**Theory and Practice of Large Scale Assessment**

|  |  |
| --- | --- |
| Program | **Doctor of Philosophy in Education** |
| Course Code | **ED 766** |
| Credit Hours | **3** |
| Duration | **15 \* 2 = 30 sessions** |
| Prerequisites | **None** |
| Resource Person | **Dr. Fariha Gull** |
| Counseling Timing |  |
| Contact | E-mail: [fariha.gul@umt.edu.pk](mailto:fariha.gul@umt.edu.pk) |

**Chairman/Director Programme signature………………. Dean’s signature…………**

**Introduction**

This course provides an introduction to large-scale testing, presenting an overview of the various tasks that employees perform in testing organizations, city-wide testing bureaus, professional licensing and certification boards, statewide educational testing programs, testing units that are part of state merit systems, etc. The course should be useful for students considering working for such organizations in a variety of capacities (e.g., item writers, statisticians, psychometricians, researchers, testing program managers), employees currently working in these organizations who would like to increase their understanding of the field, and students who want to gain an understanding of the challenges of creating and administering large-scale tests.

**Learning Outcomes**

Following are the aims and objectives of the course;

* the various tasks involved in constructing a large-scale test (e.g., defining a content domain; conducting a job analysis or practice analysis to provide the basis for a test framework; writing test specifications, items, and scoring procedures; assembling item pools and tests; conducting reviews of items for content and fairness considerations; piloting/field testing items; evaluating the quality of items in an item pool and the psychometric properties of a test)
* documents guiding the design and use of large-scale tests (Standards for Educational and Psychological Testing, Code of Fair Testing Practices in Education, Principles for the Validation and Use of Personnel Selection Procedures, Uniform Guidelines on Employee Selection Procedures, Code of Professional Responsibilities in Educational Measurement, Ethical Standards of the American Educational Research Association, Ethical Principles of Psychologists and Code of Conduct, The Rights and Responsibilities of Test Takers)
* key terms used in writing and discussing large-scale testing
* reliability, validity, and errors of measurement and their roles in the construction, selection, interpretation, and use of large-scale tests
* different approaches for scoring, scaling, norming, equating, and linking large-scale tests
* different methods for setting defensible standards/cut scores on large-scale tests
* the steps involved in standardizing a test (e.g., preparing standardized directions, administering tests under standardized conditions, employing standardized scoring procedures, developing materials to interpret test scores)
* what is involved in preparing supporting documentation for a test so that test users will have the information they need to make sound judgments regarding the test’s nature and quality, the scores it produces, as well as the interpretations based on the scores
* ways that large-scale tests can be modified to accommodate individuals with various disabilities, and individuals from diverse linguistic backgrounds
* issues of fairness in large-scale testing
* the rights and responsibilities of test takers
* the responsibilities of test users
* key issues that arise in large-scale testing
* key court cases, laws, and legislation that affect large-scale testing programs

**Textbook:**

Following book on educational assessment and measurement will be used as textbook for the course. Copy of this book can be borrowed from library or can be collected from course instruction for photocopy.

Standards for educational and psychological testing. Washington, DC: American Educational Research Association. ISBN: 0-935302-35-6

