References

Advisory Panel on the Scholastic Aptitude Test Score Decline. (1977). *On further examination.* New York: College Entrance Examination Board.

Allen, N.S., Holland, P.W., & Thayer, D. (1994a). *A missing data approach to estimating distributions of scores for optional test sections* (Research Report 94-17). Princeton, NJ: Educational Testing Service.

Allen, N.S., Holland, P.W., & Thayer, D. (1994b). *Estimating scores for an optional section using information from a common section* (Research Report 94-18). Princeton, NJ: Educational Testing Service.

American College Testing Program (ACT) (1989). *Preliminary technical manual for the Enhanced ACT Assessment.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME) (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

American Psychological Association (APA) (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Andrulis, R.S., Starr, L.M., & Furst, L.W. (1978). The effects of repeaters on test equating. *Educational and Psychological Measurement, 38,* 341–349.

Angoff, W.H. (1953). Test reliability and effective test length. *Psychometrika, 18,* 1–14.

Angoff, W.A. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education. (Reprinted as W. A. Angoff, *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service, 1984.)

Angoff, W.H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291–300.

Angoff, W.H., & Cowell, W.R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23,* 327–345.

Baker, F.B. (1992a). *Item response theory parameter estimation techniques.* New York: Marcel Dekker.

Baker, F.B. (1992b). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F.B. (1993). Equate 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement*, *17*, 20.
Baker, F.B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*, 147–162.
Baker, E.L., O'Neil, H.F., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*, 1210–1218.
Baxter, G.P., Shavelson, R.J., Goldman, S.R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, *29*, 1–17.
Beaton, A.E., & Gonzalez, E. (1994). *Comparing the NAEP trial state assessment results with the IAEP international results.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
Bishop, Y.M.M., Fienberg, S.E, & Holland, P.W. (1975). *Discrete multivariate analysis. Theory and practice.* Cambridge, MA: MIT Press.
Blommers, P.J., & Forsyth, R.A. (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Boston: Houghton Mifflin.
Bloxom, B., & McCully, R. (1992). *Initial operational test and evaluation of forms 18 and 19 of the Armed Services Vocational Aptitude Battery.* Monterey, CA: Defense Manpower Data Center.
Bloxom, B., McCully, R., Branch, R., Waters, B.K., Barnes, J., & Gribben, M. (1993). *Operational calibration of the circular-response optical-mark-reader answer sheets for the Armed Services Vocational Aptitude Battery (ASVAB).* Monterey, CA: Defense Manpower Data Center.
Bloxom, B., Pashley, P.J., Nicewander, W.A., & Yan, D. (1995). Linking to a large scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20*, 1–26.
Brandenburg, D.C., & Forsyth, R.A. (1974). Approximating standardized achievement test norms with a theoretical model. *Educational and Psychological Measurement*, *34*, 3–9.
Braun, H.I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, *13*, 1–18.
Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic.
Brennan, R.L. (Ed.) (1989). *Methodology used in scaling the ACT Assessment and P-ACT+.* Iowa City, IA: American College Testing.
Brennan, R.L. (1990). *Congeneric models and Levine's linear equating procedures* (ACT Research Report 90-12). Iowa City, IA: American College Testing.
Brennan, R.L. (1992). The context of context effects. *Applied Measurement in Education*, *5*, 225–264.
Brennan, R.L., & Kolen, M.J. (1987a). A reply to Angoff. *Applied Psychological Measurement*, *11*, 301–306.
Brennan, R.L., & Kolen, M.J. (1987b). Some practical issues in equating. *Applied Psychological Measurement*, *11*, 279–290.
Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement*, *22*, 13–20.
Budescu, D. (1987). Selecting an equating method: Linear or equipercentile? *Journal of Educational Statistics*, *12*, 33–43.
Burke, E.F., Hartke, D., & Shadow, L. (1989). *Print format effects on ASVAB test score performance: Literature review* (AFHRL Technical Paper 88-58). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
Carlin, J.B., & Rubin, D.B. (1991). Summarizing multiple-choice tests using three informative statistics. *Psychological Bulletin*, *110*, 338–349.

Cizek, G.J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*, 8–20.

Cohen, A.S., & Kim, S.H. (1993). *A comparison of equating methods under the graded response model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Congressional Budget Office. (1986). *Trends in educational achievement.* Washington, DC: Author.

Cook, L.L. (1994). *Recentering the SAT score scale: An overview and some policy considerations.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Cook, L.L., & Eignor, D.R. (1991). An NCME instructional module on IRT equat-

Cook, L.L., & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225–244.

Cope, R.T. (1986). *Use versus nonuse of repeater examinees in common item linear equating with nonequivalent populations* (ACT Technical Bulletin 51). Iowa City, IA: American College Testing.

Cope, R.T., & Kolen, M.J. (1990). *A study of methods for estimating distributions of test scores* (ACT Research Report 90-5). Iowa City, IA: American College Testing.

Crouse, J.D. (1991). *Comparing the equating accuracy from three data collection designs using bootstrap estimation methods.* Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

Cureton, E.F., & Tukey, J.W. (1951). Smoothing frequency distributions, equating tests, and preparing norms. *American Psychologist*, *6*, 404.

Darroch, J.N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, *43*, 1470–1480.

de Boor, C. (1978). *A practical guide to splines* (Applied Mathematical Sciences, Volume 27). New York: Springer-Verlag.

Divgi, D.R. (1985). A minimum chi-square method for developing a common metric

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

Eignor, D.R. (1985). *An investigation of the feasibility and practical outcomes of preequating the SAT verbal and mathematical sections* (Research Report 85-10). Princeton, NJ: Educational Testing Service.

Eignor, D. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT* (Research Report 93-55). Princeton, NJ: Educational Testing Service.

Eignor, D.R., & Stocking, M.L. (1986). *An investigation of the possible causes for the inadequacy of IRT preequating* (Research Report 86-14). Princeton, NJ: Educational Testing Service.

Eignor, D.R., Way, W.D., & Amoss, K.E. (1994). *Establishing the comparability of the NCLEX using CAT with traditional NCLEX examinations.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*, 171–191.

Englehard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Ercikan, K. (1994). *Linking state tests to NAEP.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Fairbank, B.A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement*, *11*, 245–262.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Ferrara, S. (1993). *Generalizability theory and scaling: Their roles in writing assessment and implications for performance assessments in other content areas.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Fitzpatrick, A.R., & Yen, W.M. (1993). *The psychometric characteristics of choice items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Fitzpatrick, A.R., Ercikan, K., Yen, Y.M., & Ferrara, S. (1994). *The consistency between ratings collected in different test years.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Gilmer, J.S. (1989). The effects of test disclosure on equated scores and pass rates. *Applied Psychological Measurement*, *13*, 245–255.

