment as recommended by Curry & Hager (1987).

One method of assessment within the outcomes assessment process is the completion of a locally developed comprehensive exam to assess the understanding of the core content and application of Psychology (Ebel, 1980). For ease of administration, the culminating assessment tool will take the format of a multiple choice examination, with a proposal to develop short answer items to tap into higher levels of cognitive functioning. The items are to be developed by faculty and other subject matter experts to allow consistency between the curriculum and the objectives of the department (Ewell, 1987; Banta & Moffet, 1987; Curry & Hager, 1987). The culminating assessment is being developed in order to assess student understanding of

Psychology core areas of knowledge and application, as well as understanding of personal and academic skills necessary for proper application, such as research design, statistics, presentation, writing, and ethical considerations. The results from this exam will provide an additional measure of student knowledge that should correspond with the behaviors that are observed in the classroom in accordance with the curriculum requirements of goal 1. The primary purpose of the CSUSB culminating exam is for it to be a diagnostic tool to provide feedback regarding the quality of education. By using a quantifiable measure, the department can gauge the degree to which it is meeting its own stated educational goals and objectives for its students and eliminate assumptions that may be inaccurate regarding student learning. The test can also be used to provide feedback to students about their learning, but great care should be taken to not use the results in a punitive way or to place blame when using the exam as a source of evaluation, grading, or minimum requirement gateway test for students (Banta & Moffet, 1987).

CSUSB has elected internal development of the culminating exam because it allows test development to be

consistent with the objectives of the curriculum, encourages faculty involvement (Pennycuick, 1990 as cited in Willis, 1994), and costs less to administer (Ewell, 1987; Banta & Moffet, 1987; Curry & Hager, 1987). External tests have been criticized as not representative of a specific curriculum, making them inappropriate for a curriculum review (Sheehan, 1994; Ewell, 1987). The use of the Graduate Requirement Exam-Psychology Subject Test (GRE) had been considered by CSUSB, but it is intended for graduate school selection and best discriminates at the upper range of student ability. Additionally, the norms provided are for graduate school bound students (Curry & Hager, 1987). And, although scores could be compared across institutions, the College-Level Examination Program (CLEP) was dismissed because it is too long and costly to administer and the scoring reports do not provide enough detail to allow curriculum section evaluation (Ewell, 1987). The ACT is for college admission, not for assessment purposes. The ACT College Outcomes Measures Project (COMP) is not a good measure of what is being taught within a specific curriculum, so also not appropriate for curriculum evaluation (Curry & Hager, 1987).

The six methods of assessment have been designed to evaluate the attainment of the multiple goals and objectives at CSUSB. As recommended in the literature, the faculty involvement was encouraged and a faculty committee clearly delineated how each objective would be most appropriately addressed with curriculum requirements (Willis, 1994; Halpern, 1987). See Appendix A, Table 2 to view a matrix of the relationship between course requirements and individual objectives. Appendix A, Table 3 shows the relationship between the remaining assessment tools and individual objectives. The goals and objectives will also be described below with their relationship to the culminating exam. Because these assessment methods have not yet been used at CSUSB, there is an expectation that they may need to be revised after piloting to better meet the needs of the department, as suggested by Ewell (1987).

The outcomes assessment is a complete assessment tool where each method of assessment will be developed and validated according to the needs it should be fulfilling individually, and as part of the outcomes assessment program (Ewell, 1987). The focus of this project is to develop the culminating assessment portion of the

Psychology major outcomes assessment. A description of all the methods within the Outcomes Assessment program at CSUSB are described in Appendix A. Please review to understand the context in which the culminating examination was developed and will be used.

In summary, the review of the literature included the exploration of Outcomes Assessment as a legitimate means to measure student learning, the merits of the comprehensive exam as part of an outcomes assessment, and the practical applications of such a practice in university environments. It continued to describe the process of how to develop a specific exam for a university including the comparative analysis of two modern test theories, steps in test development, analyses of psychometric properties and considerations for determining validity of the exam in the given context.

Objective of this Project

A comprehensive multiple-choice examination was designed to assess knowledge of graduating Seniors of the Psychology Department at CSUSB. Consistent with the experience of other universities in the literature, initial applications for the exam were to use it as a diagnostic

tool to determine strengths and weaknesses in the curriculum, instruction, and student learning (Banta & Moffet, 1987, Halpern, 1987, 1988; Ewell, 1987). An additional benefit of the initiative is the ability for educators to enhance the curriculum based on the test results, since the curriculum drives and promotes the standards of performance to which students and professors can aspire. Consistent with recommendations in the literature (Curry & Hager, 1987; Ewell, 1987), the culminating assessment was designed for the Psychology Department to be one method of assessment within a multi-method framework for assessing educational outcomes for students (see Appendix A for Assessment of Educational Outcomes for Psychology Majors at California State University, San Bernardino). The development of the exam adhered to the guidelines presented in the literature section to allow the development of a test that is fair, representative of what is taught, and able to assess the complex levels of learning acquired by participants in the major. Moreover, the test content was designed to reflect the content domain of the educational objectives of a Psychology undergraduate degree and constructed to be

practical in application, with simple administration and scoring.

There are two parts of this study. One is the development of a reliable and valid comprehensive exam appropriate for the target audience of graduating Seniors in the CSUSB Psychology Department. The other is the evaluation of student performance on the exam as it relates to ability, content objectives, and the curriculum of the major. An exam was developed, piloted and revised to evaluate the soundness of the exam and the corresponding student performance.

CHAPTER TWO

PILOT STUDY

Introduction

A pilot study was conducted to assess the properties of the exam. From the results of the pilot exam it was expected to find the need for revision of the exam items, as is standard procedure in test development (Crocker & Algina, 1986; Drasgow & Hulin, 1990; Suen, 1990). Moreover, it was expected to find evidence of some of the gains in student knowledge due to the Psychology curriculum in which they have participated at CSUSB (Willis, 1994; Lien, 1971; Krueger & Heisserer, 1987). Hypotheses were formulated on around expectations in student performance based on studies done at other universities like Trenton State College (Curry & Hager, 1988).

Hypotheses for Pilot Study

Student Performance

Hypothesis 1. Academic major: Majors in Psychology will have significantly higher scores than non-majors.

Hypothesis 2. Grade level: Seniors in Psychology will score significantly higher than Freshmen, Sophomores, and

Juniors in Psychology.

Hypothesis 3. Psychology courses: Students who have taken more Psychology classes will score significantly higher on total test score than students who have taken fewer Psychology courses; test scores should correlate positively with the number of Psychology courses completed.

Hypothesis 4. GPA: Students with higher grade point averages (GPA) will perform significantly higher on the test than students with lower GPAs.

Method

Participants

Participants were 94 college students at CSUSB, four cases were removed from the analyses because of insufficient demographic information. 85% of participants were female, the remaining 15% were male. The mean age was 27.41 with a range of 19 to 54. Participants reported a mean GPA of 3.2 with a range of 2.0 to 4.0. They also indicated the number of Psychology courses taken with a mean of 4.86 and a range of 1 to 13. See Appendix B, Table 4 for Demographics of Sample Populations. Information on academic major and grade level was reported inconsistently

due to formatting errors in the demographics sheet (see Appendix D to view the instrument).

Test Development

A one hundred item multiple choice examination was planned as a portion of the overall Outcomes Assessment initiative for CSUSB Psychology Department. It was administered to university students in a classroom setting. Volunteers were given a ninety-minute timeframe in which to complete the exam, an informed consent, and a demographics sheet (see Appendices F & G). A debrief statement was also given to each participant when he or she completed the assessment (see Appendix D).

To aid in the construction of a valid test for the Psychology Department, the outcomes assessment faculty committee sought the help of the CSUSB Industrial-Organizational Psychology program graduate students whose expertise in measurement, test development, and validation within the guidelines of psychological principles would help keep the costs of test development down. The exam was constructed according to process of test development outlined in the literature section (Crocker & Algina, 1986; Drasgow & Hulin, 1990; Suen, 1990) and special

consideration was given to meeting professional standards outlined by the American Psychological Association.

Planning the Test. The purpose of the exam was identified as the need to assess knowledge of students graduating from the Psychology undergraduate program in order to identify strengths and weaknesses of curriculum, instruction, and student learning. A complete list of the objectives of the content domain and an appropriate sampling of that domain were specified to provide the basis for a pool of items to be developed. The domain was defined in the department objectives to include the fields of Psychology in terms of facts, concepts, applications, integration of ideas, and implications. Criteria were outlined by the department faculty outcomes assessment committee in the Assessment of Educational Outcomes for Psychology Majors at California State University, San Bernardino (see Appendix A). The test was constructed using these content objectives as the definitive guide.

Test Specifications. After the content, population, and use were identified, decisions were made about the most appropriate method of assessment, type of format, and desired item level and difficulty. A multiple-choice

specific criteria identified as the sampling domain of the test. Many core knowledge items representing the fields of Psychology and the objectives expected to be addressed by the culminating exam were gathered by using actual exam items found on midterms and finals in the various courses offered within the department.

Different types of test items were created to include direct questions, incomplete statements, completion, negative ordering, and combined response (Bloom et al., 1981). More items than needed were written to allow the best items to remain in the final version of the test. The goal of item writing, as recommended in test development practices, was to keep items independent, cover important material, and write questions simply and clearly. More specifically, the stem of each question was aimed to clearly pose a single problem and be written concisely with enough information to make only one of the alternatives correct. Distractors were written to be plausible, but demonstrably incorrect or less correct then the key answer (Psychology 644, 1998, class handout) to see if students have learned the material, with no intent to confuse them or trick them into picking the wrong answer (CSU Outcomes

Assessment Meeting, 1999). Moreover, the purpose of distractors was to attract students who have not mastered the objective, while correct answers should have been obvious to students who have mastered the objectives.

Items were also checked for accuracy, clarity, relevance, quality of distractors, grammatical correctness, appropriate reading level, possible bias of the items, duplication, and item interdependence (Crocker and Algina, 1986).

Pilot. Items were piloted with a representative sample and item analyses were conducted following the approach of Classical Test Theory. Statistical properties of the items were identified to make decisions for revision and removal based on Bloom's guidelines. Plans were made to revise the exam with the results from the pilot, the field test, and the reliability and validity data.

Statistical Properties. With the small, local target population for the comprehensive exam, tailored to the CSUSB specific curriculum, and with no immediate intentions of a computer application, CTT was chosen as the appropriate test theory to follow for the development of the exam. Participant responses were computed using the

using the ITEMAN program (Assessment Systems Corporation, 1995). Its dichotomous scoring graded items correct or incorrect. The output included item number identification, proportion of correct responses, item discriminability, point biserial correlations, and an evaluation of item statistics. Fifteen descriptive indices were given including number of examinees, central tendency and distribution, alpha, average item-scale correlations, and scale intercorrelations. Although other research warns that the results from this type of item analysis yield discrimination and difficulty indices that are sample specific (Donnelly, 1994), it was deemed acceptable for the application in the specific university environment and local target population. From the indices of item difficulty and item discriminability, decisions to revise or remove items were made.

Results

Test Development

Statistical Properties. Item analyses revealed the range of difficulty, or percentage correct by item was .10 to .97. Based on Bloom's criteria (Bloom et al., 1981) fifteen of the one hundred items were identified as being

too easy while eleven were too difficult. Please see Appendix B, Table 5 for a grouped frequency chart of the item distribution into the levels of difficulty recommended by Bloom (Bloom et al., 1981). Item discriminability ranged from .03 to .67, with a median of .34. Point biserial correlations were corrected for spuriousness and revealed a range of -.16 to .58 with a mean item total of .27 and a mean biserial of .37. Please see Appendix B for item by item determinations based on the item analyses.

Reliability coefficients were provided from item analyses revealing an overall alpha of .90, indicating a high degree of internal consistency. Reliability analyses for the sub sections of the test were conducted, representing each objective, ranged from .02 to .75 (see Table 6, Appendix B). Objective 13, student understanding of individual differences and Objective 20, student understanding of the impact of their own behaviors had the lowest alphas. Concerns about the meaning of these reliability coefficients are addressed in the discussion section of this study.

Student Performance

Descriptive Statistics. The mean score for the exam was 59 with a range of 27 to 86. The standard deviation was 13.65 and a normal distribution was evident (see Table 2 for Whole Test Statistics).

Hypothesis 1. Academic major: Hypothesis 1 proposed that majors in Psychology would have significantly higher scores than non-majors. This hypothesis was not measured due to insufficient demographics information.

Hypothesis 2. Grade level: Hypothesis 2 proposed that Seniors in Psychology would score significantly higher than Freshmen, Sophomores, and Juniors in Psychology. This hypothesis was not measured due to insufficient demographics information.

Hypothesis 3. Psychology courses: Hypothesis 3 proposed that students who have taken more classes in Psychology would score significantly higher on total test score than students who have taken fewer Psychology courses. Correlations were examined between the demographic information and the exam scores. With a correlation of $r = .55$ ($p < .001$), the results were significant in support of Hypothesis 3.

Hypothesis 4. GPA: Hypothesis 4 proposed that students with higher GPAs would perform significantly higher on the test than students with lower GPAs. Students who reported higher grade point averages also had higher total scores on the exam ($r = .55$, $p < .001$).

An ad hoc partial correlation between number of Psychology courses taken and total test score, controlling for GPA, indicated a relationship of $r = .61$ ($p < .001$). Additionally, when GPA was correlated with overall test score and self-reported number of courses taken was controlled for, a correlation of $r = .58$ ($p < .001$) was seen.

Discussion

Test Development

Statistical Properties. Item discriminability was evaluated for each item and it was discovered that some items were very discriminating between high and low scorers, while other items were too easy and some distractors appeared to be plausible. These results indicated the need for rewording, more difficult distractors, or simply removal. Additionally, the item analyses revealed a few items that may have been too

difficult for the sample population used. Some participants commented that the questions seemed too specific. Overall results looked as if that the test items were fairly attainable and appeared to assess primarily the knowledge level of Bloom's the educational objectives. The main study aimed for higher levels of cognitive evaluation to better assess student learning. Additionally, the main study aimed to include additional upperclassman undergraduates who have had more courses in Psychology, to better evaluate if the items were the appropriate level of difficulty for the target population.

After reviewing the results, one other consideration came to mind. The structure of the answer scheme allowed for random guessing which could artificially inflate scores by allowing for up to 25% of correct responses, according to laws of probability. Results like this may provide misleading feedback for curriculum development, so tendencies for guessing were looked into in the main study.

The high measure of overall internal consistency had not been expected. The researchers had not been certain that the exam was a homogeneous composition of a single construct because the test is composed of a broad range of

concepts in Psychology. Moreover, higher reliability coefficients had been anticipated within each objective as opposed to between the objectives. However, an inspection of the covariance matrices suggested the magnitude of the reliability may have been more influenced by the length of each sub-test rather than the covariance between the items. Objective 13, student understanding of individual differences and Objective 20, student understanding of the impact of their own behaviors had the lowest alphas, most likely because each scale only contained two items. Some of the more complex objectives were difficult to measure with the multiple choice format and may not be appropriately measured in this portion of the OA process.

Student Performance

Descriptive Statistics. The quality of the descriptive statistics and the resulting hypothesized correlations would have been much better had the demographics sheet been clearer and more thorough.

Hypothesis 1. Academic major: Hypothesis 1 proposed that majors in Psychology would have significantly higher scores than non-majors. This hypothesis was not measured. The attempt to determine if the students were Psychology

majors in the demographics section of the pilot study was compromised because of the awkward formatting of the fill in section on the demographics form leading to many omitted responses.

Hypothesis 2. Grade level: Hypothesis 2 proposed that Seniors in Psychology would score significantly higher than Freshmen, Sophomores, and Juniors in Psychology. This hypothesis was not measured. Grade level was not clearly requested on the demographics form, thwarting the ability to test hypotheses 2. Although there was incomplete demographic data to perform these analyses accurately, the fairly high scores on the exam and researcher's knowledge of the pilot sample group make it possible to assume that the pilot group closely resembles the target population, so results should be generalizeable.

Hypothesis 3. Psychology courses: Hypothesis 3 proposed that students who have taken more classes in Psychology would score significantly higher on total test score than students who have taken fewer Psychology courses. The fairly strong correlations between test score and number of Psychology courses taken ($r = .55$) supports the assertion that information was learned, and might be

interpreted as a measure of value added by the Psychology department, in support of hypothesis 3.

Hypothesis 4. GPA: Hypothesis 4 proposed that students with higher GPAs would perform significantly higher on the test than students with lower GPAs. The strong correlations between the test score and grade point average ($r = .61$) support hypothesis 4. Students who did well on some items did well on many others also, possibly representing an underlying ability. This notion is further supported by the unexpected high alpha.

Limitations of the Pilot Study

In addition to some of the shortcomings listed previously in the discussion, there were a few errors in test construction. For example, item number nine had an asterisk next to the correct answer, resulting in ninety-three percent of respondents getting the item correct. A few typographical errors were also reported. Students gave feedback that there were some questions they considered too long or too detailed. Had the test development process followed the procedures outlined in the literature section more rigorously, these errors may have been prevented (Crocker & Algina, 1986; Drasgow & Hulin, 1990).

Nevertheless, some students commented that they didn't realize how much information they had been exposed to, nor how much they had forgotten, and a few were very interested in knowing their scores on the exam, consistent with research by Ewell (1987) and Juhler, et al. (1998). Further, as a tool for the university, if students are aware of an outcome assessment as part of their curriculum, it is believed that they will strive to retain more of the knowledge they are exposed to as undergraduate students of CSUSB (Bloom, et al., 1981; Juhler, et al., 1998; Loughhead, 1997).

Despite numerous shortcomings, the findings of this pilot study were encouraging; there was progress made towards developing a fair outcomes assessment for Psychology undergraduates that is representative of the curriculum offered as well as the stated learning objectives of the Psychology Department.