**Recommended material for further reading:**

1. Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). Handbook of test development. New York, NY: Taylor & Francis. ISBN: 978-0-415- 62602-6
2. Brennan, R. L. (Ed.). (2006). Educational measurement (4th ed). Westport, CT: Praeger. ISBN: 0-275-98125-8
3. Wise, L. L., & Plake, B. S. (2016). Test design and development following the standards for educational and psychological testing. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.). (2016). Handbook of test development (pp. 19-39). New York, NY: Taylor & Francis.
4. Perie, M. & Huff, K. (2016). Determining content and cognitive demand for achievement tests. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.). (2016). Handbook of test development (pp. 119-143). New York, NY: Taylor & Francis. (Note: read this if you are interested in getting more information about evidence-centered design and the use of performance-level descriptors (esp. in K-12 assessment)
5. Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 307-324). Westport, CT: Praeger Publishers.
6. Raymond, M. R. (2016). Job analysis, practice analysis, and the content of credentialing examinations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.). (2016). Handbook of test development (pp. 144-164). New York, NY: Taylor & Francis
7. Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. Applied Measurement in Education, 14(4), 369-415.
8. Raymond, M. R. (2002). A practical guide to practice analysis for credentialing examinations. Educational Measurement: Issues and Practice, 21(3), 25-37.
9. Raymond, M. R. (2005). An NCME instructional module on developing and administering practice analysis questionnaires. Educational Measurement: Issues and Practice, 24(2), 29-42.
10. Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 324-353). Westport, CT: Praeger Publishers. (Note: Read the sections of this chapter entitled “Item Development,” Item Review,” “Item Evaluation and Test Assembly,” “Test Review,” “Item Banking,” and “Quality Control.”)
11. Plake, B. S. (1995). Differential item functioning in licensure tests. In J. C. Impara (Ed.), Licensure testing: Purposes, procedures and practices (pp. 205-218). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.
12. Kane, M. T. (2016). Validation strategies. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.). (2016). Handbook of test development (pp. 64-79). New York, NY: Taylor & Francis.
13. Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? Educational Measurement: Issues and Practice, 29(1), 3-13.
14. Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. Educational Measurement: Issues and Practice, 28(1), 3-9.
15. Parkes, J. (2007) Reliability as argument. Educational Measurement: Issues and Practice, 26(4), 2-10. Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. Educational Measurement: Issues and Practice, 10(1), 37-45.
16. McCallin, R. C. (2016). Test administration. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.) Handbook of test development 2nd edition (pp. 567-584). New York, NY: Routledge. Zenisky, A. L., & Hambleton, R. K. (2016). A model for good practice in score reporting. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.) Handbook of test development 2nd edition (pp. 585-602). New York, NY: Routledge.
17. Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. Educational Measurement: Issues and Practice, 26(1), 36-43.
18. Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 355-386). Westport, CT: Praeger Publishers.
19. Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.
20. Kolen, M. J., & Brennan, R. L. (1995). Introduction and concepts. In Test equating: Methods and practices (pp. 1- 27). New York: Springer-Verlag, Inc.
21. Tong, Y., & Kolen, M. J. (2010). Scaling: An ITEMS module. Educational Measurement: Issues and Practice, 29(4), 39-48.
22. Cizek, G. J. & Earnest, D. S. (2016). Setting performance standards on tests. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.) Handbook of test development 2nd edition (pp. 212-237). New York, NY: Routledge
23. Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 433-470). Westport, CT: Praeger Publishers.
24. Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 221-256). Westport, CT: Praeger Publishers.
25. Cizek. G. J. (1999). Detecting cheating on tests. In Cheating on tests: How to do it, detect it, and prevent it (pp. 127- 150). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
26. Cizek. G. J. (1999). Responding to cheating. In Cheating on tests: How to do it, detect it, and prevent it (pp. 151- 162). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
27. Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. Applied Measurement in Education, 17(2), 145-220.
28. Hambleton, R. K. (2007, June 18). A new challenge: Making test score reports more understandable and useful [PowerPoint presentation].

**Grade Evaluation Criteria**

Following is the criteria for the distribution of marks to evaluate final grade in a semester.

**Marks Evaluation Marks in percentage**

Class activities 15

Mid Term exam 25

Class presentation 30

Final term Project 30

Total **100**

**Calendar of Course contents to be covered during semester**

|  |  |  |
| --- | --- | --- |
| **Week** | **Activity objectives of lesson**  **Assessment procedure** | **Reference** |
| **1** | Test Design and Development | Purpose of test and development of test specification |
| **2** | Test Design and Development | Conducting job and practice analyses Defining a test framework |
| **3** | Test Design and Development | Writing items and scoring procedures Item selection procedures; assembling item pools Expert reviews of items for content concerns Sensitivity/fairness review procedures Field testing/pilot testing procedures |
| **4** | Determining the psychometric | properties of items and of the test Test assembly Writing instructions to test takers Writing standardized procedures for test administration Conducting studies of differential item functioning |
| **5** | Validity | Definitions of validity Establishing a conceptual framework for investigating validity Making decisions about types of validity evidence to gather Articulating propositions to test and conducting logical analyses to evaluate those propositions Sources of validity evidence Integrating validity evidence Developing a “validity argument” Construct underrepresentation, construct-irrelevant variance |
| **6** | Reliability I | Precision and Errors of Measurement Definitions of reliability and measurement error Flexibility large-scale testing |
| **7** | Reliability II | large-scale testing and increasing measurement error Characteristics of measurement error Summarizing reliability data Interpreting reliability data |
| **8** | ***Midterm*** |  |
| **9** | Test Administration, Scoring, Reporting, and Interpretation Adhering to standardized test administration procedures | Monitoring test takers to eliminate opportunities for cheating Managing testing environments Protecting the security of test materials Training scorers Assuring accuracy of scoring Providing appropriate interpretations of scores Protecting confidentiality when transmitting scores |
| **10** | Scores, Scales, Norms, Score Linking, and Cut Scores – Part 1 | Scaling Norming Equating Linking tests that measure different constructs to establish score comparability |
| **11** | Scores, Scales, Norms, Score Linking, and Cut Scores Establishing defensible cut scores | Scales, Norms, Score Linking, and Cut Scores Establishing defensible cut scores |
| **12** | Fairness in Testing | General views of fairness Threats to fair and valid interpretations of test scores Minimizing construct-irrelevant components through test design and testing adaptations |
| **13** | The Rights and Responsibilities of Test Users | Test security Access to test results and to understandable information about a test Testing irregularities Informed consent requirements Protection from improper disclosure Forms of cheating |
| **14** | Psychological Testing and Assessment Test selection and administration Test score interpretation Collateral information used in psychological testing and assessment Types of psychological testing and assessment Purposes of psychological testing and assessment | Test score interpretation Collateral information used in psychological testing and assessment Types of psychological testing and assessment Purposes of psychological testing and assessment |
| **15** | Uses of Tests for Program Evaluation, Policy Studies, and Accountability | Evaluation of programs and policy initiatives Test-based accountability systems Issues in program and policy evaluation and accountability Additional considerations |
| **16** | ***Final Term Exams*** |  |