Gordon, B., Englehard, G., Gabrielson, S., & Bernkopf, S. (1993). *Issues in equating*

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.
Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Boston: Kluwer.
Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.
Han, T. (1993). *Comparison of IRT observed-score equating with both IRT true-score and classical equipercentile equating.* Unpublished doctoral dissertation, Southern Illinois University, Carbondale, IL.
Hanson, B.A. (1989). Scaling the P-ACT+. In R.L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 57–73). Iowa City, IA: American College Testing.
Hanson, B.A. (1990). *An investigation of methods for improving estimation of test score distributions* (ACT Research Report 90–4). Iowa City, IA: American College Testing.
Hanson, B.A. (1991a). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics, 16*, 93–100.
Hanson, B.A. (1991b). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (ACT Research Report 91–5). Iowa City, IA: American College Testing.
Hanson, B.A. (1991c). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*, 391–408.
Hanson, B.A. (1992). *Testing for differences in test score distributions using log-linear models.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
Hanson, B.A. (1993). *A missing data approach to adjusting writing sample scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.
Hanson, B.A., Zeng, L., & Kolen, M.J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17*, 225–237.
Hanson, B.A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Research Report 94-4). Iowa City, IA: American College Testing.
Harnischfeger, A., & Wiley, D.E. (1975). *Achievement test score decline: Do we need to worry?* Chicago: CEMREL.
Harris, D.J. (1986). A comparison of two answer sheet formats. *Educational and Psychological Measurement, 46*, 475–478.
Harris, D.J. (1987). *Estimating examinee achievement using a customized test.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
Harris, D.J. (1988). *An examination of the effect of test length on customized testing using item response theory.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
Harris, D.J. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*, 35–41.
Harris, D.J. (1991a). *Equating with nonrepresentative common item sets and nonequivalent groups.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.
Harris, D.J. (1991b). *Practical implications of the context effects resulting from the use of scrambled test forms.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.
Harris, D.J. (1993). *Practical issues in equating.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Harris, D.J., & Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195–240.
Harris, D.J., & Kolen, M.J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*, 35–43.
Harris, D.J., & Kolen, M.J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement, 50*, 61–71.
Harris, D.J., & Welch, C.J. (1993). *Equating writing samples.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.
Harris, D.J., Welch, C.J., & Wang, T. (1994). *Issues in equating performance assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
Holland, P.W., & Rubin, D.B. (1982) *Test equating.* New York: Academic.
Holland, P.W., & Thayer, D.T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Technical Report 87-79). Princeton, NJ: Educational Testing Service.
Holland, P.W., & Thayer, D.T. (1989). *The kernel method of equating score distributions* (Technical Report No. 89-84). Princeton, NJ: Educational Testing Service.
Holland, P.W., & Thayer, D.T. (1990). *Kernel equating and the counterbalanced design.* Paper presented at the annual meeting of the American Educational Research Association, Boston.
Holland, P.W., King, B.F., & Thayer, D.T. (1989). *The standard error of equating for the kernel method of equating score distributions* (Technical Report 89-83). Princeton, NJ: Educational Testing Service.
Houston, W., & Sawyer, R. (1991). Relating scores on the Enhanced ACT Assessment and the SAT test batteries. *College and University, 66*, 195–200.
Hung, P., Wu, Y., & Chen, Y. (1991). *IRT item parameter linking: Relevant issues for the purpose of item banking.* Paper presented at the International Academic Symposium on Psychological Measurement, Tainan, Taiwan.
Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. *Journal of Educational Measurement, 31*, 125–141.
IMSL (1991). *Fortran subroutines for mathematical applications. Math/library.* Houston, TX: Author.
Jaeger, R.M. (1981). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement, 18*, 23–38.
Jarjoura, D., & Kolen, M.J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10*, 143–160.
Keats, J.A., & Lord, F.M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59–72.
Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). New York: Macmillan.
Kiplinger, V.L., & Linn, R.L. (1994). *Linking statewide tests to the National Assessment of Educational Progress: Stability of results.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
Klein L.W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197–206.
Kolen, M.J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1–11.
Kolen, M.J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25–44.

Kolen, M.J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement*, *9*, 209–223.

Kolen, M.J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice*, *7*, 29–36.

Kolen, M.J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement*, *28*, 257–282.

Kolen, M.J., & Brennan, R.L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement*, *11*, 263–277.

Kolen, M.J., & Harris, D.J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, *27*, 27–39.

Kolen, M.J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika*, *52*, 43–59.

Kolen, M.J., Hanson, B.A., & Brennan, R.L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*, 285–307.

Kolen, M.J., Zeng, L., & Hanson, B.A. (1994). *Conditional standard errors of measure-*

Lord, F.M. (1982a). Item response theory and equating—A technical summary. In P.W. Holland and D.B. Rubin (Eds.), *Test equating* (pp. 141–148). New York: Academic.

Lord, F.M. (1982b). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*, 165–174.

Lord, F.M. (1982c). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6*, 463–472.

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

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 452–461.

Loyd, B.H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education, 4*, 11–22.

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement, 17*, 179–193.

Loyd, B., Englehard, G., & Crocker, L. (1993). *Equity, equivalence, and equating: Fundamental issues and proposed strategies for the National Board for Professional Teaching Standards.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

MacCann, R.G. (1990). Derivations of observed score equating methods that cater to populations differing in ability. *Journal of Educational Statistics, 15*, 146–170.

Maier, M.H. (1993). *Military aptitude testing: The past fifty years* (DMDC Technical Report 93-007). Monterey, CA: Defense Manpower Data Center.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Marco, G.L. (1981). Equating tests in the era of test disclosure. In B.F. Green (Ed.), *New directions for testing and measurement: Issues in testing—coaching, disclosure, and ethnic bias* (pp. 105–122). San Francisco: Jossey-Bass.

Marco, G.L., & Abdel-Fattah, A.A. (1991). Developing concordance tables for scores on the Enhanced ACT Assessment and the SAT. *College and University, 66*, 187–194.

Marco, G.L., Petersen, N.S., & Stewart, E.E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.), *New horizons in testing* (pp. 147–176). New York: Academic.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Mazzeo, J., & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests. A review of the literature* (College Board Report 88-8). New York: College Entrance Examination Board.

Mazzeo, J., Druesne, B., Raffeld, P.C., Checketts, K.T., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations* (College Board Report 91-5). New York: College Entrance Examination Board.

McKinley, R.L., & Schaeffer, G.A. (1989). *Reducing test form overlap of the GRE* *troduction and comparability of the computer adaptive GRE general test.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Mislevy, R.J., & Stocking, M.L. (1989). A consumers guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–75.

Morgan, R., & Stevens, J. (1991). *Experimental study of the effects of calculator use in the advanced placement calculus examinations* (Research Report 91-5). Princeton, NJ: Educational Testing Service.

Morris, C.N. (1982). On the foundations of test equating. In P.W. Holland and D.B. Rubin (Eds.), *Test equating* (pp. 169–191). New York: Academic.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Parshall, C.G., & Kromrey, J.D. (1993). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Pashley, P.J. & Phillips, G.W. (1993). *Toward world class standards. A research study linking international and national assessments.* Princeton, NJ: Educational Testing Service.