CHAPTER THREE

MAIN STUDY

Introduction

The pilot served as a pretest to remove or modify poor questions. As planned, a revision of the items from the first exam was conducted as the main study. The revisions included corrections and qualitative enhancements to the items and the inclusion of more items that fit the criteria of Bloom's taxonomy of educational objectives (Bloom et al., 1981; Lien, 1971; Brown, 1983). Based on the results from this pilot administration, changes were made to the instrument. To start, the demographic information sheet was revised to be clearer in the information requested from students. For example, academic major/non-major was clearly indicated, grade level was requested, and course numbers were listed next to course titles when asking which classes students have completed.

Typographical errors on the exam were corrected. Formatting was improved, items were revised and removed. Additionally, to account for guessing without statistical corrections and to provide better feedback for curriculum enhancements, an 'E' option was added to the exam, giving

the participants the option to express "I have not been taught this material". This option was expected to reduce guessing among novices and to provide more accurate feedback in terms of curriculum.

The revised exam was also administered to a larger student population and to more Psychology majors to ensure it had appropriate levels of item difficulty and discriminability and to establish validity and reliability coefficients. Since the exam had one hundred items, the new sample consisted of more than 500 participants, following Nunnally's rule of thumb to have five to ten times as many participants as items (Nunnally, 1967, as cited in Crocker and Algina, 1986).

The use of Item Response Theory was researched further. While the merits of the approach are relevant to some portions of the project as well as to potential future adaptations of this application, there is not sufficient benefit over Classical Test Theory and its simplicity for the current study for finalizing the construction of the instrument.

Assuming the revisions made to the exam would strengthen it as a valid assessment of student, department,

and curriculum performance, the focus then shifted to improving hypotheses around student performance. The main focus of an OA is to measure student gains from the program and to measure program effectiveness (Halpern, 1987, 1988). Two obvious methods to assess this are to compare scores across grade levels and between Psychology majors and non-Psychology majors. Surprisingly, one study indicated that raw scores between Seniors and Freshman were not as great as expected (Curry & Hager, 1987), and none of the literature found specifically compared majors to non-majors to identify areas of common knowledge. Due to invalid demographics data these were both areas that were unable to be tested in the pilot. Special efforts were put forth to ensure sufficient data was provided to allow these analyses to occur in the present study. Despite results at Trenton State University, where differences between raw scores of Seniors and scores of Freshmen were much smaller than they expected (Curry & Hager, 1987), and lack of conclusive results from the pilot study, differences are anticipated between scores of Seniors and Freshmen at CSUSB. Additionally, when testing within a particular curriculum, speculation may arise about the uncertain results of

transfer students. Trenton State University found no significant differences between transfer students and native students. Since CSUSB does have a significant transfer student population, any differences would be interesting to note. This relationship was not tested with the pilot but was better addressed with the main study.

The present study focused on two major issues: (a) improvements made to revise the test in terms of item construction, reliability, and validity and (b) student performance as an indication of knowledge learned and strength of the CSUSB curriculum. To evaluate test construction, a validity approach was taken and more specifically, the study proposed the following six hypotheses on student performance.

Hypotheses for Main Study

Student Performance

Hypothesis 1. Academic major: There will be significant differences in mean test scores among Psychology majors and non-Psychology majors.

Hypothesis 2. Grade level: There will be significant mean differences of total scores among grade levels.

Hypothesis 3. Freshmen: There will be no difference among scores of Freshmen students; Psychology major or non-Psychology major.

Hypothesis 4. Transfer students: There will be no significant differences between the scores of transfer students and native students.

Hypothesis 5. Bloom's taxonomy: Students who have had more Psychology courses will get more items right in higher levels of Bloom's taxonomy of comprehension and application.

Hypothesis 6. "E" option: An inverse relationship is expected between the number of Psychology courses taken and the use of the "E" response indicating the examinee had not been taught the material.

Method

Participants

Participants were 521 college students at CSUSB recruited from Capstone classes to get students from a variety of academic majors and grade levels. Of the 493 demographics sheets where academic major was filled in, 34.6% were Psychology majors. The remaining 65.3% were grouped as non-Psychology majors. 501 participants reported

their grade level: 29.3% were Freshmen, 20% were Sophomores, 25% were Juniors and 25.7% were Seniors. Of the 428 participants that indicated their gender on the demographics form, 77% were female and 23% were male. The ages of the participants were measured in discrete categories, in brief, 47% were 16 to 20 years old, 17.1% were 21 to 23, 16.7% were 24 to 30, and 19.3 percent were over 30. Participants also identified the Psychology courses they had taken with a mean of number of courses of 1.78, a range of 0 to 12, and a standard deviation of 2.73 (there were 276 cases that did not list themselves as taking any Psychology courses, even though some of them were Psychology majors). See Table 4 (Appendix B) for demographic charts.

Test Development

Test Revision. As a result of the pilot exam, 15 items were revised, 46 we retained as is, and 39 removed. Decisions to revise or remove items were based on inappropriate level of difficulty, results from Spearman-Brown prophecy on item reliability, tricky wording, duplication of content, and inadequate representation of the content domain. Attempts to utilize Bloom's taxonomy

more fully were also made. Five raters, all who possessed a Masters degree or Ph.D level education in Organizational Psychology, divided items into Bloom's taxonomy with inter-rater agreement of at least 60% per item (see Appendix B, Table 5). The items were also categorized into a matrix of subscales (areas of content, department objectives, and goals) based on SME submissions (faculty and graduate students). Please see Appendix B, Table 8 for the content categorization by objective.

Test Administration. A one hundred item paper and pencil culminating assessment was revised and field tested with university students over multiple administrations. All participants were asked to complete a demographics sheet, informed consent, and the one hundred item multiple choice exam within a ninety-minute timeframe (Appendix D). After the assessment each participant received a debrief statement of the study (Appendix D). This form of the exam represents the final version of the exam after revisions from the pilot. Please see results from pilot study.

Reliability and Validity Analyses. A greater focus was placed on determining reliability and validity for the final version of the exam in terms of the context and

application. Specifically, the pattern of relationships in student performance was evaluated and used as a validation approach to the test's development.

Results

Test Development

Statistical Properties. Item analyses were conducted in the Statistical Package for Social Sciences (SPSS) on items with respect to level of difficulty (percent correct) and item-total correlation (point biserial). The range of item difficulty for the total main study population was .09 to .74, and the point biserial results were .07 to .58. Item statistics from the revised exam were compared to item statistics of the pilot exam (please see Appendix B).

Rater categorization of the items into Bloom's taxonomy resulted in 49 items in level 1, 38 items in level 2, and 13 in level three. Comparison of these results to a post hoc categorization of the pilot exam items yielded a very similar distribution (Appendix B, Table 5).

Test reliability and subscale reliability estimates were conducted to determine internal consistency of the items. The overall alpha was .95 and the subscale reliability coefficients ranged from .49 to .91 (Appendix

B, Table 6). These reliability coefficients should be taken with caution, however, because internal inconsistency is necessary but not sufficient information in test development. Concerns are addressed further in the discussion section of this study.

Comparison of confidence internals revealed overlap in the range of scores between the two exams, but only for Junior and Senior Psychology majors. When all participants from the main study were included in the comparison, the exams were much less comparable. See Figure 2.

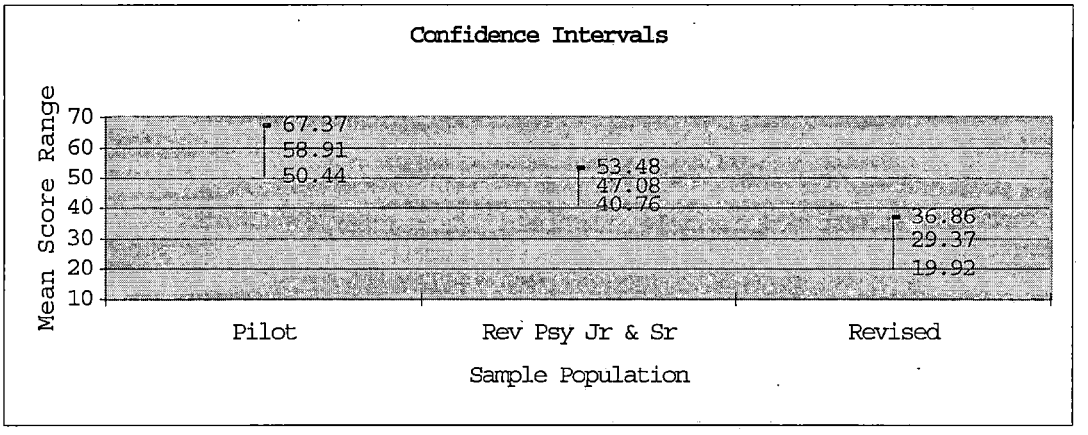


Figure 2. Confidence Intervals Between Sample Populations

Content validity was evaluated by matching exam items to CSUSB objectives to ensure all content areas were addressed in the exam. Inter-rater reliability was not assessed for this because most items were submitted by

faculty members as representing identified objectives. Objective criteria of performance have not yet been established for this assessment.

Construct validity was assessed through the expected relationships stated in the hypotheses. The relationships between test scores and Psychology majors, grade level, Bloom's taxonomy, and use of the "E" option were considered the nomological net necessary to evaluate patterns in relationships and create confidence in the exam's ability to measure what it was purporting to measure (Bobko, 1995). Results from the relationships examined were listed within each hypothesis.

Student Performance

Descriptive Statistics. The mean score on the exam for all participants was 29.39 with a range of 1 to 88 and a standard deviation of 16.38. The mean score for Psychology majors was higher at 39.07, with a range of 2 to 88 and a standard deviation of 18.01 (see Table 2 for whole test statistics and Appendix C, Table 13 for comprehensive comparisons of score means and medians by grade level, academic major and transfer status).

Table 2. Whole Test Statistics: Pilot and Revised Exam

Whole Test Statistics	Pilot	Revised Psych Jr & Sr Only	Revised All Participants
N IN GROUP	90	108	521
MEAN	58.91	47.08	29.39
MEDIAN	27	41	27
STANDARD DEVIATION	13.65	13.16	16.38
MINIMUM	27	2	1
MAXIMUM	86	88	88
COEFFICIENT ALPHA	.90	.94	.94
SEM	4.32	3.22	4.01

Hypotheses 1-3. Academic major and grade level: A 4 x 2 between-subjects ANOVA was conducted in SPSS to evaluate the effects of academic major and grade level on overall test score on the culminating exam. The means and standard deviations for test score as a function of academic major are presented in Figure 3. The results for the ANOVA indicated a significant main effect for academic major, F (1, 480) = 101.00, $p < .001$, partial $\eta^2 = .17$, a significant main effect for grade level, F (3, 480) = 79.97, $p < .001$, partial $\eta^2 = .33$, and a significant interaction between grade level and academic major, F (3, 480) = 11.28, $p < .001$, partial $\eta^2 = .07$.

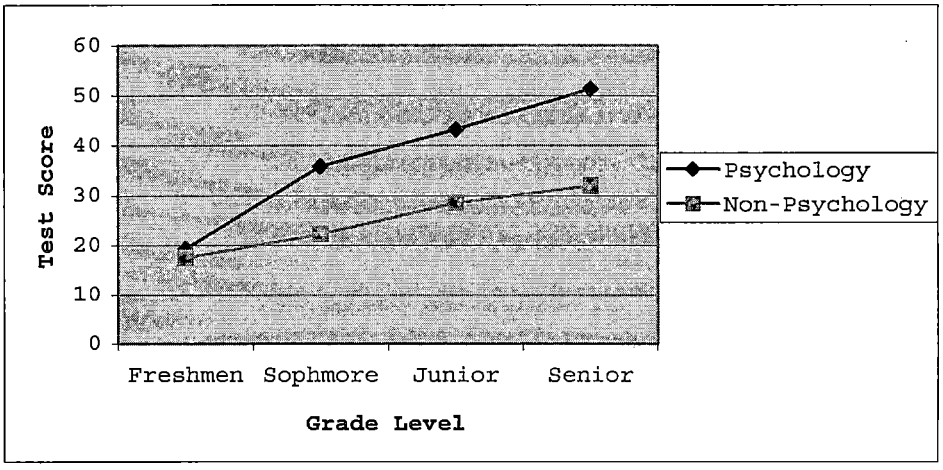


Figure 3. Changes in Mean Score by Grade Level and Academic Major

Because the interaction between grade level and academic major was significant, grade level main effect was ignored and instead the grade level simple main effects were examined. To control for Type I error across the four simple main effects, alpha for each was set at .01 (.05/4 = .01). There was no significant difference in test scores between academic major for Freshmen, F (1, 480) = .63, p = .427, partial η^2 = .001, but there were significant differences between academic major and grade level for Sophomores, F (1, 480) = 19.22, p < .001, partial η^2 = .038, Juniors, F (1, 480) = 45.46, p < .001, partial η^2 = .087, and Seniors, F (1, 480) = 77.39, p < .001, partial η^2 = .139.

Hypothesis 4. Transfer students: A one-way analysis of covariance (ANCOVA) was conducted. The independent variable, transfer status, included two levels: non-transfer and transfer. The dependent variable was the score on the culminating exam and the covariate was grade level. The homogeneity of slopes assumption yielded a significant interaction, F (1, 410) = 4.12, $p < .05$, partial η^2 .01. A significant interaction between the grade level and transfer status suggests that the differences in groups on the dependent variable vary as a function of the covariate and the results from an ANCOVA would not be meaningful and should not be conducted.

However, because the effect size (η^2 .01) was so small, it was decided to run the ANCOVA anyway. Because the assumption was not met, results from this analysis should be interpreted with caution. The ANCOVA was significant, F (1, 411) = 6.64, $p < .01$, partial η^2 .02, indicating a mean difference in test scores among transfer status after holding grade level constant (see Appendix C, Table 10). Standard mean and median scores for transfer students by grade level and by academic major can be viewed in Appendix C, Table 13.

Hypothesis 5. Bloom's taxonomy: Most of the exam items were categorized into Bloom's level one and level two by SME raters, with only a few items in level three (see Appendix B, Table 5). Correlations were computed to identify relationships between number of Psychology courses taken and student ability to answer questions correctly in higher levels of analysis within Bloom's taxonomy. The relationship for level one was r. = .33 and for level two r. = .42.

Hypothesis 6. "E" option: The mean use of the "E" option for Psychology students was 23.5, with a standard deviation of 23.3. Bivariate correlation coefficients were computed between the "E" option usage and the number of Psychology courses taken reported by participants. With a correlation of -.13, the results of the correlation were statistically significant at $p < .01$.

Discussion

Test Development

Statistical Properties. Item analyses were computed to demonstrate item difficulty and point biserial. A comparison between item statistics on the pilot exam and the revised exam can be reviewed in Appendix B. In brief,

the item statistics don't show a great improvement in terms of difficulty, discriminability, or utilization of Bloom's taxonomy. Although items were revised, there was still no improvement in assessing higher levels of cognitive functioning. It is probable that adding short answer questions would be a constructive step towards better assessing Bloom's level 3 through 5 (Bloom et al., 1981).

What was improved, and would only be evident in a qualitative review of content validity, is the better coverage of the broad content domain of Psychology. More items were added to better cover the fields of Psychology and items that were duplicative were removed from the exam. However, Brown (1983) had cautioned against adding items just to represent the content domain, rather they should be included only with the intent to measure the attainment of a particular objective. As seen in the subscale reliabilities and the limited number of items within some objectives, this suggestion could have been heeded more.

Internal consistency was examined for the entire exam and within content subscales. The overall alpha was .95, suggesting high internal consistency for the instrument. However, while determining internal reliability is

necessary for constructing an exam, it is not sufficient (Bobko, 1995). Reliability coefficients should be examined over time and multiple administrations of this exam with only the target population.

Subscale reliability coefficients ranged from .49 to .91 (please see Appendix B, Table 6). The analysis of the subscales posed a few qualitative questions though. When reviewing the items that were submitted to meet the criteria of certain content subscales as identified by the CSUSB Objectives, many questions appeared appropriate to fit into multiple objectives (which might also explain the high overall alpha). For example, question ninety-seven on the revised exam asks: "The tendency to assume that people that have one positive characteristic must have others as well is called..." The content of this question falls into the area of Social Psychology or a field of study in Objective 2. It could also be placed in objective 19, which covers accurately perceiving the behavior of others. With items like this, the decision was made to place them in the less broad categories outside of Objectives 1 and Objectives 2. As a result, subscale reliabilities were run again at the goal level according to CSUSB criteria. Even

then, it was questioned as to whether the objectives were mutually exclusive to the goals.

Within goals, certain objectives might fit better together than was outlined in the CSUSB Objectives. For example, Objective 4, statistics and data analysis was listed as part of Goal 1, Basic processes, methodology, fields, and applications of Psychology. However, it seems to have a better fit in Goal 2, Intellectual and technology skills, where there is another objective on understanding tables and graphs. Had the data collected not been dichotomous, perhaps a factor analysis could have been conducted to provide further clarity for the division of the goals and objectives. It's likely that there was cross contamination across objectives and goals in terms of the subscale measures. Not withstanding, Psychology has a large and broad content domain, and the CSUSB objectives did provide a structure in which to begin to assess the relevant areas as represented in the program's curriculum.

Standard error of measurement was calculated for the revised exam in two forms. It was done for the overall sample population, and then recalculated to include only Psychology Juniors and Seniors to produce results that

would be more equivalent to the pilot exam population. Confidence intervals were identified for all categories. The overlap can be seen in Figure 2 between the pilot exam and the revised exam upperclassmen Psychology population. The overlap range of error suggests that the exams could possibly be considered equivalent. However, it is difficult to tell because of the fluctuation in sample population performance.