Pashley, P.J., Lewis, C., & Yan, D. (1994). *Statistical linking procedures for deriving point estimates and associated standard errors.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137–156.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Press, W.H., Flannery, B.P., Teukolsky, S.A., & Vetterling, W.T. (1989). *Numerical recipes. The art of scientific computing (Fortran version).* Cambridge, UK: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raymond, M.R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement, 30,* 253–268.

Reese, C. (1992). *Development of a computer-based test for the GRE general test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Reinsch, C.H. (1967). Smoothing by spline functions. *Numerische Mathematik, 10,* 177–183.

Rosenbaum, P.R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40,* 43–49.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometrika Monograph No. 17) Richmond, VA Psychometrics Society.

Segall, D.O. (1993). *Score equating verification analyses of the CAT-ASVAB.* Briefing presented to the Defense Advisory Committee on Military Personnel Testing, Williamsburg, VA.

Skaggs, G. (1990). *Assessing the utility of item response theory models for test equat-*

ing. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Skaggs, G., & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495–529.

Spray, J.A., Ackerman, T.A., Reckase, M.D., & Carlson, J.E. (1989). Effect of medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261–271.

Stocking, M.L., & Eignor, D.R. (1986). *The impact of different ability distributions on IRT preequating* (Research Report 86-49). Princeton, NJ: Educational Testing Service.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Tenopyr, M.L., Angoff, W.H., Butcher, J.N., Geisinger, K.F., & Reilly, R.R. (1993). Psychometric and assessment issues raised by the Americans with Disabilities Act (ADA). *The Score, 15*, 1–15.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501–519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577.

Thomasson, G.L., Bloxom, B., & Wise, L. (1994). *Initial operational test and evaluation of forms 20, 21, and 22 of the Armed Services Vocational Aptitude Battery (ASVAB)* (DMDC Technical Report 94-001). Monterey, CA: Defense Manpower Data Center.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*, 15–20.

Wainer, H., & Mislevy, R.J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65–102). Hillsdale, NJ: Erlbaum.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103–118.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159–195.

Wainer, H., Dorans, N.J., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 233–286). Hillsdale, NJ: Erlbaum.

Wainer, H., Wang, X., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees choice? *Journal of Educational Measurement, 31*, 183–199.

Wang, T., & Kolen, M.J. (1994). *A quadratic curve equating method to equate the first three moments in equipercentile equating* (ACT Research Report 94-2). Iowa City, IA: American College Testing.

Way, W.D., & Tang, K.L. (1991). *A comparison of four logistic model equating methods*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Way, W.D., Forsyth, R.A., & Ansley, T.N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Applied Measurement in Education, 2*, 15–35.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST users guide*. Princeton, NJ: Educational Testing Service.

Wingersky, M.S., Cook, L.L., & Eignor, D.R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (Research Report 87–24). Princeton, NJ: Educational Testing Service.

Woodruff, D.J. (1986). Derivations of observed score linear equating methods based on test score models for the common item nonequivalent populations design. *Journal of Educational Statistics, 11*, 245–257.

Woodruff, D.J. (1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 13*, 257–261.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

Yen, W.M. (1983). Tau-equivalence and equipercentile equating. *Psychometrika, 48*, 353–369.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement, 17*, 177–186.

Zeng, L. (1995). The optimal degree of smoothing in equipercentile equating with postsmoothing. *Applied Psychological Measurement.*

Zeng, L., & Cope, R.T. (1995). Standard errors of linear equating for the counterbalanced design. *Journal of Educational and Behavioral Statistics.*

Zeng, L., & Kolen, M.J. (1994). IRT scale transformations using numerical integration. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Zeng, L., Hanson, B.A. & Kolen, M.J. (1994). Standard errors of a chain of linear equatings. *Applied Psychological Measurement.*

Zwick, R. (1991). Effects of item order and context on estimation of NAEP Reading Proficiency. *Educational Measurement: Issues and Practice, 10*, 10–16.

Appendix A: Answers to Exercises

Chapter 1

1.1.a. Because the top 1% of the examinees on a particular test date will be the same regardless of whether or not an equating process is used, equating likely would not affect who was awarded a scholarship.

1.1.b. In order to identify the top 1% of the examinees during the whole year, it is necessary to consider examinees who were administered two forms as one group. If the forms on the two test dates were unequally difficult, then the use of equating could result in scholarships being awarded to different examinees as compared to just using the raw score on the form each examinee happened to be administered.

1.2. Because Form X_3 is easier than Form X_2, a raw score of 29 on Form X_3 indicates the same level of achievement as a raw score of 28 on Form X_2. From the table, a Form X_2 raw score of 28 corresponds to a scale score of 13. Thus, a raw score of 29 on Form X_3 also corresponds to a scale score of 13.

1.3. Because the test is to be secure, items that are going to be used as scored items in subsequent administrations cannot be released to examinees. Of the designs listed, the common-item nonequivalent groups design with external common items can be most easily implemented. On a particular administration, each examinee would receive a test form containing the scored items, a set of unscored items that had been administered along with a previous form, and possibly another set of unscored items to be used as a common-item section in subsequent equatings. Thus, all items that contribute to an examinee's score would be new items that would never be reused. The single group design with counterbalancing (assuming no differential order effects) and random groups design also could be implemented using examinees from other states. For example, using the random groups design, forms could be spiraled in another state which did not require that the test be released. The equated forms could be used subsequently in the state that required disclosure. The common-item nonequivalent groups design with internal common items may also be used in this way.

1.4. Random groups design. This design requires that only one form be administered to each examinee.

1.5. Only the common-item nonequivalent groups design can be used. Both the random groups and single group designs require the administration of more than one form on a given test date.

1.6. a. Group 2. b. Group 1. c. The content of the common items should be representative of the total test; otherwise, inaccurate equating might result.

1.7. Statement I is consistent with an observed score definition. Statement II is consistent with an equity definition.

1.8. Random. Systematic.

Chapter 2

2.1. $P(2.7) = 100\{.7 + [2.7 - (3 - .5)][.9 - .7]\} = 74$;
$P(.2) = 100\{0 + [.2 - (0 - .5)][.2 - 0]\} = 14$;
$P^{-1}(25) = (.25 - .2)/(.5 - .2) + (1 - .5) = .67$;
$P^{-1}(97) = (.97 - .90)/(1 - .90) + (4 - .5) = 4.2$.

2.2. $\mu(X) = 1.70$; $\sigma(X) = 1.2689$; $\mu(Y) = 2.30$; $\sigma(Y) = 1.2689$; $m(x) = x + .60$; $l(x) = x + .60$.

2.3. $\mu[e_Y(x)] = .2(.50) + .3(1.75) + .2(2.8333) + .2(3.50) + .1(4.25) = 2.3167$;
$\sigma[e_Y(x)]$
$= \sqrt{[.2(.50^2) + .3(1.75^2) + .2(2.8333^2) + .2(3.50^2) + .1(4.25^2)] - 2.3167^2}$
$= 1.2098$.

2.4. Note: $\mu(X) = 6.7500$; $\sigma(X) = 1.8131$; $\mu(Y) = 5.0500$; $\sigma(Y) = 1.7284$. See Tables A.1 and A.2.

2.5. The mean and linear methods will produce the same results. This can be seen by applying the formulas. Note that the equipercentile method will not produce the same results as the mean and linear methods under these conditions unless the higher order moments (skewness, kurtosis, etc.) are identical for the two forms.

2.6. $21.4793 + [(23.15 - 23)/(24 - 23)][22.2695 - 21.4793] = 21.5978$.

2.7. $1.1(.8x + 1.2) + 10 = .88x + 1.32 + 10 = .88x + 11.32$.

2.8. In general, the shapes will be the same under mean and linear equating. Under equipercentile equating, the shape will be the same only if the shape of the Form X and Form Y distributions are the same. Actually, the shape of the Form X scores converted to the Form Y scale will be approximately the same as the shape of the Form Y distribution.

Table A.1. Score Distributions for Exercise 2.4.

x	$f(x)$	$F(x)$	$P(x)$	y	$g(y)$	$G(y)$	$Q(y)$
0	.00	.00	.0	0	.00	.00	.0
1	.01	.01	.5	1	.02	.02	1.0
2	.02	.03	2.0	2	.05	.07	4.5
3	.03	.06	4.5	3	.10	.17	12.0
4	.04	.10	8.0	4	.20	.37	27.0
5	.10	.20	15.0	5	.25	.62	49.5
6	.20	.40	30.0	6	.20	.82	72.0
7	.25	.65	52.5	7	.10	.92	87.0
8	.20	.85	75.0	8	.05	.97	94.5
9	.10	.95	90.0	9	.02	.99	98.0
10	.05	1.00	97.5	10	.01	1.00	99.5

Table A.2. Equated Scores for Exercise 2.4.

x	$m_Y(x)$	$l_Y(x)$	$e_Y(x)$
0	−1.7000	−1.3846	.0000
1	−.7000	−.4314	.7500
2	.3000	.5219	1.5000
3	1.3000	1.4752	2.0000
4	2.3000	2.4285	2.6000
5	3.3000	3.3818	3.3000
6	4.3000	4.3350	4.1500
7	5.3000	5.2883	5.1200
8	6.3000	6.2416	6.1500
9	7.3000	7.1949	7.3000
10	8.3000	8.1482	8.7500

Chapter 3

3.1. Note: $e_Y(x_i) = 28.3$; $t_Y(x_i) = 29.1$; $\hat{e}_Y(x_i) = 31.1$; $\hat{t}_Y(x_i) = 31.3$.

a. $29.1 - 28.3 = .8$. b. $31.1 - 28.3 = 2.8$. c. $31.3 - 28.3 = 3.0$. d. We cannot tell from the information given—we would need to have an indication of the variability of sample values over many replications, rather than the one replication that is given. e. Unsmoothed at $x_i = 26$. f. We cannot tell from the information given—we would need to have an indication of the variability of sample values over many replications, rather than the one replication that is given.

3.2. Mean, standard deviation, and skewness.

3.3. For Form Y, $C = 7$ is the highest value of C with a nominally significant χ^2. So, of the models evaluated, those with $C \leq 7$ would be eliminated. The model with the smallest value of C that is not eliminated using a nominal significance level of .30 is $C = 8$. For Form X, $C \leq 5$ are eliminated. $C = 6$ is the smallest value of C that is not eliminated.

3.4. Using equation (3.11), $\hat{d}_Y(28.6) = 28.0321 + 1.0557(.6) - .0075(.6)^2 + .0003(.6)^3$ [redacted]

3.5. Conversions for $S = .20$ and $S = .30$. Conversions for $S = .75$ and $S = 1.00$. It would matter which was chosen if Form X was used later as the old form for equating a new form, because in this process the unrounded conversion for Form X would be used.

3.6. It appears that the relationships for all S-parameters examined would fall within the ± 2 standard error bands. The identity equating relationship would fall outside the bands from 4 to 20 (refer to the standard errors in Table 3.2 to help answer this question).

3.7. For $N = 100$ on the Science Reasoning test, the identity equating was better than any of the other equating methods. Even with $N = 250$ on the Science Reasoning test, the identity equating performed as well as or better than any of the equipercentile methods. One factor that could have led to the identity equating appearing to be relatively better with small samples for the Science Reasoning test than for the English test would be if the two Science Reasoning forms were more similar to one another than were the two English forms. In the extreme case, suppose that two Science Reasoning forms were actually identical. In this case, the identity equating always would be better than any of the other equating methods.

Chapter 4

4.1. Denote $\mu_1 \equiv \mu_1(X)$, $\sigma_1 \equiv \sigma_1(X)$, etc. We want to show that $\sigma_s^2 = w_1\sigma_1^2 + w_2\sigma_2^2 + w_1w_2(\mu_1 - \mu_2)^2$. By definition, $\sigma_s^2 = w_1\mathbf{E}_1(X - \mu_s)^2 + w_2\mathbf{E}_2(X - \mu_s)^2$. Noting that $\mu_s = w_1\mu_1 + w_2\mu_2$ and $w_1 + w_2 = 1$,

$$\begin{aligned} w_1\mathbf{E}_1(X - \mu_s)^2 &= w_1\mathbf{E}_1(X - w_1\mu_1 - w_2\mu_2)^2 \\ &= w_1\mathbf{E}_1[(X - \mu_1) + w_2(\mu_1 - \mu_2)]^2 \\ &= w_1\mathbf{E}_1(X - \mu_1)^2 + w_1w_2^2(\mu_1 - \mu_2)^2 \\ &= w_1\sigma_1^2 + w_1w_2^2(\mu_1 - \mu_2)^2. \end{aligned}$$

By similar reasoning,

$$w_2\mathbf{E}_2(X - \mu_s)^2 = w_2\sigma_2^2 + w_1^2w_2(\mu_1 - \mu_2)^2.$$

Thus,

$$\begin{aligned} \sigma_s^2 &= w_1\mathbf{E}_1(X - \mu_s)^2 + w_2\mathbf{E}_2(X - \mu_s)^2 \\ &= w_1\sigma_1^2 + w_1w_2^2(\mu_1 - \mu_2)^2 + w_2\sigma_2^2 + w_1^2w_2(\mu_1 - \mu_2)^2 \\ &= w_1\sigma_1^2 + w_2\sigma_2^2 + (w_1 + w_2)w_1w_2(\mu_1 - \mu_2)^2 \\ &= w_1\sigma_1^2 + w_2\sigma_2^2 + w_1w_2(\mu_1 - \mu_2)^2 \end{aligned}$$

4.2. To prove that Angoff's $\mu_s(X)$ gives results identical to equation (4.17), note that $\mu_s(V) = w_1\mu_1(V) + w_2\mu_2(V)$, and recall that $w_1 + w_2 = 1$. Therefore, Angoff's $\mu_s(X)$ is

$$\begin{aligned}\mu_s(X) &= \mu_1(X) + \alpha_1(X|V)[w_1\mu_1(V) + w_2\mu_2(V) - \mu_1(V)]\\ &= \mu_1(X) + \alpha_1(X|V)[-w_2\mu_1(V) + w_2\mu_2(V)]\\ &= \mu_1(X) - w_2\alpha_1(X|V)[\mu_1(V) - \mu_2(V)],\end{aligned}$$

which is equation (4.17) since $\gamma_1 = \alpha_1(X|V)$.