To the extent possible, validity was examined with the recommendations of Messick's unitarian concept in mind. Content validity was not measured empirically, although items were categorized by faculty submissions under certain specific CSUSB objectives that they had helped identify. The measure of internal consistency of the associated subscales indirectly supports the notion that content was placed correctly within objectives. Efforts were made to add questions to more thoroughly cover the content as described by the objectives, specifically the fields of Psychology. The result was fewer questions in areas that had been covered more thoroughly in the pilot exam (see Appendix B, Table 8). Construct validity is discussed further after the discussion on the hypotheses.

Because data from the target would be more complete after multiple administrtions over time, the norms provided from this study should be considered only as psuedo-normative data. They were computed to establish norm-references to allow comparisons to start to be made within the CSUSB population, please see Appendix B. These norms should be considered with caution, however, because of the sample population used. It is believed that Psychology upperclassmen in the main study took the exam less seriously than their counterparts in the pilot exam. The mean scores are lower which is probably less a result of the revision to the exam and more a result of the broader sample population since the exam item difficulties were not increased substantially nor were higher levels of Bloom's taxonomy utilized more (Table 5, Appendix B).

Student Performance

Descriptive Statistics. Mean differences in overall student scores from the pilot exam to the revised exam were not directly comparable, neither were the item statistics because different populations took the exam and performed at different levels on it. To equate the results between editions of the exam, it was necessary to ferret out the

performance of Psychology major Juniors and Seniors from the total sample population of the main study. While this seems intuitive to identify performance of the intended audience, these results provide substantial support for the documented limitations of CTT analyses being sample dependent (Weiss & Davison, 1981; Suen, 1990). To view a comparison of demographic data between the pilot sample population and the main study sample population, please see Table 4 (Appendix B). One may wonder why the sample population included so may students not representing the target population. The inclusion of novices was intentional to confirm that the knowledge being assessed was not common knowledge, but knowledge specifically acquired through participation in the CSUSB Psychology undergraduate program. Discussion on those findings follows.

Hypothesis 1. Academic major: there was a significant main effect for academic major. Psychology students scored significantly higher on the exam than non-majors. These results suggest that Psychology students are learning more about Psychology than their non-Psychology major peers, indicating an impact from the curriculum.

Hypothesis 2. Grade level: there was a significant main effect for grade level. Further pair wise comparisons among grade level simple main effects revealed significant differences in test scores as grade level increased. Sophomores scored higher than Freshmen, Juniors scored higher than Sophomores, and Seniors scored higher than Juniors.

The pair wise comparisons for grade level and academic major found main effects between all grades except Freshman. While Sophomores did better than Freshmen, Sophomores in Psychology scored higher than non-major Sophomores. This trend continued with greater differences seen at each higher grade level. The differences are clearly illustrated in Table 9 (Appendix C). Although differences were expected between majors, the grade level differences were greater than expected. Some explanations might include that students gain general knowledge, increased maturity, and more test experience as they progress through more years of school.

Hypothesis 3. Freshmen: The pair wise comparisons conducted confirmed that there were no differences between total scores of Freshmen whether they were Psychology

majors or not. These findings are not surprising, because Freshmen have not had much exposure to education yet, so low scores would be expected for major and non-major students.

Hypothesis 4. Transfer students: despite the violation of homogeneity of slopes, the results of the ANCOVA suggest that there was a relationship between transfer status and test scores, more decidedly when holding grade level constant. Pair wise comparisons of mean score revealed surprising results. When not adjusted for the covariate, grade level, mean score for native students was 23.2, much lower than for transfer students, 37.6. Upon reflection, these results make sense, because the total population of native students would include more Freshmen than the transfer population. Accordingly, when results are computed again, controlling for grade level, the mean scores were more comparable: 27.6 for native and 32.3 for transfer. However, these results are not sufficient support for hypothesis 4, that there would be no significant difference between native and transfer students. In fact, the researchers would encourage another look at comparisons between these populations.

Hypothesis 5. Bloom's taxonomy: the correlations were fairly strong and the results indicate that students with more exposure to Psychology courses are better able to answer questions that require higher cognitive functioning. However, the researchers had hoped to have more questions on the exam that represented levels three, four, and five on the taxonomy. Perhaps questions tapping into these levels of cognitive functioning are best achieved through short answer assessments (Bloom et al., 1981).

Hypothesis 6. "E" option: the results revealed that students who had taken more Psychology classes were more likely to attempt to answer the exam item. While their answers may not necessarily have been correct, the results suggest that they at least had the confidence or partial knowledge to attempt to answer the questions. This was encouraging, because at Trenton State, Curry & Hager (1987) found that student effort varied when there was no incentive to take the test nor any performance expectations, as was the case in this study. However, with the lower mean test scores, it is believed that motivation may have been an issue for the CSUSB sample population.

One point of admission should be provided about the "E" option as applied to this exam. While the researchers believe the use of the "E" option was a viable addition to the exam, it was not applied consistently. It was not available on the first 20 matching questions, the result was forced choice answers for the first 20% of the exam. So, the inflated scores at the beginning of the exam could be due to the higher probably of getting an item right or, just as easily, as a consequence of the length of the test resulting in participants burning out before completing it or running out of time. Time may have been a factor in the lower scores among Psychology Seniors as compared to the pilot exam. The University of Northern Colorado reported a skew in their data because Seniors who may have answered questions correctly did not have enough time to finish the test (Sheehan & Granrud, 1995).

With the content defined by the faculty and the strength of most of the relationships in the variables in the nomological net, it is believed that this exam measured student knowledge of Psychology as it was intended to.

Limitations

Although most of these items are mentioned in the context of the discussion, a brief summary of limitations is provided for those interested in further research on this project.

As mentioned earlier, the use of CTT may have contributed to results that are circular: all statistical calculations were linked to CSUSB students and objectives. It would be prudent to revisit department objectives to ensure they are meeting the criteria of external stakeholder and internal needs of students. Also, a follow-up re-evaluation of the exam and the other OA assessments could then ensure CSUSB is measuring the right objectives. Additionally, department goals and objectives appeared to have substantial overlap, evident in a qualitative review and subscale reliability analyses that had been overlooked in the pilot phase of this project.

The "E" option was not available on the first 20 matching questions of the revised exam. Additionally, ethnicity was inadvertently not requested in the demographics information. This level of information should

be added to allow the school to address any differing needs of ethnic groups or other subpopulations.

Implications

With the potential for the comprehensive exam to be put in place at CSUSB, there are numerous potential benefits to the university, students, and external stakeholders alike. OA results could be compiled to provide a complete picture of student performance and curriculum strengths and weaknesses.

By examining aggregate data and trends in responses, the university can maximize the diagnostic potential of the comprehensive exam could allow weak areas to be identified and addressed with the faculty, as well as to confirm areas of strength (Astin, 1987; Halpern,1987, 1988). The objective data can provide an opportunity for the faculty to reflect on teaching methods and priorities and determine if adjustments ought to be made in the curriculum or if greater emphasis is needed in the classroom (Lien, 1071; Krueger & Heisserer, 1987).

Continued faculty involvement would benefit further enhancements to the OA program. Periodic reviews and updates should be completed on the exam to maintain

security, to remain current with issues and policies, and to ensure correspondence with curriculum or department objectives. It might also be beneficial for the department to share items, objectives, and curriculum objectives with other universities to facilitate further enhancements (Halpern, 1987, 1988).

The student data can also begin to be tracked. It could be administered to incoming Freshmen as well as graduating Seniors to get a gauge of the quantitative differences in knowledge. Tracking of data for pre and post testing could also be practiced to establish criterion scores and to develop composite norms after multiple administrations. The test, if administered to incoming transfer students, could also be used as a diagnostic tool for course advisement before students take redundant courses (Halpern, 1987).

Over time, the scores could be correlated to the probability of other events like graduating on time, graduating with honors, going to graduate school, or performing well on the job (Astin, 1987) Predictive information like this can give students tangible goals for achievement as well as provide useful predictive

information to the university. Moreover, if students were made aware of the expectation of passing a comprehensive exam, it would give them the opportunity to increase study time and resulting scores using the exam as a tool to encourage student learning and retention of information (Astin, 1987; Ebel, 1980; and Tuckman, 1998). Fore knowledge of the exam and expectations of performance on the exam might motivate performance and cognitive engagement (Pintrich & Schrauben in Tuckman, 1998; Loughhead, 1997), encourage studying (Ebel, 1980: Halpin & Halpin, 1982), and reinforce learning and retention (Brown, 1983; Tuckman, 1998; Juhler, et al., 1998).

Further, the faculty could measure student perceptions towards learning, based on student knowledge of the culminating assessment requirement. A survey might be developed asking participants how much time they spend studying, an estimation of how much time they feel their studying was impacted due to the test, and how they would rate their attention to retaining knowledge due to the culminating assessment (Sheehan, 1994; Hattie & Jaeger, 1994). Additionally, the information from the survey tool could lead to a more accurate interpretation of exam

scores. With this level of student benefit, the exam may truly provide a sense of accomplishment or closure at the end of the degree program giving a sense of earning the degree (Loughhead, 1997; Anderson et al., 1884).

Summary and Conclusions

The comprehensive exam was developed to measure increases in student knowledge as a result of their education from the CSUSB Psychology department. It was concluded that, yes, graduating Psychology students know more about Psychology than Freshmen entering the program and they know more about Psychology than their non-Psychology peers. But, does it demonstrate that there is value to employers when students graduate from the school? The researchers believe that the results from this study do put CSUSB one step closer to concluding that. The comprehensive examination measured content objectives that pertained to the application of real studies and theories that can be generalized into life beyond Psychology classes. Paired with the other OA assessments to evaluate the development of intellectual skills, appreciation of diversity, interpersonal skills, and ethical standards, the OA program for CSUSB promises to provide a dynamic measure

to demonstrate some of the net gains in student knowledge, development, and preparation for the real world in their chosen field.

APPENDIX A

UNIVERSITY OBJECTIVES

APPENDIX A: UNIVERSITY OBJECTIVES
Assessment of Educational Outcomes for Psychology Majors at California State University, San Bernardino

Mission Statement
Sections Relevant to Undergraduate Education

I. Preamble

We are a teaching and learning community dedicated to the actualization of human potential, an appreciation and respect for individual uniqueness, diversity and achievement, and the pursuit of knowledge within psychology as a scientific discipline and profession.

II. Whom We Serve

The primary recipients of the psychology department's services are students enrolled at California State University, San Bernardino. Students' needs are met by the faculty's active pursuit of teaching and scholarly excellence.

Since students have a diversity of interests and goals, an ever-present danger for departments of psychology is the temptation to attempt to satisfy all demands of the marketplace. The department will judiciously integrate various market demands and student interests to develop a realistic selection of courses/programs for which we have the resources to maintain high quality instruction.

III. Department's Services

A. Teaching

The goal of the psychology department is to shape and enhance student perspectives through knowledge of the basic processes and fields of psychology. The department will educate our students in a manner consistent with their personal interests and career objectives, and in a manner representing and upholding our professional responsibilities and guidelines. As a result we expect to create a positive reputation and market demand for graduates of our programs.

Description of the B.A. Degree in Psychology. The general objectives of the Bachelor of Arts in psychology are to present the theoretical and methodological aspects of psychology to the undergraduate majoring in this field and to provide service courses as electives for students throughout the university. The primary purpose of the degree program is to provide the student with up-to-date, broad-based knowledge of the science of psychology. The degree program should: 1) prepare students for a variety of professional careers that benefit from undergraduate psychology preparation, 2) provide a strong general education that prepares students for careers emphasizing writing, critical thinking, quantitative analysis and interpersonal skills, 3) prepare students for paraprofessional careers in counseling or related fields where graduate training may

not be necessary, and 4) provide an excellent base for those intending to enter graduate school.

Goals and Objectives for Psychology Majors

Goal 1. Basic processes, methodology fields, and applications of psychology The primary goal of the psychology department is to provide psychology students with substantial understanding of the major theoretical and methodological aspects of psychology and psychological inquiry.

Objective 1: Students will have an understanding of differences and similarities among psychology's schools of thought as exemplified by the cognitive, behavioral, psychodynamic, humanistic, and biological perspectives.

This objective develops a breadth and depth of knowledge of psychology's basic processes. The development of this knowledge is assessed in courses on the basic processes of psychology, and ensured by the breadth of basic processes classes students are required to complete for the psychology

Objective 2: Students will have an understanding of the different fields of study and applications of psychology

This objective develops a breadth and depth of knowledge of psychology's various fields and applications, and current thinking in these areas. The development of this knowledge is assessed in courses on the fields of psychology, applications of psychology, and electives that include coverage of psychology and current culture. This knowledge is ensured by the breadth of required classes students are required to complete for the psychology degree.

Objective 3: Students will have an understanding of basic principles of research design.

This objective develops a depth of knowledge in basic research design. The development of this knowledge is assessed in the psychology department's sequence of quantitative and experimental methods courses required for all psychology majors.

Objective 4: Students will have an understanding of concepts in data analysis as applied to statistical decision making and hypothesis testing.

This objective develops a depth of knowledge of statistical techniques and reasoning. The development of this knowledge is assessed in the psychology department's sequence of quantitative and experimental methods courses required for all psychology majors.

The Undergraduate Program Committee continually reviews the psychology course offerings to ensure the department is offering courses that provide a breath and depth of knowledge in all of

these areas. In addition, information collected from graduated students ensures the relevance of knowledge to applied experiences

Goal 2. Intellectual and technology skills The psychology curriculum will enhance intellectual skills that are broadly applicable in work and graduate school, including written and oral communication, quantitative and computer skills, and critical thinking.

Objective 5: Students will have effective writing skills in multiple formats, especially the style of the American Psychological Association.

This objective includes the development of written communication skills and knowledge of the American Psychological Association styles. The development of these skills and knowledge is assessed in the many courses requiring students to create a written product.

Objective 6: Students will have the ability to defend and explain ideas orally clearly and without dogma.

This objective includes the development of oral communication skills and critical thinking skills. The development of these skills is assessed in the many courses requiring students to develop and present information on chosen topics.

Objective 7: Students will have the ability to create and understand tables and graphs.

The development of this ability is assessed in many of the required and elective courses that use graphical information to present course information.

Objective 8: Students will have information gathering skills, which includes electronic searches of data bases.

The development of these skills is assessed in multiple classes that require students to independently gather information for the development of papers and projects.

Objective 9: Students will have the ability to make appropriate generalizations from empirical findings.

This objective includes the development of critical thinking skills and the ability to apply statistical reasoning and probability theory. These skills and abilities are assessed in the multiple classes that require students to interpret empirical findings from the published literature as well as classes that require students to interpret and communicate findings from their own projects.

Objective 10: Students will have the ability to support conclusions with reasons and evidence.

This objective includes the development of critical thinking skills. The development of these skills is assessed in the many courses requiring students to develop and present information on chosen topics.

Goal 3. Diversity The Psychology Department will help students develop an appreciation and respect for individual uniqueness and diversity and individual differences in human behavior.

Objective 11: Students will have an understanding of topics in prejudice and discrimination, and appreciation of diversity.

This objective develops an understanding of prejudice and discrimination and the relevant psychological processes, as well as and appreciation of diversity through the understanding of relevant psychological processes and multiple definitions of "normal". This knowledge is assessed in various required and elective courses that include discussions in this area. The development of knowledge is ensured by the inclusion of diversity issues in multiple required and elective courses.

Objective 12: Students will have an understanding of applications of psychology to contemporary issues such as violence, mental illness, homelessness, or physical handicaps

This objective develops an understanding of role of psychology in contemporary societal issues. This knowledge is assessed in various required and elective courses that include discussions in this area. The development of knowledge is ensured by the inclusion of diversity issues in multiple required and elective courses.

Objective 13: Students will have an understanding of individual differences and their assessment and measurement.

This objective develops an understanding and appreciation of individual differences in personality, values, motives, abilities, and skills. This knowledge is assessed in various required and elective courses that include discussions in this area.

Goal 4. Commitment to learning We strive to advance the intellectual development of our students by engaging them in activities that will encourage a commitment to life long learning.

Objective 14: Students will have curiosity about human behavior and skills to studying its causes.

This objective includes the development of critical thinking skills that allow students to develop and pursue their own ideas and interests, as well as develop the motivation and persistence to complete the development of ideas. The development of these skills and

abilities is assessed in the many psychology classes at require the development of individual and group papers and projects.
Objective 15: Students will participate in learning activities that foster intellectual growth.

This objective includes student participation in reading, media selections, and other learning activities that develop critical thinking skills, written and oral communication skills, and research skills that allow students to develop and pursue their own ideas and interests. The development of these skills and abilities is assessed in the many psychology classes that require the development of individual and group papers and projects.

Goal 5. Interpersonal skills Psychology students will develop their interpersonal skills so that they can participate in and lead groups.

Objective 16: Students will work effectively and cooperatively in groups.

This objective includes the development of social skills that allow students to accurately perceive themselves and others, and act in a cooperative manner. The development of these social skills will be assessed in classes that require students to work with others

Objective 17: Students will adapt to organization rules and procedures.

This objective includes the development of social skills that allow students to accurately perceive themselves and others, and act appropriately. The development of these social skills will be assessed in classes that require students to work with others

Objective 18: Students will demonstrate confidence and leadership.

This objective includes the development of leadership skills. The development of these skills will be assessed in classes that require students to work with others

Objective 19: Students will develop the ability to accurately perceive the behavior of others.

This objective includes the development of social skills that allow students to accurately perceive themselves and others. The development of these social skills will be assessed in classes that require students to work with others

Objective 20: Students will understand the impact of own behavior on others.

This objective includes the development of social skills that allow students to accurately perceive themselves and others. The development of these social skills will be assessed

in classes that require students to work with others

Goal 6. Ethical standards Psychology students will develop an understanding of high ethical standards across academic and professional settings.

Objective 21: Students will behave in accordance with professional and ethical

This objective develops knowledge about professional and ethical conduct including an understanding of department and university policies regarding academic standards. This knowledge is assessed in several classes that include the discussion of ethics, as well as the observation of student behavior in all of their psychology classes.

Objective 22: Students will behave in accordance with APA standards covering all aspects of research activity.

This objective develops knowledge APA standards. This knowledge is assessed in several classes that include the discussion of ethics, as well as the observation of student behavior in all of their psychology classes.

Assessment

In order to assess the psychology departments efforts to reach each of the twenty two objectives we will be using six different assessments tools: course requirements, culminating assessment, sample exit interviews, alumni survey, professional development experiences, and peer review. This combination of assessment tools has been designed to provide multiple assessments of the knowledge, skills, and abilities that students should developed at CSUSB based on our list of objects. In addition, the assessment provides information about the transferability of these knowledge, skills, and abilities to students professional lives after leaving CSUSB.

Assessment 1. Course requirements

The undergraduate psychology curriculum has been designed to ensure that psychology majors are exposed to a breadth and depth of knowledge of psychology, methodology, and application. In addition, the diversity of assignments ensures the development of basic technology and intellectual skills. Table 3 list3 each of the objectives and the categories of classes which assess those objectives. Grades on exams, presentations, and class assignments indicate the degree of understanding and skill developed by each student. Psychology faculty completed surveys to indicate which objectives were covered in their classes. Categories of classes include:

- Introduction - Students are required to take both of the following courses: Psyc 100. Introductory Psychology (4) Psyc 101. Psychology as a Major (1)

- Basic Processes - Students are required to complete three courses from the

following list:

Psyc 100. Introductory Psychology (4) Psyc 101. Psychology as a Major (1)
Psyc 428 Advanced Seminar in Psychology: Industrial and Organizational (4)

- Psychology Electives - Students are required to complete four upper division electives. In addition, they have the option of completing for lower level electives to fulfill overall requirements for graduation.

Psyc 105. Critical Thinking (4)
Psyc 115. Personal and Social Development (2)
Psyc 120. Career Development (2)
Psyc 270. Infant and Toddler Development (4)
Psyc 280. Early Childhood Development (5)
Psyc 301. Psychology of Human Sexuality (4)
Psyc 302. Management and Organizational Behavior (4)
Psyc 303. Parenting and Family Relations (4)
Psyc 305. Psycholinguistics (4)
Psyc 318. Health Psychology (4)
Psyc 320. Psychology of Middle Childhood (4)
Psyc 324. Developmental Psycho biology (4)
Psyc328. Psychology of Adolescent Development (4)
Psyc 329. Psychology of Adulthood and Aging (4)
Psyc 331. The Psychology of Women (4)
Psyc 332. Biofeedback (4)
Psyc 333. Drugs and Behavior (4)
Psyc 334. Addiction and Recovery (4)
Psyc 340. Prejudice, Race, and Racism (4)
Psyc 345. Cross-Cultural Psychology (4)
Psyc 349. The Psychology of Gays and Lesbians (4)
Psyc 350. Development of Exceptional Children (4)
Psyc 358. Cognitive Development (4)
Psyc 366. Computer Applications in Psychology (4)
Psyc 370. Topics in Psychology (2)
Psyc 372. Psychology of Death and Dying (4)
Psyc 375. Seminar in Psychology (4)
Psyc 384. Personality and Social Development (4)
Psyc 387. Community Psychology (4)
Psyc 391. Psychopathology of Childhood (4)
Psyc 395. Psychology of Consciousness (4)
Psyc 400. Child Assessment (5)
Psyc 460. Adult Assessment (4)
Psyc 540. Work, Retirement, and Leisure (4)
Psyc58l. Seminar in Industria1/Organizational Psychology (2)

Psyc 595. Independent Study (2 to 4)
Psyc 597. Honors Seminar (2)
Psyc 598. Honors Project

Assessment 2. Culminating Assessment

Psyc 357. History and Systems of Psychology (4)
Psyc 360. Cognitive Psychology (4)
Psyc 362. Learning and Motivation (4)
Psyc 363. Biological Psychology (4)
Psyc 364. Perception (4)

- Fields of Psychology - Students are required to complete two courses from the following list:

Psyc 201. Developmental Psychology (4)
Psyc 355. Industrial Psychology (4)
Psyc 382. Psychology of Social Behavior (4)
Psyc 385. Personality (4)
Psyc 390. Abnormal Psychology (4)

- Applications of Psychology - Students are required to complete one course from the following list:

Psyc 315. Communication Processes (5)
Psyc 318. Health Psychology (4)
Psyc 351. Behavior Modification: Principles and Applications
Psyc 377. Tests and Measurements (4)
Psyc 386. Introduction to Psychotherapy (4)
Psyc 575. Internship in Psychology (4)

- Research Methods and Statistical Analysis - Students are required to complete both of the following courses:

Psyc 210.
Psyc 311.

In addition students have the option to take addition methods courses which provide advanced experience in Psyc 410 Advanced Psychological Statistics, and the advanced lab courses.

- Advanced Culminating Seminar or Laboratory - Students are required to complete one course from the following list:

Psyc 431 Experimental Psychology: Developmental (6)
Psyc 432 Experimental Psychology: Clinical (6)
Psyc 433 Experimental Psychology: Physiological (6)
Psyc 434 Experimental Psychology: Social (6)
Psyc 435 Experimental Psychology: Personality (6)
Psyc 436 Experimental Psychology: Learning and Motivation (6)
Psyc 437 Experimental Psychology: Cognition and Perception (6)

Psyc 438 Experimental Psychology: Industrial and Organizational (6)
Psyc 421 Advanced Seminar in Psychology: Developmental (4)
Psyc 422 Advanced Seminar in Psychology: Clinical (4)
Psyc 423 Advanced Seminar in Psychology: Physiological (4)
Psyc 424 Advanced Seminar in Psychology: Social (4)
Psyc 425 Advanced Seminar in Psychology: Personality (4)
Psyc 426 Advanced Seminar in Psychology: Learning and Motivation (4)
Psyc 427 Advanced Seminar in Psychology: Cognition and Perception (4)

All psychology majors will be required to take a comprehensive exam during their advanced culminating class to assess the understanding of the core content and application of psychology. The exam will:

- will consist primarily of multiple choice items with the possibility of short answers and/or performance items
- will be administered during the final exam or last week of the quarter
- will include core elements of psychology and their application
- will be developed locally, using the faculty as subject matter experts about which
content to for
- will be used to provide feedback for the department and the student
- will not be used as a source of evaluation or grading for the student
- may be computer administered to give students experience with computerized exam

Assessment 3. Sample Exit Interview

A small randomly selected sample of graduating seniors will participate in yearly exit interviews. The interview:

- will ask job applicants to evaluate the relevance of knowledge, skills and abilities
developed as a psychology major (KSAs) in obtaining entry-level employment
- will ask graduate school applicants to evaluate the relevance of KSAs in the admissions process
- will ask graduates to evaluate the adequacy of advising regarding major and minor fields, degree requirements, and university regulations/policies
- will ask graduates to evaluate the level of intellectual challenge experiences in the
major
- will ask graduates to indicate immediate and log-term career goals/plans
- will ask graduates to indicate if they would choose the major again if they were beginning undergraduate study
- will ask graduates to assess the strengths of the psychology major
- will ask graduates to offer suggestions toward enhancing the undergraduate program in psychology

Assessment 4. Alumni Survey

In the Spring of each academic year the department will mail questionnaires to selected Alumni at 1 year, 3 years, and 5 years post-B.A. The questionnaire will provide feedback to the department about the transference and applicability of KSAs developed that the CSUSB psychology department. In addition, we will ask for feedback on the adequacy of the academic advising they received at CSUSB, and recommendations for additional information and courses they feel the psychology department should focus on developing.

Assessment 5. Professional Development Experiences

The psychology major should provide opportunities for students to participate in professional activities within the field and to achieve a modicum of proficiency in this regard. Professional development experiences include:

- membership in one or more of the psychology clubs
- honors program
- collaboration with faculty research
- attendance and presentations at professional conferences or workshops
- membership in professional organizations
- fieldwork opportunities such as internships and externships

Assessment of these experiences will include an ongoing monitoring of the percentage of students that participate in professional development experiences.

Assessment 6. Peer Review

A peer review instrument will be created to provide students and the department with feedback from other students. This instrument will be administered at the end of their culminating advanced lab or seminar where students work on group projects. Group members will rate their colleagues on:

- application of knowledge of the core content of Psychology
- application of intellectual skills
- application of technology and methodology skills
- interpersonal understanding
- interest and commitment to learning
- ethical standards

The different methods assess multiple goals and objectives, Table 3 indicates the relationship between course requirements and individual objectives. Table 3a indicates the relationship between the remaining assessment tools and individual objectives. There are no a priori criteria for success using these assessment tools; the psychology department will use these tools to assess current levels and monitor future efforts to improve upon goal attainment.

Table 3. Relationship of Objectives and Assessments in Courses

Obj.	Introduction	Basic Processes	Fields of Psychology	Applications	Research Methods	Advanced Sem/Lab	Electives
1	X	X				X	X
2	X		X	X		X	X
3	X	X		X	X	X	X
4	X				X	X	
5	X	X	X	X	X	X	X
6	X	X	X	X	X	X	X
7	X	X	X	X	X	X	X
8	X	X	X	X	X	X	X
9		X	X	X	X	X	
10	X	X	X	X	X	X	X
11	X	X				X	X
12	X	X					X
13	X		X	X	X		X
14	X	X	X	X	X	X	X
15	X	X	X	X	X	X	X
16						X	X
17	X					X	X
18						X	X
19	X	X				X	X
20	X	X				X	X
21	X	X	X	X	X	X	X
22	X	X	X	X	X	X	X

Note: X indicates that the objectives are assessed in these courses

Table 3a. Relationship of Objectives and Assessment Tools

Objective #	Culminating Assessment	Exit Interview	Alumni Survey	Professional Development	Peer Review
1	X	X	X		X
2	X	X	X		X
3	X	X			X
4	X	X			X
5					X
6		X			X
7	X	X			
8		X	X		X
9	X	X			X
10		X			X
11	X	X			
12	X	X			
13	X				
14		X			X
15		X	X		X
16				X	X
17		X	X	X	X
18				X	X
19	X	X			X
20	X	X	X	X	X
21		X		X	X
22	X	X		X	X

APPENDIX B

TEST STATISTICS

APPENDIX B: TEST STATISTICS

Table 4 Demographics of Sample Populations

		Pilot Exam		Revised Exam	
Demographic Area	Specific Options	Pilot Exam	Valid Percent	Main Study Revised Exam	Valid Percent
N		90		521	
Gender	Male	14	15%	97	22.6%
	Female	76	85%	331	77.3%
Major	Psych	NA	NA	171	34.7%
	Non-Psych	NA	NA	322	65.3%
Grade Level	Freshman	NA	NA	147	29.3%
	Sophomore	NA	NA	100	20.0%
	Junior	NA	NA	125	25.0%
	Senior	NA	NA	129	25.7%
Transfer Status	Transfer	NA	NA	191	46.1%
	Non-Transfer	NA	NA	223	53.9%
Age	Mean	27.41			
	Range	19 to 54			
	17 to 20			211	47.0%
	21-23			77	17.1%
	24-30			75	16.7%
	31 and over			78	19.2%
GPA	Mean	3.2			
	Range	2.0-4.0			
	3.5-4.0			108	23.0%
	3.0-3.4			178	37.9%
	2.5-2.9			126	26.5%
	2.0-2.4			48	10.2%
	1.5-1.9			10	2.1%
# Psych Courses	Mean	4.86		1.78	
	Range	1 to 13		0 to 12	
	Standard Dev			2.73	
Area of Psych Interest	Biological	2	2%	8	6.1%
	Clinical	29	32%	47	35.3%
	Cognitive	9	10%	6	4.5%
	Developmental	33	37%	24	18.0%
	Experimental	NA		5	3.7%
	I/O	3	3%	11	8.3%
	Personality	13	15%	8	6.1%
	Social	NA		24	18.0%
Satisfaction with CSUSB	Great	21	23.9%	33	17.7%
	Very Good	19	21.6%	55	29.6%
	Good	34	38.6%	82	44.1%
	Poor	4	4.6%	4	2.2%
	Very Poor	2	2.2%	1	.5%
	NA	8	9.1%	11	5.9%

Table 5. Bloom's Taxonomy Revised Exam Inter-rater Agreement

	Rater 1	*Rater 2*	*Rater 3*	*Rater 4*	*Rater 5*
Rater 1	1				
Rater 2	0.51	1.00			
Rater 3	0.62	0.68	1.00		
Rater 4	0.50	0.63	0.62	1.00	
Rater 5	0.56	0.53	0.50	0.48	1

Average Correlation:
.56

Average Percent Agreement:80.6%

Table 5a. Bloom's Levels for Item

Level	Pilot	Revised
1: Knowledge	47 items	49 items
2: Comprehension	40 items	38 items
3: Application	13 items	13 items
4: Analysis	--	--
5. Synthesis	--	--
6: Evaluation	--	--

Table 6
Subscale Reliabilities

		Pilot Exam		Revised Exam		
CSUSB Goals	CSUSB Content Objs	Pilot Exam Subscale Reliability Coeffic.	Number of Items	Revised Exam Subscale Reliability Coeffic.	Number of Items	*Alternate subscale rel. coeffic.*
Goal 1	Obj. 1 Obj. 2 Obj. 3 Obj. 4	.38 .75 .67 .35	6 22 16 13	.76 .85 .84 .700 Obj 1-4 .91	36 19 12 8	*Obj 1-3* *.90*
Goal 2	Obj. 7 Obj. 9	.47 --	9 --	.61 .49 Obj 7, 9 .68	3 3	*Obj 4,7, 9* *.82*
Goal 3	Obj. 11 Obj. 12 Obj. 13	.40 .55 .10	5 14 2	.65 -- .58 Obj. 11, 13 .77	5 1 4	
Goal 5	Obj. 19 Obj. 20	.26 .02	2 4	.66 -- .66	3 --	
Goal 6	Obj. 22	.54	7	.61 .61	5	*Goal 5, 6* *.76*
OVRL LAlpha		.90		.95		

Table 8. Content Categorization by Objective

CSUSB Comprehensive Exam Objectives	Pilot Exam Items	Revised Exam Items
Objective 1: Psychological Theories	1-6	1-20, 21, 25, 26, 35, 41, 44, 63, 71, 75, 76, 80, 81, 84, 90, 91, 95, 100
Objective 2: Fields of Study	7-28	23, 30, 31, 32, 33, 46, 59, 65, 68, 73, 77, 78, 82, 83, 87, 88, 93, 98
Objective 3: Research Design	29-44	28, 34, 37, 38, 39, 43, 45, 47, 48, 50, 53, 89
Objective 4: Statistics and Data Analysis	45-57	49, 51, 52, 54, 55, 56, 66, 67
Objective 7: Tables & Graphs	58-66	58, 60, 62
Objective 9: Empirical Generalizations & Conclusions		27, 64, 69
Objective 11: Diversity	67-71	70, 72, 74, 77, 78
Objective 12: Applications & Special Issues	72-85	86
Objective 13: Individual Differences & Measurement	86-87	22, 24, 29, 92
Objective 19: Perceive Behaviors of Others	88-89	61, 79, 97
Objective 20: Impact of Own Behavior	90-93	
Objective 22: APA Standards	94-100	40, 94, 92, 96, 99

APPENDIX B Continued
ITEM ANALYSES

	Pilot Exam				Revised Exam		
Item #	Prop Correct	Disc Index	Point Biserial	Comments/Recommendations	Revised Exam	Prop Corr N = 530	Point Biserial N=530
1	0.6	0.27	0.26	Distracter A- Revise or remove	21	0.53	0.38
2	0.39	0.53	0.43	Good Discr & Diff. Reworded	13RW	0.43	0.46
3	0.61	0.12	0.07	Too easy	X		
4	0.74	0.41	0.4	Removed - relevance of content?	X		
5	0.7	0.34	0.25	Too easy. A & D are weak distracters	X		
6	0.86	0.3	0.38	Reworded	76RW	0.26	0.31
7	0.88	0.3	0.31	Reworded - too easy	57	0.23	0.16
8	0.33	0.64	0.58	High Discriminability & high Difficulty	65	0.40	0.56
9	0.93	0.08	0.15	Should have retained	X		
10	0.85	0.22	0.25	Reworded - too easy	5RW	0.25	0.30
11	0.71	0.3	0.27	Removed - too much Freud	X		
12	0.76	0.41	0.39	Removed - relevance of content?	X		
13	0.85	0.15	0.08	Too easy with weak distracters	X		
14	0.57	0.45	0.37	Removed	X		
15	0.87	0.26	0.34	Too easy with poor distracters	26	0.45	0.45
16	0.48	0.42	0.25	Removed	X		
17	0.8	0.44	0.4	Removed	X		
18	0.59	0.41	0.26	Fairly strong item	14	0.48	0.53
19	0.33	0.57	0.5	Difficult two-answer item	28	0.27	0.31
20	0.72	0.56	0.48	Good discr and difficulty	29	0.21	0.27
21	0.27	0.39	0.39	Removed - "D" too attractive cter	X		
22	0.51	0.27	0.16	Reworded	31	0.12	0.24
23	0.97	0.07	0.23	Too easy	X		
24	0.85	0.15	0.22	Little discr and low difficulty	X		
25	0.5	0.49	0.36	Reworded	33RW	0.39	0.36
26	0.53	0.34	0.27	Reworded	87RW	0.74	0.44
27	0.23	0.39	0.37	Too difficult	35	0.48	0.47
28	0.66	0.52	0.35	Reworded	30RW	0.19	0.35
29	0.43	0.38	0.28	Improve "C" and "D"	36	0.42	0.50
30	0.56	0.31	0.23	Removed	X		
31	0.49	0.46	0.24	Improve "D"	37	0.33	0.30
32	0.59	0.38	0.29	Reworded	38RW	0.21	0.13
33	0.39	0.67	0.45	Practically a perfect question	39	0.11	0.29
34	0.56	0.49	0.33	Common Knowledge	X		
35	0.7	0.56	0.43	Common Knowledge	X		
36	0.71	0.3	0.34	Reworded	42RW	0.30	0.29