To prove that Angoff's $\sigma_s^2(X)$ gives results identical to equation (4.19), note that

$$\sigma_s^2(V) = w_1\sigma_1^2(V) + w_2\sigma_2^2(V) + w_1w_2[\mu_1(V) - \mu_2(V)]^2.$$

(This result is analogous to the result proved in Exercise 4.1.) Therefore, Angoff's $\sigma_s^2(X)$ is

$$\begin{aligned}\sigma_s^2(X) &= \sigma_1^2(X) + \alpha_1^2(X|V)\{w_1\sigma_1^2(V) + w_2\sigma_2^2(V)\\ &\quad + w_1w_2[\mu_1(V) - \mu_2(V)]^2 - \sigma_1^2(V)]\}\\ &= \sigma_1^2(X) + \alpha_1^2(X|V)[-w_2\sigma_1^2(V) + w_2\sigma_2^2(V)]\\ &\quad + w_1w_2\alpha_1^2(X|V)[\mu_1(V) - \mu_2(V)]^2\\ &= \sigma_1^2(X) - w_2\alpha_1^2(X|V)[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\alpha_1^2(X|V)[\mu_1(V) - \mu_2(V)]^2,\end{aligned}$$

which is equation (4.19) since $\gamma_1 = \alpha_1(X|V)$. Similar proofs can be provided for $\mu_s(Y)$ and $\sigma_s^2(Y)$.

4.4. The Tucker results are the same as those provided in the third row of Table 4.4. For the Levine method, using equations (4.58) and (4.59), respectively,

$$\gamma_1 = \frac{6.5278^2 + 13.4088}{2.3760^2 + 13.4088} = 2.9401$$

$$\gamma_2 = \frac{6.8784^2 + 14.7603}{2.4515^2 + 14.7603} = 2.9886.$$

Note that

$$\mu_1(V) - \mu_2(V) = 5.1063 - 5.8626 = -.7563 \qquad \text{and}$$

$$\sigma_1^2(V) - \sigma_2^2(V) = 2.3760^2 - 2.4515^2 = -.3645.$$

Therefore, equations (4.17)–(4.20) give

$$\mu_s(X) = 15.8205 - .5(2.9401)(-.7563) = 16.9323$$

$$\mu_s(Y) = 18.6728 + .5(2.9886)(-.7563) = 17.5427$$

$$\sigma_s^2(X) = 6.5278^2 - .5(2.9401^2)(-.3645) + .25(2.9401^2)(-.7563^2) = 45.4237$$

$$\sigma_s^2(Y) = 6.8784^2 + .5(2.9886^2)(-.3645) + .25(2.9886^2)(-.7563^2) = 46.9618.$$

Using equation (4.1),

$$l_{Ys}(x) = \sqrt{46.9618/45.4237}\,(x - 16.9323) + 17.5427 = .33 + 1.02x.$$

4.5. Using the formula in Table 4.1,

$$\rho_1(X,X') = \frac{\gamma_1^2[\sigma_1(X,V) - \sigma_1^2(V)]}{(\gamma_1 - 1)\sigma_1^2(X)},$$

where $\gamma_1 = \sigma_1^2(X)/\sigma_1(X,V)$. For the illustrative example,

$$\gamma_1 = 6.5278^2/13.4088 = 3.1779 \quad \text{and}$$

$$\rho_1(X,X') = \frac{3.1779^2(13.4088 - 2.3760^2)}{(3.1779 - 1)6.5278^2} = .845.$$

Similarly,

$$\rho_2(Y,Y') = \frac{\gamma_2^2[\sigma_2(Y,V) - \sigma_2^2(V)]}{(\gamma_2 - 1)\sigma_2^2(Y)},$$

where $\gamma_2 = \sigma_2^2(Y)/\sigma_2(Y,V)$. For the illustrative example,

$$\gamma_2 = 6.8784^2/14.7603 = 3.2054$$

$$\rho_2(Y,Y') = \frac{3.2054^2(14.7603 - 2.4515^2)}{(3.2054 - 1)6.8784^2} = .862.$$

4.6.a. From equation (4.38), the most general equation for γ_1 is $\gamma_1 = \sigma_1(T_X)/\sigma_1(T_V)$. It follows that

$$\gamma_1 = \frac{(K_X/K_V)\sigma_1(T_V)}{\sigma_1(T_V)} = \frac{K_X}{K_V}.$$

Similarly, $\gamma_2 = K_Y/K_V$.

4.6.b. Under the classical model, the γs are ratios of actual test lengths; whereas under the classical congeneric model, the γs are ratios of effective test lengths.

4.7. All of it [see equation 4.82].

4.8. No, it is not good practice from the perspective of equating alternate forms. All other things being equal, using more highly discriminating items will cause the variance for the new form to be larger than the variance for previous forms. Consequently, form differences likely will be a large percent of the observed differences in variances, and equating becomes more suspect as forms become more different in their statistical characteristics. These and related issues are discussed in more depth in Chapter 8.

4.9. From equation (4.59),

$$\gamma_2 = \frac{\sigma_2^2(Y) + \sigma_2(Y,V)}{\sigma_2^2(V) + \sigma_2(Y,V)}.$$

Recall that, since γ_2 is for an external anchor, $\sigma_2(E_Y, E_V) = 0$. Replacing the quantities in equation (4.59) with the corresponding expressions in equation set (4.70) gives

$$\begin{aligned}\gamma_2 &= \frac{[\lambda_Y^2\sigma_2^2(T) + \lambda_Y\sigma_2^2(E)] + \lambda_Y\lambda_V\sigma_2^2(T)}{[\lambda_V^2\sigma_2^2(T) + \lambda_V\sigma_2^2(E)] + \lambda_Y\lambda_V\sigma_2^2(T)} \\ &= \frac{\lambda_Y[(\lambda_Y + \lambda_V)\sigma_2^2(T) + \sigma_2^2(E)]}{\lambda_V[(\lambda_V + \lambda_Y)\sigma_2^2(T) + \sigma_2^2(E)]} \\ &= \lambda_Y/\lambda_V.\end{aligned}$$

4.10.a. Since $X = A + V$,

$$\sigma_1(X, V) = \sigma_1(A + V, V) = \sigma_1^2(V) + \sigma_1(A, V).$$

The assumption that $\rho_1(X, V) > 0$ implies that $\sigma_1(X, V) > 0$. Since $\sigma_1^2(V) \geq 0$ by definition, the above equation leads to the conclusion that $\sigma_1(A, V) > 0$ and, therefore, $\sigma_1^2(V) < \sigma_1(X, V)$. Also,

$$\begin{aligned}\sigma_1^2(X) &= \sigma_1(A + V, A + V) = \sigma_1^2(A) + \sigma_1^2(V) + 2\sigma_1(A, V)\\ &= [\sigma_1^2(V) + \sigma_1(A, V)] + [\sigma_1^2(A) + \sigma_1(A, V)]\\ &= \sigma_1(X, V) + [\sigma_1^2(A) + \sigma_1(A, V)].\end{aligned}$$

Since $\sigma_1^2(A) \geq 0$ by definition and it has been shown that $\sigma_1(A, V) > 0$, it necessarily follows that $\sigma_1(X, V) < \sigma_1^2(X)$. Consequently, $\sigma_1^2(V) < \sigma_1(X, V) < \sigma_1^2(X)$.

4.10.b. $\gamma_{1T} = \sigma_1(X, V)/\sigma_1^2(V)$, which must be greater than 1 because $\sigma_1(X, V) > \sigma_1^2(V)$. Now, $\gamma_{1L} = \sigma_1^2(X)/\sigma_1(X, V)$. To show that $\gamma_{1T} < \gamma_{1L}$, it must be shown that

$$\sigma_1(X, V)/\sigma_1^2(V) < \sigma_1^2(X)/\sigma_1(X, V) \qquad \text{or}$$

$$\sigma_1^2(X, V) < \sigma_1^2(X)\sigma_1^2(V) \qquad \text{or} \qquad \left[\frac{\sigma_1(X, V)}{\sigma_1(X)\sigma_1(V)}\right]^2 < 1,$$

which must be true because the term in brackets is $\rho_1(X, V)$, which is less than 1 by assumption.

4.10.c. Suppose that V and X measure the same construct and both satisfy the classical test theory model. If V is longer than X, then $\sigma^2(V) > \sigma^2(X)$. This,

Table A.3. Conditional Distributions of Form X Given Common-Item Scores for Population 1 in Exercise 5.1.

	v			
x	0	1	2	3
0	.20	.10	.10	.00
1	.20	.20	.10	.05
2	.30	.30	.25	.10
3	.15	.30	.25	.25
4	.10	.075	.20	.30
5	.05	.025	.10	.30
$h_1(v)$	.20	.40	.20	.20

Table A.4. Calculation of Distribution of Form X and Common-Item Scores for Population 1 Using Frequency Estimation Assumptions in Exercise 5.2.

	v					
x	0	1	2	3	$f_2(x)$	$F_2(x)$
0	.20(.20) = .04	.10(.20) = .02	.10(.40) = .04	.00(.20) = .00	.10	.10
1	.20(.20) = .04	.20(.20) = .04	.10(.40) = .04	.05(.20) = .01	.13	.23
2	.30(.20) = .06	.30(.20) = .06	.25(.40) = .10	.10(.20) = .02	.24	.47
3	.15(.20) = .03	.30(.20) = .06	.25(.40) = .10	.25(.20) = .05	.24	.71
4	.10(.20) = .02	.075(.20) = .015	.20(.40) = .08	.30(.20) = .06	.175	.885
5	.05(.20) = .01	.025(.20) = .005	.10(.40) = .04	.30(.20) = .06	.115	1.00
$h_2(v)$	.20	.20	.40	.20		

Table A.5. Cumulative Distributions and Finding Equipercentile Equivalents for $w_1 = .5$ in Exercise 5.3.

x	$F_s(x)$	$P_s(x)$	y	$G_s(y)$	$Q_s(y)$	x	$e_{Ys}(x)$
0	.1000	5.00	0	.0925	4.62	0	.04
1	.2400	17.00	1	.3000	19.62	1	.87
2	.4850	36.25	2	.5150	40.75	2	1.79
3	.7300	60.75	3	.7525	63.38	3	2.89
4	.8925	81.12	4	.9000	82.62	4	3.90
5	1.0000	94.62	5	1.0000	95.00	5	4.96

.2308. Because the residuals tend to be negative in the middle and positive at the ends, the regression of X on V for Population 1 appears to be nonlinear. Similarly, for Population 2, the mean residuals for the regression of Y on V are .2385, −.1231, −.2346, .3538, also suggesting nonlinear regression. This nonlinearity of regression would likely cause the Tucker and Braun-Holland methods to differ [illegible]

Percentile Rank of $v = .375$ in Population $2 = 17.5$; $Q_2^{-1}(17.5) = .975$. Thus, $x = 1$ is equivalent to $y = .975$ using chained equipercentile. For $x = 3$; $P_1(x = 3) = 62.50$; 62.5th percentile for V in Population $1 = 1.625$; Percentile Rank of $v = 1.625$ in Population $2 = 45$; $Q_2^{-1}(45) = 2.273$. Thus, $x = 3$ is equivalent to $y = 2.273$ using the chained equipercentile method.

Chapter 6

6.1. For the first item, using equation (6.1),

$$p_{ij} = .10 + (1 - .10)\frac{\exp[1.7(1.30)(.5 - -1.30)]}{1 + \exp[1.7(1.30)(.5 - -1.30)]} = .9835.$$

For the two other items, $p_{ij} = .7082$, and .3763.

6.2. For $\theta_I = .5$, $f(x = 0) = .0030$; $f(x = 1) = .1881$; $f(x = 2) = .5468$; $f(x = 3) = .2621$.

6.3.a. From equation (6.4), $b_{Jj} = Ab_{Ij} + B$ and $b_{Jj^*} = Ab_{Ij^*} + B$. Subtract the second equation from the first to get $b_{Jj} - b_{Jj^*} = A(b_{Jj} - b_{Jj^*})$, which implies that $A = (b_{Jj} - b_{Jj^*})/(b_{Ij} - b_{Ij^*})$.

6.3.b. From equation (6.3), $a_{Jj} = a_{Ij}/A$. Solving for A, $A = a_{Ij}/a_{Jj}$.

Table A.6. IRT Observed Score Equating Answer to Exercise 6.5.

Probability of Correct Answers and True Scores

θ_i	Item $j=1$	$j=2$	$j=3$	$j=4$	$j=5$	τ
Form X						
−1.0000	.7370	.6000	.2836	.2531	.2133	2.0871
.0000	.8799	.9079	.4032	.2825	.2678	2.7414
1.0000	.9521	.9867	.6881	.4965	.4690	3.5925
Form Y						
−1.0000	.7156	.6757	.2791	.2686	.2074	2.1464
.0000	.8851	.8773	.6000	.3288	.2456	2.9368
1.0000	.9611	.9642	.9209	.5137	.4255	3.7855

Form X Distribution

x	$f(x\|\theta=-1)$	$f(x\|\theta=0)$	$f(x\|\theta=1)$	$f(x)$	$F(x)$	$P(x)$
0	.0443	.0035	.0001	.0159	.0159	.7966
1	.2351	.0646	.0052	.1016	.1175	6.6734
2	.3925	.3383	.0989	.2766	.3941	25.5831
3	.2524	.3990	.3443	.3319	.7260	56.0064
4	.0690	.1704	.4009	.2134	.9394	83.2720
5	.0068	.0244	.1506	.0606	1.0000	96.9718