37	0.71	0.41	0.37	Distracter-A is weak	43	0.12	0.19
38	0.5	0.38	0.35	Review distracter-A	44	0.28	0.31
39	76	0.41	0.39	Distracter-A is weak	45	0.17	0.30
40	0.37	0.42	0.33	Removed Review distracter-A	X		
41	0.35	0.42	0.18	Review distracter-C	47	0.25	0.41
42	0.13	0.14	0.1	High difficulty	48	0.34	0.44
43	0.46	0.38	0.27	Reworded	49RW	0.40	0.51
44	0.53	0.2	0.15	Low discriminability	50	0.35	0.47
45	0.33	0.46	0.39	Review distracter-A	52	0.40	0.50
46	0.21	0.21	0.14	High difficulty; low discriminability	X		
47	0.55	0.24	0.14	Reworded	53RW	0.18	0.31
48	0.96	0.15	0.26	Low difficulty & high discriminability	54	0.12	0.22
49	0.4	-0.16	-0.16	Removed	X		
50	0.59	0.12	0.08	Removed	X		
51	0.35	0.42	0.33	None	55	0.11	0.28
52	0.51	0.23	0.2	Low discriminability modify distr-C	56	0.25	0.39
53	0.3	0.42	0.37	None	51	0.18	0.12
54	0.2	0.06	0.13	Removed - High diff; low disc	X		
55	0.19	0.03	0.03	Removed - High diff; low disc	X		
56	0.6	0.31	0.28	Distracter-D is weak	60	0.25	0.41
57	0.1	0.07	0.01	High difficulty; low discriminability	62	0.33	0.52
58	0.71	0.23	0.19	Distracter-B is weak	54	0.13	0.22
59	0.59	0.38	0.31	Distracter-A is weak	64	0.27	0.36
60	0.79	0.41	0.43	Distracter-B is weak	66	0.21	0.28
61	0.73	0.3	0.25	Distracter-B is weak	67	0.18	0.33
62	0.62	0.42	0.24	Removed - Duplicate with revised Q	X		
63	0.72	0.34	0.23	Distracters-B & C are weak	69	0.17	0.31
64	0.66	0.23	0.17	Distracter-D was selected by 0 examinees	70	0.30	0.49
65	0.56	0.31	0.23	Removed	X		
66	0.89	0.19	0.14	Low difficulty	72	0.32	0.47
67	0.8	0.33	0.38	Removed - Too easy	X		
68	0.49	0.35	0.19	Possibly alter distracters	74	0.17	0.43
69	0.87	0.22	0.25	Removed - Too easy	X		
70	0.94	0.15	0.26	Removed - Too easy	X		
71	0.61	0.49	0.43	Reworded	77RW	0.21	0.16
72	0.79	0.41	0.33	Discr acceptable, Difficulty okay	78	0.58	0.54
73	0.6	0.63	0.57	High Discrim, Point biserial high	79	0.19	0.36
74	0.71	0.38	0.27	Removed - Too easy	X		
75	0.6	0.45	0.35	Distracters evenly selected low scorers	81	0.23	0.27
76	0.77	0.45	0.39	Removed	X		
77	0.59	0.13	0.16	Reword-too long, pt biserial low	83	0.15	0.35
78	0.69	0.27	0.26	Change "C" option to "rabbits"	84	0.39	0.46

79	0.22	0.24	0.22	Removed	X		
80	0.88	0.11	0.14	Too easy, low discrim & low difficulty	86	0.34	0.49
81	0.19	0.07	0	Removed - controversial?	X		
82	0.76	0.37	0.32	C not a good distracter	88	0.39	0.58
83	0.44	0.41	0.31	D is strong distracter for both levels	89	0.15	0.32
84	0.47	0.24	0.18	D is too good of a distracter	90	0.44	0.37
85	0.69	0.41	0.32	No one selected B	91	0.13	0.23
86	0.67	0.08	0.02	Reworded	12RW	0.43	0.49
87	0.47	0.53	0.38	Removed	X		
88	0.37	0.2	0.17	Removed	X		
89	0.55	0.31	0.28	Reworded	61RW	0.19	0.42
90	0.67	0.3	0.22	Removed - too simple	X		
91	0.78	0.45	0.4	Removed - duplicate	X		
92	0.56	0.13	0.08	Check "A" as distracter	97	0.37	0.53
93	0.48	0.42	0.31	Removed - tricky	X		
94	0.49	0.34	0.26	Removed	X		
95	0.77	0.41	0.4	Low scorers correct .56	99	0.36	0.54
96	0.22	0.1	0.08	Removed	X		
97	0.54	0.45	0.42	Reworded	94RW	0.31	0.37
98	0.62	0.6	0.47	Good distribution for low scorers	96	0.37	0.45
99	0.91	0.11	0.19	Removed	X		
100	0.6	0.41	0.29	"C" Good distracter for low scorers	40	0.25	0.23

APPENDIX B Continued
Revised Exam Item Analyses

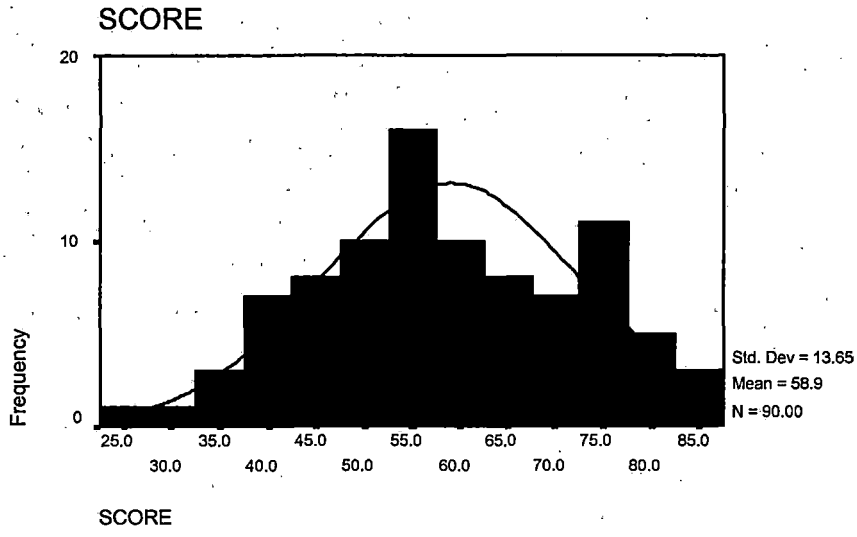
Item	Item Difficulty	Point Biserial
1	0.53	0.38
2	0.43	0.46
3	0.32	0.33
4	0.29	0.27
5	0.29	0.35
6	0.26	0.31
7	0.23	0.16
8	0.40	0.56
9	0.14	0.12
10	0.25	0.30
11	0.17	0.07
12	0.13	0.12
13	0.40	0.44
14	0.29	0.42
15	0.45	0.45
16	0.29	0.42
17	0.26	0.29
18	0.48	0.53
19	0.27	0.31
20	0.21	0.27
21	0.42	0.49
22	0.12	0.24
23	0.36	0.38
24	0.54	0.53
25	0.39	0.36
26	0.74	0.44
27	0.48	0.47
28	0.19	0.35
29	0.42	0.50
30	0.15	0.21

31	0.33	0.30
32	0.21	0.13
33	0.11	0.29
34	0.70	0.46
35	0.18	0.27
36	0.30	0.29
37	0.12	0.19
38	0.28	0.31
39	0.17	0.30
40	0.48	0.30
41	0.25	0.41
42	0.34	0.44
43	0.40	0.51
44	0.35	0.47
45	0.40	0.50
46	0.24	0.24
47	0.18	0.31
48	0.12	0.22
49	0.13	0.33
50	0.19	0.31
51	0.11	0.28
52	0.25	0.39
53	0.18	0.12
54	0.71	0.42
55	0.17	0.37
56	0.25	0.41
57	0.33	0.52
58	0.13	0.22
59	0.27	0.36
60	0.21	0.28
61	0.18	0.33
62	0.09	0.16
63	0.17	0.31
64	0.30	0.49
65	0.15	0.33

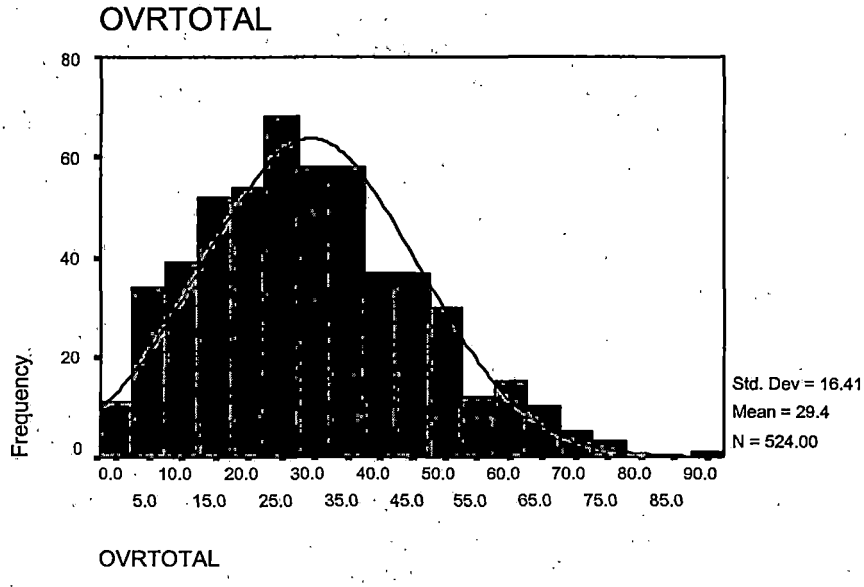
66	0.32	0.47
67	0.37	0.46
68	0.17	0.43
69	0.26	0.45
70	0.38	0.41
71	0.21	0.16
72	0.58	0.54
73	0.19	0.36
74	0.23	0.37
75	0.23	0.27
76	0.23	0.47
77	0.15	0.35
78	0.39	0.46
79	0.26	0.47
80	0.34	0.49
81	0.27	0.51
82	0.39	0.58
83	0.15	0.32
84	0.44	0.37
85	0.13	0.23
86	0.43	0.49
87	0.31	0.32
88	0.47	0.53
89	0.19	0.42
90	0.28	0.31
91	0.39	0.53
92	0.37	0.53
93	0.33	0.48
94	0.25	0.55
95	0.36	0.54
96	0.14	0.32
97	0.31	0.37
98	0.37	0.45
99	0.38	0.55
100	0.25	0.23

APPENDIX B Continued
Score Distribution Graphs

Pilot Exam



Revised Exam, All Participants



APPENDIX B Continued
Norms Psychology Majors Only

N	171
Missing	0
Percentiles	
	OVERALL TOTAL
10	12.2
20	23.4
25	26.0
30	30.0
40	34.8
50	40.0
60	45.0
70	50.0
75	52.0
80	54.6
90	62.8
95	67.4
99	79.4

APPENDIX C

TESTS OF BETWEEN SUBJECT EFFECTS

APPENDIX C: Tests of Between Subject Effects

Table 9. ANOVA Results Main Study Hypotheses 1-3
4 x 2 ANOVA

Descriptive Statistics

	Grade Level	Major	Mean	Std. Deviation	N
Test Score	Freshman	Non-Psychology	17.4902	9.8972	102
		Psychology	19.2683	12.0914	41
		Total	18.0000	10.5597	143
	Sophomore	Non-Psychology	22.2179	11.7281	78
		Psychology	35.7895	16.4809	19
		Total	24.8763	13.8046	97
	Junior	Non-Psychology	28.4559	11.1079	68
		Psychology	43.1786	13.7287	56
		Total	35.1048	14.3390	124
	Senior	Non-Psychology	31.9167	13.2152	72
		Psychology	51.2885	12.5438	52
		Total	40.0403	16.0676	124
	Total	Non-Psychology	24.2188	12.7145	320
		Psychology	39.0179	18.0015	168
		Total	29.3135	16.3266	488

Table 10. Transfer student / native student mean scores controlling for grade level

	TRNSFR	Mean	Std. Error
OVRTOTAL	Native	27.6297	1.081
	Transfer	32.3643	1.196

Table 11. ANCOVA Results Main Study Hypothesis 4

Transfer student / native student

	TRNSFR	Mean	Std. Deviation	N
OVRTOTAL	Native	23.1839	13.0776	223
	Transfer	37.5550	16.2171	191
	Total	29.8140	16.2594	414

Table 12. ANCOVA Transfer

Dependent Variable: OVRTOTAL

Source	Type III Sum of Squares	df	Mean Square	F	Sig	Eta Sq	Noncent. Param.	Observed Power
Corrected Model	32385.92	2	16192.96	86.65	.00	.30	173.31	1.000
Intercept	9544.98	1	9544.98	51.08	.00	.11	51.08	1.000
GRADE LEV	11137.88	1	11137.88	59.60	.00	.13	59.60	1.000
TRNSFR	1241.36	1	1241.36	6.64	.01	.02	6.64	.730
Error	76798.75	411	186.85					
Total	477179.00	414						
Corrected Total	109184.67	413						

a Computed using alpha = .05

b R Squared = .297 (Adjusted R Squared = .293)

APPENDIX C Continued

Table 13. Transfer and Native Student Means by Major and Grade Level

Transfer Status	Grade Level	Major	N	Mean	Median	Std. Dev	Min	Max
Native	Freshman	Non Psych	89	17.09	17.00	9.22	1	38
		Psych	32	20.34	23.50	10.76	2	39
		Total	121	17.95	18.00	9.71	1	39
	Sophomore	Non Psych	41	23.80	24.00	11.16	1	53
		Psych	13	25.23	20.00	15.90	3	53
		Total	54	24.14	23.00	12.31	1	53
	Junior	Non Psych	17	32.58	36.00	11.10	11	47
		Psych	7	36.14	37.00	17.97	8	66
		Total	24	33.62	36.00	13.14	8	66
	Senior	Non Psych	12	34.33	31.50	8.65	20	49
		Psych	10	43.20	41.50	15.84	25	74
		Total	22	38.36	33.50	12.93	20	74
	Total	Non Psych	159	21.77	22.00	11.60	1	53
		Psych	62	26.83	25.00	15.93	2	74
		Total	221	23.19	23.00	13.12	1	74
Transfer	Freshman	Non Psych	--	--	--	--	--	--
		Psych	2	14.00	14.00	12.72	5	23
		Total	2	14.00	14.00	12.72	5	23
	Sophomore	Non Psych	13	31.76	34.00	15.98	7	67
		Psych	2	57.50	57.50	7.77	52	63
		Total	15	35.20	34.00	17.47	7	67
	Junior	Non Psych	44	28.18	30.00	11.19	1	52
		Psych	37	46.67	46.00	13.05	20	88
		Total	81	36.62	34.00	15.16	1	88
	Senior	Non Psych	44	29.81	28.50	13.02	6	54
		Psych	46	49.76	51.00	13.03	23	76
		Total	90	40.01	40.00	16.37	6	76
	Total	Non Psych	101	29.35	30.00	12.60	1	67
		Psych	87	47.80	47.00	13.93	5	88
		Total	188	37.89	37.00	16.10	1	88
Total	Freshman	Non Psych	89	17.08	17.00	9.22	1	38
		Psych	34	19.97	23.00	10.77	2	39
		Total	123	17.88	18.00	9.71	1	39
	Sophomore	Non Psych	54	25.72	25.00	12.79	1	67
		Psych	15	29.53	21.00	18.71	3	63
		Total	69	26.55	25.00	14.22	1	67
	Junior	Non Psych	61	29.40	30.00	11.25	1	52
		Psych	44	45.00	43.50	14.24	8	88
		Total	105	35.94	36.00	14.72	1	88
	Senior	Non Psych	56	30.78	31.00	12.29	6	54
		Psych	56	48.58	49.50	13.65	23	76
		Total	112	39.68	39.50	15.72	6	76
	Total	Non Psych	260	24.72	24.00	12.53	1	67
		Psych	149	39.08	39.00	18.02	2	88
		Total	409	29.95	29.00	16.29	1	88

APPENDIX D

INSTRUMENTS

APPENDIX D:INSTRUMENTS
Informed Consent
(Pilot & Main Study)

Thank you for taking your time to participate in this study. Your time is greatly appreciated. Tammy West Reichley, Masters Student of Industrial/Organizational Psychology, California State University San Bernardino, is conducting this study in part for her Master's thesis on Psychology Outcomes Assessment, under the supervision of Dr. Janet Kottke. The purpose of this research is to develop a comprehensive exam for graduating seniors in the Psychology Department of CSUSB.

To be qualified for as a participant, you must be at least 18 years old and a student of CSUSB classified in any Major.

Your participation includes completing the attached exam. It should take about one hour to complete. All of your responses will remain anonymous and will be used for research purposes only. You are strongly encouraged to respond to all items, yet if you feel unable or unwilling to respond to a particular item, please fill in option "E". Participation in this study is completely voluntary and if you have a need to withdraw, you will not be penalized.

This study has been approved by the Institutional Review Board at California State University, San Bernardino. If you have any questions, please contact Tammy West Reichley at (626) 302-5324.

Please indicate your voluntary participation in this study by placing an "X" on the line below and filling in today's date.

Anonymous Participation: _____
Today's Date: _____

Thank you again for your participation.

APPENDIX D Continued
Pilot Study Exam

1. John Watson's statement, "Give me a dozen healthy infants..." was an extreme statement about the importance of _____ in the development of personality.
a. scientific direction
b. good health
c. hereditary factors
d. environmental influences

2. Psychology is considered by historians to have been founded in 1879, name the founder and the country.
a. John Locke; England
b. Wilhelm Wundt; Germany
c. Gustav Fechner; Germany
e. Edward Titchener; United States

3. Research in social psychology, compared to sociology, focuses on the thoughts (cognitions), feelings, and actions of
a. the individual
b. society
c. the group
d. cultural institutions

4. A psychologist who examines whether one remembers the beginning, middle, or end of a list better is probably a _____ psychologist.
a. Social
b. Cognitive
c. Developmental
d. Neurological

5. A psychologist who assesses the effects of noise levels in a factory and the incidence of accidents is probably a(n)_____ psychologist.
a. social
b. behavioral
c. industrial
d. commercial

6. A former student of Freud, this theorist proposed that humans pass through 8 stages of psychosocial development.
a. Erikson
b. Piaget
c. Pavlov
d. Skinner

7. Maslow used the term _____ in describing growth oriented needs to become all one can be.
a. frames of reference
b. existentialism
c. self-actualization
d. conditions of worth

8. Gordon was an active infant who approached strangers easily and showed curiosity easily in new situations. The characteristics reflect Gordon's:
a. personality
b. temperament
c. social referencing
d. adaptability

9. Psychologists who study personality usually define it as
a. the qualities that make a person lively or interesting
b. the person's emotional and behavioral components provided by heredity
c. the reasonably stable patterns of emotions, motives, and behavior that distinguish people from one another*
d. those characteristics that most clearly distinguish an individual at maturity

10. What theory supports the notion that learning occurs through imitation (modeling) of, and identification with other people?
a. psychoanalytic theory
b. psychosocial theory
c. information processing
d. social leaning theory

11. In avoiding conflicts with the id and superego, what does the ego use to adjust reality and make it less stressful?
a. pleasure principles
b. defense mechanisms
c. reality principles
d. free association

12. A psychoanalyst places the most importance on which of the following processes?
a. extrinsic
b. intrinsic
c. unconscious
d. conscious

13. The basic unit of the nervous system is the:
a. soma
b. glial cell
c. neuron
d. axon

14. Systematic errors in reasoning are called...
a. overgeneralizations
b. self-punishment
c. negative feedback
d. cognitive distortions

15. Periods of REM sleep are most closely associated with:
a. dreaming
b. sleepwalking
c. insomnia
d. sleep apnea

16. Barbara expected little out of one of her students because she had been advised by a colleague that the student had performed really poorly in the previous year. When the student met her negative expectation, the result was a(n):
a. self-fulfilling prophecy
b. expectation-outcome match
c. goodness of fit
d. situational, social-order comparison

17. Specialists who study aging are called:
a. thanatologists
b. gerontologists
c. oncologists
d. endocrinologists

18. Noam Chomsky is a pioneering influence in what area of psychology?
a. psychoanalysis
b. psycholinguistics
c. behavior modification
d. health psychology

19. When it comes to solving problems, _____ is a systematic strategy that guarantees a correct solution, while ____ is a more general rule of thumb that may be quick but could also lead to errors.
a. an algorithm ; a heuristic
b. means-ends analysis ; trial and error
c. protocol analysis ; mental modeling
d. functional fixedness ; mental set

20. A response time is a commonly used measure in cognitive psychology. The logic behind this is:
A. If two processes are different, they should take a different amount of time to complete.
B. If two processes are different, they should take a same amount of time to complete.
C. If two processes are the same, they should take a different amount of time to complete.
D. The response time is easier to remember.

21. Jung's collective unconscious contains:
a. the same sexual and aggressive material found in the Freudian unconscious
b. primitive images of the human species such as the young hero or the wise old man
c. various roles that a person's self has played
d. developmental themes from the earliest childhood until the present time

22. The "common cold" of psychological problems is:
a. generalized anxiety
b. depression
c. hypochondriasis
d. phobia

23. Every night Harry spends several hours checking and rechecking the locks on his doors and windows, in a standard procedural fashion. Harry's persistent adherence to his routine may be most indicative of a(n)
a. panic attack
b. phobia
c. obsessive-compulsive disorder
d. generalized anxiety disorder

24. Compared to a Type A personality, a Type B personality might be expected to
a. be less intelligent
b. be older
c. be less driven for success
d. earn more money

25. If you study hard because doing so removes the likelihood that you will fail the next exam, your efforts are being strengthened by a:
a. negative reinforcer
b. dread of failure
c. positive reinforcer
d. reliance on rituals

26. How words are strung together, or ordered, in a language is its:
a. semantics
b. morphology
c. phonology
d. syntax

27. According to Bandura, an adequate measure of efficacy would measure:
a. positive and negative feelings about oneself
b. a general belief about one's ability to do things well
c. belief that good performance will lead to specific outcomes
d. belief concerning one's ability to perform a specific task

28. What is the developmental sequence of Piaget's cognitive stages?
a. preoperational, concrete operational, formal operational, sensorimotor
b. sensorimotor, preoperational, concrete operational, formal operational
c. preoperational, sensorimotor, formal operational, concrete operational
d. sensorimotor, formal operational, preoperational, concrete operational

29. In an experiment on the effects of group size on helping in an emergency, group size would be the _____ variable
a. dependent
b. independent
c. confounding
d. extraneous

30. If an observed effect has a very low probability of having occurred by chance (only 5 times in 100) then it is NOT statistically significant.
a. true
b. false

31. Type I error is
a. the probability of rejecting the null hypothesis when it is true
b. the probability of failing to reject the null hypothesis when it is false
c. the probability of accepting the null hypothesis
d. the probability of rejecting the null hypothesis

32. An operational definition of a variable is
a. easy to measure
b. limited to observable behavior only

c. one that has a causal effect on behavior or mental processes
d. a definition of a variable in terms of the methods used to create or measure that variable

33. In an experiment studying the effects of different drugs on reaction time, each subject is presented with all of the conditions in the experiment. This is an example of a
a. factorial design.
b. within subjects design
c. between subjects design
d. quasi-experimental design

34. Which of the following is NOT a method of data collection?
a. observing and recording behaviors as they occur in a contrived laboratory setting (i.e., laboratory observations)
b. observing and recording behaviors as they occur in a natural or every day setting (naturalistic observations)
c. observing and recording behaviors of many individuals as they spontaneously occur (multi-participant observations)
d. Case studies, in which the researcher does an in-depth report on the life history, attitudes, etc. of a single individual

35. Participants who go through an experiment but do not receive any treatment or manipulation are called the_____ subjects.
a. placebo
b. random
c. control
d. experimental

36. When Oscar exercises for a long time, he experiences a euphoria similar to that elicited by morphine. Which of the following is probably responsible for this sensation?
a. serotonin
b. dopamine
c. acetocholine
d. endorphins

37. A classic study of gifted children begun by Lewis Terman more that 75 years ago is an example of a:
a. cross-cultural study
b. cross-sequential study
c. age cohort study
d. longitudinal study

38. Reinforcing closer and closer approximations to the response we wish to condition is:
a. conditioning
b. attachment
c. shaping
d. accommodation

39. The period when a behavior is measured before a treatment is introduce is called...
a. treatment
b. reversal
c. baseline
d. probe

40. How many levels of variation should be included in an independent variable?
a. Only one level if possible
b. As many different levels as possible
c. At most two different levels
d. At least two different levels

41. In an experiment, Eugene wants to find out the extent to which he can accurately state that the observed effect measured by the dependent variable is due only to variation in the independent variable. In other words, Eugene is trying to achieve
a. construct validity
b. ecological validity
c. external validity
d. internal validity

42. In a _____, two or more independent variables are always presented in combination.
a. random assignment
b. Between subjects design
c. Repeated measures design
d. Factorial design

43. When TWO groups are being compared in a psychology experiment, the proper statistical test for analyzing the results is ____. When there are THREE OR MORE groups in an experiment the proper statistical test is ____.
a. t-test ; analysis of variance
b. chi-square test ; multivariate test
c. correlation ; linear regression
d. a parametric test ; a nonparametric test

44. In a within subjects design, what procedure controls for the order effects of treatment?
a. randomizing
b. matching
c. counterbalancing
d. blocking

45. The process of using sample data to answer general questions about a population is called
a. parameter
b. statistic
c. descriptive statistics
d. inferential statistics

46. In a normal curve...
a. the mean and mode are not the same
b. 99.5% of the cases fall between -2 and +2 standard deviation
c. the mean, mode, and median are the same
d. 68% of the cases fall between the -2 and +2 standard deviations

47. Determining a person's gender would involve measurement on a _____ scale of measurement.
a. nominal
b. ordinal
c. interval
d. ratio

48. There is a good chance that we will find a _____ correlation between time spent studying and grades in school.
a. positive
b. perfect
c. weak
d. negative

49. In a statistical test, Power, 1-Beta is
a. the probability of accepting a false null hypothesis
b. a measure of external and construct validity
c. the probability of rejecting a false null hypothesis
d. a type I error

50. Extreme scores in a distribution most dramatically affect the
a. t-score
b. mode
c. mean
d. median

51. When the null hypothesis is rejected it means that
a. the hypothesis of interest was not supported
b. the hypothesis of interest was supported
c. the theory was discredited
d. the theory was proved

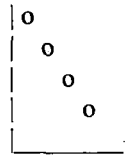
52. Consistency of scoring on tests is referred to as
a. validity
b. reliability
c. predictability
d. continuity

53. Suppose that we obtained a significant F value in a one-way ANOVA with p levels. Based on this result, we can conclude that
a. at least two out of p groups are different from each other.
b. All but p groups are different from each other.
c. None of the groups is different from each other.
d. We cannot conclude anything.

54. Which of the following is NOT a method for graphically representing frequency data?
a. range plot
b. histogram
c. bar graph
d. stem-and-leaf plot

55. What type of curve shows the relative position of individuals in a distribution?
a. a normal curve
b. a cumulative frequency curve
c. a skewed curve
d. a symmetrical frequency curve

56. Match the following graph with its approximate correlation coefficient.



a. R = +1.00
b. R = -1.00
c. R = .00
d. R = .5

57. If you plotted the following data points (1, 3, 5, 48, 50, 85, 95, 96, 98, 99, 100), the distribution would be
a. Positively skewed
b. Negatively skewed
c. Bimodal
d. Normal

58. If an anxiety test actually measures a respondent's anxiety, the test has
a. reliability
b. applicability
c. utility
d. validity

59. Harry Harlow's classic study with cloth and wire monkeys illustrates that
a. food alone is sufficient to bring about attachment
b. the need to satisfy the reinforcement drive promotes the infant's attachment to the mother
c. food alone is insufficient to bring about attachment
d. satisfaction of the hunger drive nurtures the infants attachment to the mother

60. When a headline reads "Lower incomes cause children to have lower IQ's," what mistake has the media made?
a. assuming causality from a correlational study
b. assuming the credibility of a psychologist
c. quoting from nomothetic study
d. misrepresenting an emic study

61. When psychologists say that sample results will generalize to a population, they mean that what is found for the participants in a study will hold true for:
a. all people
b. future participants
c. people in the larger group from which the sample was selected
d. all the people in the geographic area who have participated in comparable studies

62. Inhelder and Piaget designed the pendulum problem in an attempt to explore
a. formal thought
b. trial-and-error
c. applying the equivalence rule
d. questioning the examiner about the meaning of each dimension

63. Bandura's classic Bobo doll experiment focused on the effects of

a. Observational learning
b. Cognitive dissonance
c. Group-think
d. Self-actualization

64. "Negative feelings toward persons based on their membership in certain groups," is a good workable definition of (the)
a. discrimination
b. fundamental attribution error
c. prejudice
d. contact hypothesis

65. Which of the following develops form a sense of membership based on shared experiences and beliefs?
a. acculturation
b. ethnic identity
c. self esteem
d. biological gender

66. Which of the following perspectives regards culture, ethnicity, and gender as important factors?
a. humanistic
b. sociocultural
c. neurobiological
d. psychoanalytic

67. While sex is a _____ dimension of being male or female, gender is a _____ dimension of being male or female
a. sociocultural; biological
b. biological; sociocultural
c. neurobiological; humanistic
d. humanistic; neurobiological

68. What is meant by the term "double jeopardy racism"?
a. Racism in an area (e.g., the community) usually spills over into other areas (e.g., the work place)
b. Racists usually have prejudices against more than one group (e.g., prejudice against African-Americans and Jews)
c. Racism strikes twice, once because it's harder for women to find work and again because it's harder for women to get promoted when they do find work
d. Women from ethnic minorities must deal with both racism and sexism

69. When studying Asian Americans it is necessary to have a Caucasian comparison group to interpret the results.
a. true
b. false

70. What is the primary focus of cross-cultural psychology?
a. the education of children in the public schools
b. the development of teenagers through adolescence
c. the role of culture in understanding behavior
d. examining only one culture, very thoroughly

71. To be treated merely as a group representative rather than a talented individual defines
a. tokenism

b. group-think
c. egoism
d. altruism

72. The term "glass ceiling" describes
a. a barrier to understanding employee problems
b. the effect of capping top male executives' salaries
c. a barrier to moving into management ranks
d. the upper income bracket of Fortune 500 companies

73. As a member of a group, we are most likely to engage in social loafing when:
a. we experience evaluation apprehension
b. our level of arousal increases
c. we are anonymous
d. the leader is an authority figure

74. Homosexuality is viewed as deviant behavior in the DSM-IV.
a. true
b. false

75. Sandra Scarr's position that different parenting approaches do not substantially impact child outcomes is an example of _____ orientation toward psychological development.
a. humanistic
b. laissez faire
c. nature
d. nurture

76. The "misinformation effect" refers to the fact that memory for an event can be altered if inaccurate information is presented after the actual event. The misinformation effect has implications for what current debate?
a. the nature vs. nurture debate
b. the cause-and-effect controversy
c. the false-memory vs. repressed-memory debate
d. the social promotion controversy

77. Within developmental psychology, what is the major premise behind the continuity vs. discontinuity debate?
a. Developmental changes occur abruptly (in a stage-like manner) versus developmental changes occur gradually (more like a slope than like stairs)
b. Development occurs throughout life versus development discontinues after adolescence
c. Important developmental differences continue from one generation on to the next versus developmental differences discontinue after each generation
d. Mental process flow smoothly (continuity) versus mental processes are filled with pauses (discontinuity)

78. Agoraphobia is the fear of
a. spiders
b. heights
c. bunnies
d. open places

79. Which of the following is NOT true about suicide?

a. men attempt suicide more often than women
b. the completed suicide rate is 3 times higher for men than women
c. married people commit suicide less than unmarried and divorced people
d. physicians commit suicide at a higher rate than average

80. Critics of deinstitutionalization of the mentally ill believe that it has led to
a. criminalization of the mentally ill
b. increased homelessness among the mentally ill
c. increased substance abuse among the mentally ill
d. all of the above

81. Victims of childhood sexual abuse are most likely to be
a. Boys 13 to 15
b. Children 9 to 12
c. Girls 3 to 7
d. Children under 7

82. According to the cycle-of-violence hypothesis, abuse and neglect of children leads them to be
a. victims in adult life
b. insecure and withdrawn adults
c. more sympathetic to others as adults
d. predisposed to abusiveness as adults

83. A researcher studies how individuals born during the 1920s compare to individuals born during the 1940s with regard to political participation. The groups of individuals are called:
a. cohorts
b. alliances
c. support systems
d. reference groups

84. Substance abuse addictions may be classified along a continuum of dependence rather than on an absolute basis. This continuum (in order) is:
a. usage, dependence, tolerance
b. indulgence, dependence, gratification
c. usage, abuse, addiction
d. usage, addiction, abuse

85. Name given to a person who unwittingly supports and reinforces another person's addictive behavior:
a. enabler
b. junkie
c. co-addict
d. sponsor

86. Achievement motivation (nAch) was proposed by
a. Clark Hull
b. David McClelland
c. Abraham Maslow
d. Henry Murray

87. Tests that predict what an individual will be able to do later are called
a. achievement tests
b. aptitude tests

c. personality tests
d. all of the above

88. Generally speaking, observers view an actor's behavior or experience of negative events as stemming from
a. chance or bad luck
b. internal traits
c. situational causes
d. extrinsic rewards

89. The theory about the processes we try to explain a person's behavior by attributing it to situational factors and/or inferred dispositional qualities is called
a. James-Lange Theory
b. similarity theory
c. attribution theory
d. social learning theory

90. The sum of an individual's beliefs or schemas about his or her personal traits and characteristics is termed
a. the self-concept
b. autobiographical memory
c. social perception
d. self-awareness

91. What are you using when, for the purposes of understanding, you evaluate your thoughts, ambitions, and behavior against someone else's?
a. social comparison
b. conformity rules
c. the primacy effect
d. cognitive dissonance

92. The tendency to assume that people that have one positive characteristic must have others as well is called
a. homogamy
b. the halo effect
c. reciprocity
d. familiarity

93. Our initial impressions of people are generally based on _____
a. Very little information
b. Their family background
c. Positive stereotypes
d. Their body language

94. Before serving in a psychology experiment, participants must be debriefed regarding the true purpose of the research.
a. true
b. false

95. After subjects give their informed consent to participate in a research study they
a. can withdraw from the study at any time
b. must finish the entire research study

c. can debrief the experimenter if they so choose
d. must be paid at the end of the study

96. A psychologist conducted a study concerning sexual behavior. While collecting questionnaires for subjects he gave out advice on the way in which the subjects should raise their children. Which of the following is accurate?
a. as a professional, the psychologist was obligated to give out advice
b. psychologists should try to improve peoples lives when they have the opportunity
c. psychologists should avoid dispensing wisdom
d. the rule of debriefing was violated

97. Milgram's obedience experiment
a. could be replicated today
b. could not be replicated today
c. demonstrated decreased learning with the use of punishment
d. demonstrated increased learning with the use of punishment

98. The most significant purpose of the IRB approval process is:
a. to let the school know what experiments are being conducted
b. to protect participants from unethical testing
c. To provide strict regulations and control psychological research
d. To allow researchers to practice writing proposals for studies

99. All of the following are required as part of any research study EXCEPT:
a. informed consent
b. confidentiality
c. deception
d. debriefing

100. A clinical psychologist has been trying to build up the nerve to invite her client out to dinner, she should:
a. write him a letter
b. express her intentions verbally during the counseling session with the client
c. refer the client to another doctor before initiating any relationship
d. refer the client to another doctor and remember it is unethical and illegal to date a date a client

APPENDIX D Continued
Demographics Sheet Pilot Study

We want to determine if this test is representative of the information taught at CSUSB. We are not collecting names or ID numbers so your responses cannot be used to identify you. All the information you provide is anonymous and confidential. We would appreciate your taking another minute to answer the following questions.