Form Y Distribution

y	$g(y\|\theta=-1)$	$g(y\|\theta=0)$	$g(y\|\theta=1)$	$g(y)$	$G(y)$	$Q(y)$
0	.0385	.0029	.0000	.0138	.0138	.6905
1	.2165	.0490	.0020	.0892	.1030	5.8393
2	.3953	.2594	.0425	.2324	.3354	21.9178
3	.2670	.4235	.3100	.3335	.6688	50.2114
4	.0752	.2276	.4589	.2539	.9228	79.5807
5	.0075	.0376	.1866	.0772	1.0000	96.1384

Form Y Equivalents of Form X scores

x	$e_Y(x)$
0	.0772
1	1.0936
2	2.1577
3	3.1738
4	4.1454
5	5.1079

Table A.7. Answer to Exercise 6.6.

r	x	$f_r(x)$ for $r \leq 4$		Probability	
4	0	$f_4(0) = f_3(0)(1-p_4)$		$= .4430(1-.4)$	$= .2658$
	1	$f_4(1) = f_3(1)(1-p_4)$	$+ f_3(0)p_4$	$= .4167(1-.4) + .4430(.4)$	$= .4272$
	2	$f_4(2) = f_3(2)(1-p_4)$	$+ f_3(1)p_4$	$= .1277(1-.4) + .4167(.4)$	$= .2433$
	3	$f_4(3) = f_3(3)(1-p_4)$	$+ f_3(2)p_4$	$= .0126(1-.4) + .1277(.4)$	$= .0586$
	4	$f_4(4) =$	$f_3(3)p_4$	$= .0126(.4)$	$= .0050$

Table A.8. Estimated Probability of Correct Response Given $\theta = 1$ for Exercise 6.7.

Item	Scale J	Mean/sigma	Mean/mean
1	.9040	.8526	.8522
2	.8366	.8076	.8055
3	.2390	.2233	.2222
sum	1.9796	1.8835	1.8799
Hdiff		.0037	.0039
SLdiff		.0092	.0099

the common items provide direct evidence about how the new group compares to the old group for two groups of examinees that actually can be observed. In IRT equating to a calibrated pool, the only group of examinees who takes all of the common items is the new group. Thus, when equating to a pool, there is no old group with which to compare the new group on the common items, unless we rely on the assumptions of the IRT model, which is a much weaker comparison than can be made when we have two groups who actually took the common items.

6.9. Step (a) is similar, except that, with IRT, a design might be selected that involves linking to an IRT calibrated item pool. Step (b) is the same, in that the same construction, administration, and scoring procedures could be used for either type of equating method. In Step (c), IRT equating involves estimating item parameters and scaling the item parameter estimates. These steps are not needed in the traditional methods. In both types of methods, the raw scores are converted to scale scores by using statistical methods. However, traditional methods differ from the IRT methods. Also, the IRT methods might involve equating using an item pool. Steps (d), (e), and (f) are the same for the two types of methods.

Chapter 7

7.1. Answers to 7.1.a, 7.1.b, and 7.1.c are given in Table A.9. Using equation (7.10) for Exercise 7.1.d, the standard error at $x = 3$ is 1.3467. The standard error at $x = 5$ is 1.4291.

Table A.9. Bootstrap Standard Errors for Exercise 7.1a, b, and c.

	Sample				
Statistic	1	2	3	4	$\widehat{se}_{boot}$
$\hat{\mu}(X)$	4.0000	2.7500	4.2500	3.2500	
$\hat{\mu}(Y)$	3.0000	4.6667	3.6667	2.0000	
$\hat{\sigma}(X)$	2.1213	2.0463	1.9203	2.2776	
$\hat{\sigma}(Y)$	1.4142	.4714	1.8856	1.4142	
$\hat{l}_Y(x=3)$	2.3333	4.7243	2.4392	1.8448	1.2856
$\hat{l}_Y(x=5)$	3.6667	5.1850	4.4031	3.0866	.9098
$sc[\hat{l}_Y(x=3)]$	10.9333	11.8897	10.9757	10.7379	.5142
$sc[\hat{l}_Y(x=5)]$	11.4667	12.0740	11.7613	11.2346	.3639
$sc_{int}[\hat{l}_Y(x=3)]$	11	12	11	11	.5000
$sc_{int}[\hat{l}_Y(x=5)]$	11	12	12	11	.5774

7.2. Using equation (7.12),

$$\hat{var}[\hat{e}_Y(x_i)] \cong \frac{1}{[.7418-.7100]^2}\left\{\frac{(72.68/100)(1-72.68/100)(4329+4152)}{4329(4152)} - \frac{(.7418-72.68/100)(72.68/100-.7100)}{4329(.7418-.7100)}\right\} = .9084.$$

Estimated standard error equals $\sqrt{.9084} = .3014$. Using equation (7.13),

$$\hat{var}[\hat{e}_Y(x_i)] \cong 8.9393^2\frac{(72.68/100)(1-72.68/100)}{.33^2}\left(\frac{1}{4329}+\frac{1}{4152}\right) = .0687.$$

Estimated standard error equals $\sqrt{.0687} = .2621$. The differences between the standard errors could be caused by the distributions not being normal. Also, equation (7.12) assumes discrete distributions, whereas equation (7.13) assumes continuous distributions. Differences also could result from error in estimating the standard errors.

7.3. a. 150 total (75 per form). b. 228 total (114 per form). c. If the relationship

was truly linear, it would be best to use linear, because linear has less random error.

7.4. Using equation (7.11), with a sample size of 100 per form, the error variance for linear equating equals .03, and the error variance for equipercentile equals .0456. The squared bias for linear is $(1.3-1.2)^2 = .01$. Thus, the mean squared error for linear is $.03 + .01 = .04$. Assuming no bias for equipercentile, the mean squared error for equipercentile = .0456. Therefore, linear leads to less error than equipercentile. With a sample size of 1000 per form ...

7.5. a. .2629 and .4382. b. .1351 and .2683. c. .3264 and .6993. d. 96 per form and 267 per form.

7.6. The identity equating does not require any estimation. Thus, the standard error for the identity equating is 0. If the population equating is similar to the identity equating, then the identity equating might be best. Otherwise, the identity equating can contain substantial systematic error (which is not reflected in the standard error). Thus, the identity equating is most attractive when the sample size is small or when there is reason to believe that the alternate forms are very similar.

Chapter 8

8.1.a. From equation (7.18), a sample size of more than $N_{tot} = (2/.1^2)(2 + .5^2) =$ 450 total (225 per form) would be needed.

8.1.b. From equation (7.18), a sample size of more than $N_{tot} = (2/.2^2)(2 + .5^2) =$ 112.5 total (approx. 67 per form) would be needed.

8.1.c. In a situation where a single passing score is used, the passing score is at a z-score of .5, and the equating relationship is linear in the population.

8.2.a. For Forms D and following: In even-numbered years, the spring form links to the previous spring form and the fall form links to the previous spring form. In odd-numbered years, the spring form links to the previous fall and the fall form links to the previous fall.