What area of Psychology are you most interested in?

- ❑ Clinical
- ❑ Developmental
- ❑ Industrial/Organizational
- ❑ Personality
- ❑ Experimental
- ❑ Biological
- ❑ Cognitive

Please mark the courses that you have taken at another college or at CSUSB:

Class taken Elsewhere	Class taken at CSUSB
❑	❑ Intro to Psychology
❑	❑ Statistics
❑	❑ Experimental Psych
❑	❑ History and Systems
❑	❑ Social Psych
❑	❑ Human Development
❑	❑ Cognition-Perception
❑	❑ Abnormal
❑	❑ Personality
❑	❑ Industrial-Organizational
❑	❑ Tests and Measurement
❑	❑ Learning and Motivation
❑	❑ Physiological
❑	❑ Psychobiology
❑	❑ Capstone: Race & Racism
❑	❑ Capstone: Perspectives on Gender

Entry into CSUSB:

- ❑ I started CSUSB as a Freshman
- ❑ I transferred to CSUSB as a Sophomore
- ❑ I transferred to CSUSB as a Junior
- ❑ I transferred to CSUSB as a Senior
- ❑ I am a post baccalaureate student

Expected time I'll graduate:

- ❑ Fall
- ❑ Winter
- ❑ Spring
- ❑ Summer
- ❑ 1999
- ❑ 2000
- ❑ 2001
- ❑ 2002

How many quarters have you completed at CSUSB? ____

I declared myself as a psychology major in my____ year.

- ❑ Freshman
- ❑ Sophomore
- ❑ Junior
- ❑ Senior
- ❑ I am not a psychology major
- ❑ I have not yet declared but will

- ❑ Male
- ❑ Female

GPA: _____

Current age: _____

Age when I started my BA degree: _____

My overall experience as a psychology student at CSUSB has been:

- ❑ Great
- ❑ Very good
- ❑ Good
- ❑ Somewhat good
- ❑ Somewhat bad
- ❑ Bad
- ❑ Very bad

APPENDIX D Continued:
Main Study Demographics

What is your Major? _____ Please indicate your Minor, if you have one. _____

I am currently a:	Entry into CSUSB:
❑ Freshman	❑ I started CSUSB as a Freshman
❑ Sophomore	❑ I transferred to CSUSB as a Sophomore
❑ Junior	❑ I transferred to CSUSB as a Junior
❑ Senior	❑ I transferred to CSUSB as a Senior

How many quarters have you completed at CSUSB? _____Did you take a psychology class in high school? Y / N

GPA	Current age
❑ 3.5-4.0	❑ 16 or less
❑ 3.0-3.4	❑ 17-20
❑ 2.5-2.9	❑ 21-23
❑ 2.0-2.4	❑ 24-27
❑ 1.5-1.9	❑ 28-30
❑ 1.0-1.4	❑ 31-35
❑ 0.5-0.9	❑ 36-40
Gender	❑ 41-45
❑ Male	❑ 46-50
❑ Female	❑ 51 and better

Please mark the Psychology courses that you have taken at another college or at CSUSB:

Class taken Elsewhere	Class taken at CSUSB		
❑	❑	Psy 100	Intro to Psychology
❑	❑	Psy 101	Psychology as a Major
❑	❑	Psy 210	Statistics
❑	❑	Psy 311	Experimental Psych
❑	❑	Psy 357	History and Systems
❑	❑	Psy 382 or 434	Social Psych
❑	❑	Psy 201 or 421	Human Development
❑	❑	Psy 360 or 437	Cognition-Perception
❑	❑	Psy 390	Abnormal
❑	❑	Psy 385 or 435	Personality
❑	❑	Psy 355 or 438	Industrial-Organizational
❑	❑	Psy 377	Tests and Measurement
❑	❑	Psy 362 or 436	Learning and Motivation
❑	❑	Psy 433	Physiological
❑	❑	Psy 363	Psychobiology
❑	❑	316	Capstone: Race & Racism
❑	❑	325	Capstone: Perspectives on Gender

The remaining information is needed from Psychology Majors only:

What area of Psychology are you most interested in?

❑ Biological ❑ Experimental
❑ Clinical ❑ Industrial/Organizational
❑ Cognitive ❑ Personality
❑ Developmental ❑ Social

Overall experience as psychology student at CSUSB has been:

❑ Great ❑ Poor
❑ Very good ❑ Very poor
❑ Good ❑ N/A

APPENDIX D Continued
Main Study Revised Exam

Matching:

1. Erik Erikson	a. Asserted analytical psychology, which featured a collective unconscious, and numerous archetypes, both of which reflect the history of our species.
2. Karen Horney	b. Eight Stages of Psychosocial Development. Moving away from Freud's emphasis on the unconscious influences from the past this person placed a strong emphasis on social determinants of development.
3. Carl Jung	c. Criticized Freud's views as male biased and gave balance to the concepts by emphasizing cultural and interpersonal forces.
4. Alfred Adler	d. Criticized Freud for too much emphasis on sexual impulses. Offered views, termed individual psychology, that people are motivated by an inferiority complex.
	e. Wrote The Ego and the Mechanisms of Defense, listing and describing ego defense mechanisms like repression, denial, and projection.
5. Albert Bandura	a. Originated the term "group think"
6. Lawrence Kohlberg	b. Presented ideas about Social Learning Theory, emphasizing the importance of imitation.
7. Irving Janis	c. Operant Conditioning
8. B. F. Skinner	d. Originated the Theory of Moral Development to explain how children's cognitive development lays the groundwork for different levels of moral reasoning.
	e. Studied four positions of birth order, believing that birth order affects the personality.
9. Abraham Maslow	a. An ethologist who defined "imprinting," the process in the critical period of development of an organism in which that organism responds to a stimulus in a manner that will afterward be difficult to modify.
10. Konrad Lorenz	b. Helped pioneer the assessment of Need for Achievement.
11. Clark Hull	c. Framed the drive-reduction theory view that organisms seek to engage in behaviors that reduce biological drives.
12. David McClelland	d. Presented the humanistic view, emphasizing the positive potential of the individual
	e. Defined "Locus of control" referring to the primary source of a person's behavior as either internal or external.
13. Wilhelm Wundt	a. Proposed a deep structure of language that is an innate tendency to process information in linguistic form.
14. Noam Chomsky	b. Experiments focused on how children and adults mentally represent and reason about the world
15. Jean Piaget	c. Developed the first psychology lab at the University of Lepzig
16. Edward Thorndike	d. Law of Effect, that responses are stamped in and stamped out by punishment.
	e. Proposed the ACT—Adaptive Control of Thought model of memory.
17. Fritz Perls	a. Gestalt Therapy, based on the belief that people should take responsibility for themselves and focus their attention on the 'here and now.'
18. Sigmund Freud	b. Cognitive therapist who used Rational Emotive Therapy and asserted abnormal behavior is the result of irrational thoughts and beliefs.

19. Carl Rogers — c. Published The Interpretation of Dreams," reflecting his psychoanalytic view.

20. Albert Ellis — d. Practiced Client-centered Counseling, giving unconditional positive regard to all clients.

e. Designed a behavioral treatment, systematic desensitization, based on extinction of a conditioned emotional response.

Multiple Choice. Please select the best response to the following questions:

21. John Watson's statement, "Give me a dozen healthy infants..." was an extreme statement about the importance of _____ in the development of personality.

a. scientific direction
b. good health
c. hereditary factors
d. environmental influences
e. I have not been taught this material

22. The minimum amount of stimulus energy necessary for an observer to detect the stimulus is called

a. absolute threshold
b. just noticeable difference
c. receptor adaption
d. central habituation
e. I have not been taught this material

23. Individuals with monocular vision have difficulty judging

a. distance
b. color
c. depth
d. contours
e. I have not been taught this material

24. In a study, Asch asked participants to make judgments about the comparative length of three drawn lines. Confederates went first and gave false judgments in front of the group. About a third of the participants agreed with the false statements. From this study Asch demonstrated that

a. attitudes expressed verbally tend to be more heterogeneous
b. people do what they are told
c. people do not hesitate to state an opposing opinion
d. people tend to conform to a temporary reference group
e. I have not been taught this material

25. _____ are psychoactive drugs that alter consciousness, awareness, or perception.

a. Analgesics
b. Stimulants
c. Depressants
d. Psychedelics
e. I have not been taught this material

26. Periods of REM sleep are most closely associated with:

a. dreaming
b. restlessness
c. insomnia
d. sleep apnea
e. I have not been taught this material

27. On an early morning in 1964, Kitty Genovese was attacked and murdered by a man with thirty-eight witnesses peering out windows. In the thirty minutes it took to kill Kitty Genovese, not one person came to help or call the people. Research on this phenomenon has found that people are more likely to help a stranger if they are alone than if part of a large group. The behavior of the bystanders is described by researchers as:

a. altruistic
b. stranger effect
c. apathetic and uncaring
d. diffusion of responsibility
e. I have not been taught this material

28. When it comes to solving problems, _____ is a systematic strategy that guarantees a correct solution, while ____ is a more general rule of thumb that may be quick but could also lead to errors.

a. an algorithm; a heuristic
b. means-ends analysis; trial and error
c. protocol analysis; mental modeling
d. functional fixedness; mental set
e. I have not been taught this material

29. A response time is a commonly used measure in cognitive psychology. The logic behind this is:

a. If two processes are different, they should take a different amount of time to complete.
b. If two processes are different, they should take a same amount of time to complete.
c. If two processes are the same, they should take a different amount of time to complete.
d. The response time is easier to remember.
e. I have not been taught this material

30. If a little girl looked at the two rows of pennies below and said the second row had more, what stage of cognitive development would she be in according to Piaget?

a. formal operational
b. pre-operational
c. concrete operational
d. sensory motor
e. I have not been taught this material

31. The "common cold" of psychological problems is:

a. generalized anxiety
b. depression
c. hypochondriasis
d. obsessive-compulsive disorder
e. I have not been taught this material

32. In an attempt to study the salivary glands, this researcher was annoyed with the unusual reactions of the dog salivating before receiving food because they got in the way of his planned research. He called these reactions "psychic stimulations;" the reactions Pavlov had surreptitiously discovered are the:

a. Unconditioned stimulus
b. Unconditioned response
c. Conditioned stimulus
d. Conditioned response
e. I have not been taught this material

33. Mary was allowed to skip doing the dinner dishes because she got an 'A' on her biology exam. Mary's mother applied a _____ to encourage Mary's good behavior.

a. negative reinforcer
b. negative punishment
c. positive reinforcer
d. positive punishment
e. I have not been taught this material

34. Repression, denial, hysteria, and displacement are all examples of what method of dealing with stress?
a. defensive coping
b. direct coping
c. tolerance
d. confrontational approach
e. I have not been taught this material

35. According to Bandura, an adequate measure of efficacy would measure:
a. positive and negative feelings about oneself
b. a general belief about one's ability to do things well
c. belief that good performance will lead to specific outcomes
d. belief concerning one's ability to perform a specific task
e. I have not been taught this material

36. In an experiment on the effects of group size on helping in an emergency, group size would be the _____ variable
a. dependent
b. independent
c. confounding
d. extraneous
e. I have not been taught this material

37. Type I error is
a. the probability of rejecting the null hypothesis when it is true
b. the probability of failing to reject the null hypothesis when it is false
c. the probability of accepting the null hypothesis
d. the probability of rejecting the null hypothesis
e. I have not been taught this material

38. An operational definition of a variable is
a. the description of the conditioned response
b. limited to observable behavior only
c. one that has a causal effect on behavior or mental processes
d. describing a variable in terms of the methods used to create or measure that variable
e. I have not been taught this material

39. In an experiment studying the effects of different drugs on reaction time, each subject is presented with all of the conditions in the experiment. This is an example of a
a. factorial design
b. within subjects design
c. between subjects design
d. quasi-experimental design
e. I have not been taught this material

40. A clinical psychologist has been trying to build up the nerve to invite her client out to dinner, she should:
a. write him a letter so he isn't embarrassed
b. express her intentions verbally after the weekly counseling session with the client
c. refer the client to another doctor before initiating any relationship

d. refer the client to another doctor and remember it is unethical and illegal to date a client

e. I have not been taught this material

41. In Tolman's experiment with rats and complex mazes, he believed that the rats learned the spatial relationships in the maze without being rewarded. He called this:

a. S-R associations
b. Cognitive maps
c. Elicited response
d. Emitted response
e. I have not been taught this material

42. A deficiency in _____ has been linked to anxiety, mood disorders, and insomnia.

a. Serotonin
b. Dopamine
c. Acetocholine
d. Endorphins
e. I have not been taught this material

43. A classic study of gifted children begun by Lewis Terman more that 75 years ago is an example of a:

a. Cross-cultural study
b. Cross-sequential study
c. age cohort study
d. longitudinal study
e. I have not been taught this material

44. Reinforcing closer and closer approximations to the response we wish to condition is:

a. conditioning
b. attachment
c. shaping
d. accommodation
e. I have not been taught this material

45. The period when a behavior is measured before a treatment is introduced is called:

a. treatment
b. reversal
c. baseline
d. probe
e. I have not been taught this material

46. People playing a slot machine in a Las Vegas casino are most likely operating under what type of schedule of reinforcement?

a. Fixed ratio
b. Variable ratio
c. Fixed interval
d. Variable interval
e. I have not been taught this material

47. In an experiment, Eugene wants to find out the extent to which he can accurately state that the observed effect measured by the dependent variable is due only to variation in the independent variable. In other words, Eugene is trying to achieve

a. Construct validity
b. Ecological validity
c. External validity
d. Internal validity
e. I have not been taught this material

48. In a _____, two or more independent variables are always presented in combination.

a. random assignment
b. Between subjects design
c. Repeated measures design
d. Factorial design
e. I have not been taught this material

49. If Claire wanted to compare the test results between Mrs. Kotter's class and Mr. Jackson's class to test for significant differences, the appropriate statistical test for analyzing the results is ____.

a. linear regression
b. chi-square test
c. correlation
d. t-test
e. I have not been taught this material

50. In a within subjects design, what procedure controls for the order effects of treatment?

a. randomizing
b. matching
c. counterbalancing
d. blocking
e. I have not been taught this material

51. Suppose that we obtained a significant F value in a one-way ANOVA with p levels. Based on this result, we can conclude:

a. at least two out of p groups are different from each other.
b. All but p groups are different from each other.
c. None of the groups is different from each other.
d. We cannot conclude anything.
e. I have not been taught this material

52. The process of using sample data to answer general questions about a population is called

a. *parameter*
b. statistic
c. descriptive statistics
d. *inferential statistics*
e. I have not been taught this material

53. Determining a person's height would involve measurement on a _____ scale of measurement.

a. nominal
b. ordinal
c. interval
d. ratio
e. I have not been taught this material

54. There is a good chance that we will find a _____ correlation between time spent studying and grades in school.

a. positive
b. perfect
c. weak
d. negative
e. I have not been taught this material

55. When the null hypothesis is rejected it means that

a. the hypothesis of interest was not supported
b. the hypothesis of interest was supported
c. the theory was discredited
d. the theory was proved
e. I have not been taught this material

56. Consistency of scoring on tests is referred to as

a. validity
b. reliability
c. predictability
d. continuity
e. I have not been taught this material

57. After being able to satisfy the primary biological needs, the next level that must be satisfied according to Maslow's hierarchy of needs would be:

a. belongingness
b. self-actualization
c. safety
d. esteem
e. I have not been taught this material

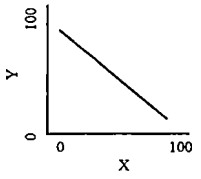
58. Which of the following is NOT a method for graphically representing frequency data?

a. range plot
b. histogram
c. bar graph
d. stem-and-leaf plot
e. I have not been taught this material

59. In a token economy, an in-patient earns tokens for healthy behaviors. Which of the following is not a reason why this type of program would be implemented?

a. To encourage patients to take responsibility for their own improvement
b. To demonstrate the value of the economy in the outside world
c. To prevent institutional neurosis
d. To allow patients visible evidence of progress
e. I have not been taught this material

60. Match the following graph with its approximate correlation coefficient.



a. R = +1.00
b. R = -1.00
c. R = .00
d. R = .5
e. I have not been taught this material

61. Bethany was held up in traffic from an accident on the freeway, resulting in her late arrival to Trisha's party. Annoyed, Trisha said, "I can't believe Bethany was so late to my party, she should have given herself more time to get here." Trisha's statement about Bethany is indicative of

a. autistic hostility
b. fundamental attribution error
c. self-serving bias
d. behavioral intentions
e. I have not been taught this material

62. If you plotted the following data points (1, 3, 5, 48, 50, 85, 95, 96, 98, 99, 100), the distribution would be

a. Bimodal
b. Negatively skewed
c. Positively skewed
d. Normal
e. I have not been taught this material