8.2.b. Form K links to Form I. Form L links to Form I. Form M links to Form L. Form N links to Form L.

8.3.a. For Forms D and following in Modified Plan 1 (changes from Link Plan 4 shown in bold italics): In even-numbered years, the spring form links to the previous spring form and the fall form links to the previous spring form. In odd-numbered years, ***the spring form links to the fall form from two years earlier*** and the fall form links to the previous fall.

For Forms D and following in Modified Plan 2: In even-numbered years, the spring form links to the previous spring form and the fall form links to the previous spring form. In odd-numbered years, ***the spring form links to the previous spring*** and the fall form links to the previous fall.

8.3.b. In Modified Plan 1, K links to I, L links to I, ***M links to J***, and N links to L. In Modified Plan 2, K links to I, L links to I, ***M links to K***, and N links to L.

8.3.c. For Modified Plan 1, Rule 1 is violated (this plan results in equating strains), and Rules 2 through 4 are met as well with this plan as with Single Link Plan 4. For Modified Plan 2, Rule 1 is achieved much better than for Modified Plan 1, Rule 2 is met better than for Single Link Plan 4 or for Modified Plan 1, and Rules 3 and 4 are met as well as for Modified Plan 1 or Single Link Plan 4. Modified Plan 2 seems to be the best of the two modified plans.

8.4. In Table 8.6, for the first 4 years the decrease in mean and increase in standard deviation were accompanied by an increase in the sample size. However, now in year 5 there is a decrease in the sample size. The Levine method results are most similar to the results when the sample size was near 1050 in year 2. For this reason, the Levine method might be considered to be preferable. However, the choice between methods is much more difficult in this situation, because a sample size decrease never happened previously. In practice, many additional issues would need to be considered.

8.5.a. Randomly assign examinees to the two modes. Convert parameter estimates for the computerized version to the base IRT scale using the random groups design. Probably two different classrooms would be needed, one for paper and pencil and one for computer.

8.5.b. Use the items that are in common between the two modes as common items in the common-item equating to an item pool design.

8.5.c. Random groups requires large sample sizes and a way to randomly assign examinees to different modes of testing. Common-item equating to an item pool requires that the common items behave the same on computerized and paper and pencil versions. This requirement likely would not be met. This design also requires that the groups taking the computerized and paper and pencil versions be reasonably similar in achievement level.

8.5.d. It is unlikely that all items will behave the same when administered by computer as when administered using paper and pencil. Therefore, the results from using this design would be suspect. At a minimum, a study should be conducted to discover the extent to which context effects affect the performance of the items.

8.5.e. The random groups design is preferable. Even with this design, it would be necessary to study whether or not the construct being measured by the test changes from a paper and pencil to a computerized mode. For example, there is evidence that reading tests with long reading passages can be affected greatly when they are adapted for computer administration. Note that with the random groups design, the effects of computerization could be studied for those items that had been previously administered in the paper and pencil mode.

8.6. Some causes due to changes in items include changes in item position, changes in surrounding items, changes in font, changes in wording, and rearranging alternatives. Some causes due to changes in examinees include changes in a field of study and changes in the composition of the examinee groups. For example, changes in country names, changes in laws, and new scientific discoveries might lead to changes in the functioning of an item. As another example, a vocabulary word like "exorcist" might become much more familiar after the release of a movie of the same name. Some causes due to changes in administration conditions include changes in time given to take the test, security breaches, changes in mode of administration, changes in test content, changes in test length, changes in motivation conditions, changes in calculator usage, and changes in directions given to examinees.

8.7. To consider equating, the forms must be built to the same content and statistical specifications. Assuming that they are, the single group design is eliminated because it would require that two forms be administered to each examinee, which would be difficult during an operational administration. The common-item nonequivalent groups design is eliminated because having many items associated with each reading passage would make it impossible to construct a content representative set of common items. The random groups design could be used. This design requires large samples, which would not be a problem in this example. Also, the random groups design is not affected by context, fatigue, and practice effects, and the only statistical assumption that it requires is that the process used to randomly assign forms was effective. Therefore, the random groups design is best in this situation. Equipercentile equating would be preferred because it generally provides more accuracy along the score scale (assuming that the relationship is not truly linear). Equipercentile equating also requires large sample sizes, which is not a problem in the situation described.

Appendix B: Computer Programs and Data Sets

We will make available some of the data sets used in this book to interested persons. We will also make available Macintosh computer programs, and associated documentation, that can be used to conduct many of the analyses in the book free of charge. To find out how to obtain the data sets program implements the cubic spline smoothing method described in Chapter 3.

2. **RG Equate** by B.A. Hanson. This program conducts equipercentile equating with log-linear smoothing, as described in Chapter 3.
3. **Usmooth** by B.A. Hanson. This program smoothes test score distributions using the log-linear smoothing method, as described in Chapter 3.
3. **CIPE** by M.J. Kolen. This program conducts observed score equating under the common-item nonequivalent groups design as described in Chapters 4 and 5. Tucker linear (external or internal common items), Levine linear observed score (internal common items only), and frequency estimation equipercentile equating with cubic spline smoothing are implemented.
4. **ST** by L. Zeng and B.A. Hanson. This program conducts IRT scale transformations using the mean/mean, mean/sigma, Stocking and Lord, and Haebara methods described in Chapter 6.
5. **PIE** by B.A. Hanson and L. Zeng. This program conducts IRT true and observed score equating using the methods described in Chapter 6.
6. **Equating Error** by B.A. Hanson. This program estimates bootstrap standard errors of equipercentile equating for the random groups design. Standard errors for both the cubic spline postsmoothing and log-linear presmoothing methods can be calculated.

Although these programs have been tested and we believe them to be free of errors, we do not warrant, guarantee, or make any representations regarding the use or the results of this software in terms of their appropriateness, correctness, accuracy, reliability, or otherwise. The entire responsibility for the use of this software rests with the user.

Index

Springer Series in Statistics

(continued from p. ii)

Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.
Reinsel: Elements of Multivariate Time Series Analysis.
Reiss: A Course on Point Processes.
Reiss: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.
Rieder: Robust Asymptotic Statistics.
Rosenbaum: Observational Studies.
Ross: Nonlinear Estimation.
Sachs: Applied Statistics: A Handbook of Techniques, 2nd edition.
Särndal/Swensson/Wretman: Model Assisted Survey Sampling.
Schervish: Theory of Statistics.
Seneta: Non-Negative Matrices and Markov Chains, 2nd edition.
Shao/Tu: The Jackknife and Bootstrap.
Siegmund: Sequential Analysis: Tests and Confidence Intervals.
Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 2nd edition.
Tong: The Multivariate Normal Distribution.
Vapnik: Estimation of Dependences Based on Empirical Data.
Weerahandi: Exact Statistical Methods for Data Analysis.
West/Harrison: Bayesian Forecasting and Dynamic Models.
Wolter: Introduction to Variance Estimation.
Yaglom: Correlation Theory of Stationary and Related Random Functions I: Basic Results.
Yaglom: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.