63. A therapist who helps clients become aware of destructive self-talk and dysfunctional beliefs about the self and also helps the client focus on him/herself in a more realistic and optimistic way is probably from what theoretical approach to therapy?

a. Cognitive
b. Psychoanalytic
c. Humanistic
d. Holistic
e. I have not been taught this material

64. Harry Harlow's classic study with cloth and wire monkeys illustrates that

a. food alone is sufficient to bring about attachment
b. the need to satisfy the reinforcement drive promotes the infant's attachment to the mother
c. food alone is insufficient to bring about attachment
d. satisfaction of the hunger drive nurtures the infants attachment to the mother
e. I have not been taught this material

65. Brandon was an active infant who approached strangers easily and showed curiosity easily in new situations. The characteristics reflect Brandon's:

a. personality
b. temperament
c. social referencing
d. adaptability
e. I have not been taught this material

66. When a headline reads "Lower incomes cause children to have lower IQ's," what mistake has the media made?

a. assuming causality from a correlational study
b. assuming that SES has anything to do with IQ
c. quoting from nomothetic study
d. misrepresenting an emic study
e. I have not been taught this material

67. When psychologists say that sample results will generalize to a population, they mean that what is found for the participants in a study will hold true for:

a. all people
b. future participants
c. people in the larger group from which the sample was selected
d. all the people in the geographic area who have participated in comparable studies
e. I have not been taught this material

68. Severe side effects from psychoactive drugs include losing voluntary control over the patient's own actions like eye blinks, shaky hands, and losing the ability to walk normal. These physical disabilities resulting from psychoactive drugs are characteristic of what disorder?

a. Neuroleptic Malignancy Syndrome
b. Tardive Dyskenesia
c. Senile Dementia
d. Grand Mal Epilepsy
e. I have not been taught this material

69. Bandura's classic Bobo doll experiment focused on the effects of

a. Observational learning
b. Cognitive dissonance
c. Group-think
d. Self-actualization
e. I have not been taught this material

70. "Negative feelings toward persons based on their membership in certain groups," is a good workable definition of (the)

a. discrimination
b. fundamental attribution error
c. prejudice
d. cultural bias
e. I have not been taught this material

71. in right handed individuals, aphasia is a disorder produced from damage to the left temporal lobe. It results in:

a. The inability to taste sweet foods
b. The inability to recognize and produce spoken language.
c. Paralysis of the right side of the body
d. The ability to remember past events, but no ability to form new memories
e. I have not been taught this material

72. Which of the following perspectives regards culture, ethnicity, and gender as important factors?

a. Humanistic
b. Sociocultural
c. Neurobiological
d. Psychoanalytic
e. I have not been taught this material

73. People who claim to have a psychological problem such as amnesia in order to escape responsibility are said to be

a. Suffering from conversion disorder
b. Afflicted with psychogenic amnesia
c. Malingering
d. Experiencing hypochondriasis
e. I have not been taught this material

74. What is meant by the term "double jeopardy racism"?

a. Racism in an area (e.g., the community) usually spills over into other areas (e.g., the work place)

b. Racists usually have prejudices against more than one group (e.g., prejudice against African-Americans and Jews)

c. Racism strikes twice, once because it's harder for women to find work and again because it's harder for women to get promoted when they do find work

d. Women from ethnic minorities must deal with both racism and sexism

e. I have not been taught this material

75. An individual who has a single X on the 23rd chromosome will have female genitalia but lack ovaries. This chromosomal problem is called

a. Klinefelter's syndrome
b. Turner Syndrome
c. XYY male
d. Trisomy-21
e. I have not been taught this material

76. According to Erikson, an adult concern to clean up and preserve the natural environment for future generations would fit into which stage of psychosocial development?

a. Autonomy vs. shame, doubt
b. Initiative vs. guilt
c. Industry vs. inferiority
d. Generativity vs. stagnation
e. I have not been taught this material

77. To be treated merely as a group representative rather than a talented individual defines

a. tokenism
b. group-think
c. egoism
d. altruism
e. I have not been taught this material

78. The term "glass ceiling" describes

a. a barrier to understanding employee problems
b. the effect of capping top male executives' salaries
c. a barrier to moving into management ranks
d. the upper income bracket of Fortune 500 companies
e. I have not been taught this material

79. As a member of a group, we are most likely to engage in social loafing when:

a. we experience evaluation apprehension
b. our level of arousal increases
c. we are anonymous
d. the leader is an authority figure
e. I have not been taught this material

80. Once you have conceptualized a problem in a certain way, you may find it difficult to see the problem in a different light. Cognitive psychologists refer to this as

a. fixation
b. stubbornness
c. analytical dilemma
d. divergent thinking
e. I have not been taught this material

81. Sandra Scarr's position that different parenting approaches do not substantially impact child outcomes is an example of _____ orientation toward psychological development.

a. humanistic
b. laissez faire
c. nature
d. nurture
e. I have not been taught this material

82. Research by Seligman involved shocking dogs unavoidably. When the dogs were later in a different cage where they could jump a barrier to avoid the shock, they didn't. According to the research, the dogs felt resigned to failure, acting under

a. Learned helplessness
b. Self-fulfilling prophecy
c. Operant conditioning
d. Psychomotor retardation
e. I have not been taught this material

83. Within developmental psychology, what is the major premise behind the continuity vs. discontinuity debate?

a. Developmental changes occur abruptly (in a stage-like manner) versus developmental changes occur gradually (more like a slope than like stairs)

b. Development occurs throughout life versus development discontinues after adolescence

c. Important developmental differences continue from one generation on to the next versus developmental differences discontinue after each generation

d. Mental process flow smoothly (continuity) versus mental processes are filled with pauses (discontinuity)

e. I have not been taught this material

84. Agoraphobia is the fear of

a. Spiders
b. Heights
c. Rabbits
d. Open places
e. I have not been taught this material

85. Cattell called learned skills such as the ability to do math problems and the size of the vocabulary _____.

a. fluid intelligence
b. crystallized Intelligence
c. "G" general intelligence
d. primary mental abilities
e. I have not been taught this material

86. Critics of deinstitutionalization of the mentally ill believe that it has led to

a. criminalization of the mentally ill
b. increased homelessness among the mentally ill
c. increased substance abuse among the mentally ill
d. all of the above
e. I have not been taught this material

87. Three year old Connor proudly exclaimed, "I hided under my bed!" His speech demonstrates an error in

a. Semantics
b. Syntax
c. Phonemes
d. Pragmatics
e. I have not been taught this material

88. According to the cycle-of-violence hypothesis, abuse and neglect of children leads them to be

a. victims in adult life
b. insecure and withdrawn adults
c. more sympathetic to others as adults
d. predisposed to abusiveness as adults
e. I have not been taught this material

89. A researcher studies how individuals born during the 1920s compare to individuals born during the 1940s with regard to political participation. The individuals within the groups are called:

a. cohorts
b. alliances
c. support systems
d. reference groups
e. I have not been taught this material

90. Substance abuse addictions may be classified along a continuum of dependence rather than on an absolute basis. This continuum (in order) is:

a. usage, dependence, tolerance
b. indulgence, dependence, gratification
c. usage, abuse, addiction
d. gratification, tolerance, dependence
e. I have not been taught this material

91. Name given to a person who unwittingly supports and reinforces another person's addictive behavior:

a. enabler
b. dealer
c. co-addict
d. sponsor
e. I have not been taught this material

92. The main reason California has outlawed the use of intelligence tests as the sole determinant to place children in special classrooms is

a. The tests don't really measure intelligence
b. The cultural bias of the test misrepresents the mental abilities of minority children
c. The tests don't measure achievement skills in language and computational ability
d. The tests were emphasizing the genetic differences in mental ability between Caucasian, Asian, Hispanic and African-American children too much.
e. I have not been taught this material

93. If you can clearly remember a unique event like kicking the game winning goal when you were in the youth soccer tournament, Anderson would call this first hand experience

a. episodic knowledge
b. semantic knowledge
c. procedural knowledge
d. declarative knowledge
e. I have not been taught this material

94. Milgram's obedience experiment, where subjects delivered an electrical shock to confederates to teach them word pairs.

a. demonstrated decreased learning with the use of punishment
b. demonstrated increased learning with the use of punishment
c. could be replicated today
d. could not be replicated today
e. I have not been taught this material

95. The sum of an individual's beliefs or schemas about his or her personal traits and characteristics is termed

a. the self-concept
b. ego
c. social perception
d. self-awareness
e. I have not been taught this material

96. The most significant purpose of the IRB approval process is:

a. to let the school know what experiments are being conducted
b. to protect participants from unethical testing
c. To provide strict regulations and control psychological research
d. To allow researchers to practice writing proposals for studies

e. I have not been taught this material

97. The tendency to assume that people that have one positive characteristic must have others as well is called

a. misinformation effect
b. the halo effect
c. error of inference
d. familiarity
e. I have not been taught this material

98. An example of a secondary sex characteristic would be:

a. ovaries
b. sexual orientation
c. facial hair
d. genitals
e. I have not been taught this material

99. After subjects give their informed consent to participate in a research study they

a. can withdraw from the study at any time
b. must finish the entire research study
c. can debrief the experimenter if they so choose
d. must be paid at the end of the study
e. I have not been taught this material

100. What part of the body prepares the person for flight or fight?

a. amygdala
b. limbic system
c. parasympathetic nervous system
d. sympathetic nervous system
e. I have not been taught this material

APPENDIX D Continued
Debrief Statement
(Pilot & Main Study)

Thank you for taking your time to participate in this study. Your time is greatly appreciated. Tammy West Reichley, Masters Student of Industrial/Organizational Psychology, California State University San Bernardino, is conducting this study in part for her Master's thesis on Psychology Outcomes Assessment, under the supervision of Dr. Janet Kottke. The purpose of this research is to develop a comprehensive exam for graduating seniors in the Psychology Department of CSUSB.

To be qualified as a participant, you must be at least 18 years old and a student of CSUSB classified in any Major.

Your participation included completing an exam testing your knowledge of Psychology. All of your responses will remain anonymous and will be used for research purposes only. Participation in this study is completely voluntary and if you will not be penalized if you withdraw.

This study has been approved by the Institutional Review Board at California State University, San Bernardino. If you have any questions, please contact Tammy West Reichley at (626) 302-5324.

REFERENCES

Anderson, W.P., Krauskopf, C.J., Rogers, M.E., & Neal, G.W. (1984). Reasons for comprehensive examinations: A re-evaluation. Teaching of Psychology, 11 (2), 78-82.

Assessment Systems Corporation (1995). ITEMAN conventional item analysis program user's manual. St. Paul Minnesota.

Assessment Systems Corporation (1995). RASCAL Rasch analysis program user's manual. St. Paul Minnesota.

Astin, A.W. (1987). Assessment, value-added, and educational excellence. In D.T. Halpern (Ed.), Student outcomes assessment: What institutions stand to gain (pp. 89-107). Edited San Francisco: Jossey-Bass, Inc.

Banta, T.W. & Moffet, M.S. (1987). Performance funding in Tennessee: Stimulus for program improvement. In D.T. Halpern (Ed.), Student outcomes assessment: What institutions stand to gain (pp. 35-44). San Francisco: Jossey-Bass, Inc.

Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). Evaluation to improve learning. McGraw-Hill, Inc.

Bobko, P. (1995). Correlation and regression: Principles and applications for Industrial/Organizational Psychology and management (pp. 37-38). McGraw-Hill, Inc.

Brown, F.G. (1983). Principles of educational and psychological testing (pp. 186-287). New York: Holt, Rinehart, and Winston.

Camera, W.J. & Brown, D.C. (1995). Educational and employment testing: Changing concepts in measurement and policy. Educational Measurement: Issues and Practice, Spring.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. Orlando: Harcourt Brace Jovanovich College Publishers.

California State University Psychology Outcomes Assessment Meeting. (1999). California State University, Bakersfield.

Curry, W. & Hager, E. (1987). Assessing general education: Trenton State College In D.T. Halpern (Ed.), Student outcomes assessment: What institutions stand to gain (pp. 57-65). San Francisco: Jossey-Bass, Inc.

Donnelly, J.P. Jr. (1994). Item analysis with ITEMAN and RASCAL. Measurement and Evaluation in Counseling and Development, 26 (4), 269-274.

Drasgow, F. & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hooper (Eds), Handbook of Industrial/Organizational Psychology (2nd ed.)(pp. 577-636). Palo Allto, CA: Consulting Psychologists Press.

Dwyer, C.A. (1998). Assessment and classroom learning: Theory and practice. Assessment in Education, 5 (1), 131-138.

Ebel, R.L. (1980). Practical problems in educational measurement (pp. 24-35, 264-266). Lexington, MA: D.C. Heath and Company.

Erwin. T.D. (1991). Assessing student learning and development: A guide to the principles, goals, and methods of determining college outcomes. San Francisco: Jossey-Bass Publishers.

Ewell, P.T. (1987). Establishing a campus-based assessment program. In Student outcomes assessment: What institutions stand to gain. Edited by D.T. Halpern (pp. 9-24). San Francisco: Jossey-Bass, Inc.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58 (3), 357-381.

Goldstein, I.L. (1993), Training in Organizations (3rd ed) (p. 114). Pacific Grove, CA: Brooks/Cole Publishing Company.

Halpern, D.F., Ed. (1987). Student outcomes assessment: What institutions stand to gain (pp. 1-8). San Francisco: Jossey-Bass Publishers.

Halpern, D.F. (1988). Assessing student outcomes for psychology majors. Journal of Educational Psychology, 15 (4), 181-186.

Halpern, D.F., Appleby, D.C., Beers, S.E., Cowan, C.L., Furedy, J.J., Halonen, J.S., Horton, C.P., Peden, B.F., and Pittenger, D.J. (1993). In T.V. McGovern (Ed.), Handbook for enhancing undergraduate education in psychology (pp. 23-46). Washington DC: American Psychological Association.

Halpin, G & Halpin, G. (1982). Experimental investigation of the effects of study and testing on student learning, retention, and ratings of instruction. Journal of Educational Psychology, 74 (1), 32-38.

Hattie, J. & Jaegar, R. (1998). Assessment and classroom learning: A deductive approach. Assessment in Education, 5 (1), 111-113.

Juhler, S.M., Rech, J.F., From, S.G., & Brogan, M.M. (1998). The effect of optional re-testing on college students' achievement in an individualized algebra course. The Journal of Experimental Education, 66 (2), 125-138.

Kimmel, S.L., Marquette, R.P., & Olsen, D.H. (1998). Outcomes assessment programs: Historical perspective and state of the art. Issues in Accounting Education, 13 (4) 851-869.

Krathwohl, D.R., Bloom, B.S., & Masia, B.B. (1956). Taxonomy of educational objectives: The classification of educational goals. New York: David McKay Company, Inc.

Krueger, D.W. & Heisserer, M.L. (1987). Assessment and involvement: Investments to enhance learning. In D.T. Halpern (Ed.), Student outcomes assessment: What institutions stand to gain (pp. 45-56). San Francisco: Jossey-Bass, Inc.

Landy, F.J. (1986). Stamp collecting versus science: Validation as hypothesis testing. American Psychologist, 41 (11), 1183-1192.

Lawler, E.E. (1994). Motivation in Work Organizations p. 25-28.
San Francisco: Jossey-Bass, Inc.

Lien, A.J. (1971). Measurement and evaluation of learning (3rd ed.). Dubuque IA: W.C. Brown Co.

Loughead, T.O. (1997). The doctoral comprehensive examination: Fine-tuning the process. Counselor Education and Supervision, 37 (2), 140-149.

Magill, S.L. (1998). Using educational outcomes and student portfolios to steer management education. Journal of Management Education, page numbers not available.

Messick, S. (1995). Validity of psychological assessment: Validation inferences from persons' responses and performance as scientific inquiry into score meaning. American Psychologist, 50 (9), 741-749.

Ndalichako, J.L. & Rogers, W.T. (1997). Comparison of finite state score theory, and item response theory in scoring multiple choice items. Educational & Psychological Measurement 57 (4), 580-590.

Ogletree, S.M. (1998). Using alumni surveys to evaluate and improve the psychology major. College Student Journal, 32, 215-221.

Popham, W.J. (1981). Modern educational measurement (p. 5). Englewood Cliffs, N.J.: Prentice Hall.

Psychology 644 (1998). Applied Psychological Measurement. Class handout, Dr. Jan Kottke.

Resnik and Goulden (1987). Assessment, curriculum, and expansion: A historical perspective. In D.T. Halpern (Ed.), Student outcomes assessment: What institutions stand to gain (pp. 77-89). San Francisco: Jossey-Bass, Inc.

Sebatame, E.M. (1998). Assessment and classroom learning: A response to Black and William. Assessment in Education, 5, (1), 123-131.

Sheehan, E.P. (1994). A multimethod assessment of the psychology major. Teaching of Psychology, 21 (2), 74-78.

Sheehan, E.P. & Granrud, C.E. (1995). Assessment of student outcomes: Evaluating an undergraduate psychology program. Journal of Instructional Psychology, 22, 366-372.

Suen, H.K. (1990). Principles of test theories (pp. 6-101). Hillside, NJ: Lawrence Erlbaum.

Tuckman, B.W. (1998). Using tests as an incentive to motivate procrastinators to study. Journal of Experimental Education, 66 (2), 141-148.

Vandament, W.E. (1987). A state university perspective on student outcomes assessment. In D.T. Halpern (Ed.), Student outcomes assessment: What institutions stand to gain 25-28. San Francisco: Jossey-Bass, Inc.

Weiss, D.J. & Davison, M.L. (1981). Review of test theory and methods. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Willis, D. (1994). School-based assessment: Underlying ideologies and their implications for teachers and learners. New Zealand Journal of Educational Studies, 29 (2), 161-174.