Literatuur

Adema, J.J., & van der Linden, W.J. (1989). Algorithms for computerized test construction of parallel tests using classical item parameters. *Journal of Educational Statistics, 15,* 129-145.

Aitchison, J., & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813-828.

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

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimation. *Journal of the Royal Statistical Society, Series B, 32,* 283-301.

Andersen, E.B. (1973a). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E.B. (1973b). *Conditional inference and models for measuring.* (Unpublished Ph.D. Thesis). Copenhagen: Mentalhygiejnisk Forlag.

Andersen, E.B. (1973c). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andersen, E.B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357-374.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andersen, E.B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Andersen, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 46,* 443-459.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In: R.L. Thorndike (red.). *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Armstrong, R.D., Jones, D.H., & Wu, I. (1992). An automated test development of parallel tests from a seed test. *Psychometrika, 57,* 271-288.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19,* 3-11.

Bartko, J.J., & Carpenter, W.T. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease, 163,* 307-317.

Bejar, I.I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software.

Berger, J.O. (1980). *Statistical decision theory: Foundations, concepts and methods.* New York: Springer.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21,* 147-152.

Bezembinder, Thom. G. G. (1970). *Van rangorde naar continuum*. Deventer: Van Loghum Slaterus.

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord, & M.R. Novick. *Statistical theories of mental test scores* (pp. 397-424). Reading: Addison-Wesley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge: The MIT Press.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. (1976). Basic issues in the measurement of change. In: D.N.M. de Gruijter, & L.J.Th. van der Kamp (red.). *Advances in psychological and educational measurement* (pp. 75-96). London: Wiley.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Psychological Measurement, 13,* 261-280.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15,* 129-145.

Bol, E., & Verhelst, N.D. (1985). Inhoudelijke en statistische analyse van een leertoets. *Tijdschrift voor Onderwijsresearch, 10,* 49-68.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bosch, L. van den, Gillijns, P., Krom, R., & Moelands, F. (1991). *Handleiding schaal vorderingen in spellingvaardigheid 1.* Arnhem: Cito.

Bradley, T.B. (1983). Remediation of cognitive deficits: A critical appraisal of the Feuerstein model. *Journal of Mental Deficiency Research, 27,* 79-92.

Braun, W.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In: P.W. Holland, & D.B. Rubin (red.). *Test equating* (pp. 9-49). New York: Academic Press.

Brennan, R.L. (1992). Elements of generalizability theory. Iowa City: ACT.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Bügel, K. (1991). Sexeverschillen in onderwijsprestaties in Nederland: Een overzicht van de literatuur en enkele nieuwe gegevens. *Pedagogische Studiën, 68,* 350-370.

Bügel, K. (1993). Tekstbegrip moderne vreemde talen: De invloed van sekse en tekstonderwerp op de scores van centrale examens. *Tijdschrift voor Onderwijs-wetenschappen, 23,* 162-176.

Bügel, K., & Glas, C.A.W. (1991). Item specifieke verschillen in prestaties tussen jongens en meisjes bij tekstbegrip examens moderne vreemde talen. *Tijdschrift voor Onderwijsresearch, 16,* 337-351.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the mulititrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Coombs, C.H. (1964). *A theory of data.* New York: Wiley.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18,* 183-204; *19,* 331-332.

Cicchetti, D.V. (1972). A new measure of agreement between rank ordered variables. *In Proceedings of the 80th Annual Convention of the American Psychological Association 7,* 17-18.

Cicchetti, D.V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129,* 452-456.

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psycholo-*
gical Measurement, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provisions for scales disagreement of partial credit. *Psychological Bulletin, 70,* 213-220.

Cornfield, J., & J.W. Tukey (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27,* 907-949.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L.J. (1971). Test validation. In: R.L. Thorndike (red.). *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L.J., & Furby, L. (1970). How we should measure "change" - or should we? *Psychological Bulletin, 74,* 68-80.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dirickx, Y.M.I., Baas, S.M., & Dorhout, B. (1987). *Operationele research*. Schoonhoven: Academic Service.

Divgi, D.R. (1981). *Two direct procedures for scaling and equating tests with item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Dixon, W.J. (red.) (1992). *BMDP statistical software manual: Vol. 1 and 2.* Berkeley: University of California Press.

Dousma, T., & Horsten, A. (1989). *Tentamineren.* Groningen: Wolters-Noordhoff.

Drenth, P.J.D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psycholo- gische test en zijn toepassingen.* Houten: Bohn Stafleu Van Loghum.

Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors.* New York: Oxford University Press.

Ebel, R.L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4,* 125-128.

Ebel, R.L. (1972). *Essentials of educational measurement.* Englewood Cliffs: Prentice-Hall.

Ebel, R.L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2,* 7-10.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement.* Englewood Cliffs: Prentice Hall.

Eggen, T.J.H.M. (1990). Innovative procedures in the calibration of measurement scales. In: W.H. Schreiber, & K. Ingenkamp (red.). *International developments in large scale*

assessment (pp.199-212). Windsor, Berkshire: NFER-NELSON.

Eggen, T.J.H.M., & Verhelst, N.D. (1992). *Item calibration in incomplete testing designs.* (Measurement and Research Department Reports 92-3). Arnhem: Cito.

Elliott, C.D., Murray, D.J., & Saunders, R. (1977). *Goodness of fit to the Rasch model as a criterion of test unidimensionality.* Manchester: University of Manchester.

Evers, A., Vliet-Mulder, J.C. van, & Laak, J. ter. (1992). *Documentatie van tests en testresearch in Nederland.* Amsterdam: Nederlands Instituut van Psychologen.

Fagot, R.F. (1991). Reliability of ratings for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement, 15,* 1-11.

Fagot, R.F. (1993). A generalized family of coefficients of relational agreement for numeri-
cal scales. *Psychometrika, 58,* 357-370.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357-370.

Feldt, L.S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education 6,* 37-49.

Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Feldt, L.S, & Brennan, R.L. (1989). Reliability. In: R.L. Linn (red.). *Educational Measure-*
ment (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

Ferguson, G.A., & Takane, Y. (1989). *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Feuerstein, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability.* Baltimore: University Park Press.

Fischer, G.H. (1972). *A step towards dynamic test-theory.* (Research Bulletin Nr. 10/72). Universität Wien: Psychologisches Institut.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-373.

Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests.* Bern: Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika, 46,* 59-77.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Fischer, G.H. (in voorbereiding). Derivations of the Rasch model. In: G.H. Fischer, & I.W. Molenaar (red.). *Rasch models: Their foundations, recent developments and applica-*

tions.

Fischer, G.H., & Scheiblechner, H. (1970). Algorithmen und programme für das probabi- listische testmodell von Rasch. *Psychologische Beiträge, 12,* 23-51.

Flanagan, J.C. (1951). Units, scores and norms. In: E.F. Lindquist (red.). *Educational measurement* (pp. 695-763). Washington, DC: American Council on Education.

Fleiss, J.L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 5, 323-327.

Fleiss, J.L., & Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika, 43,* 259-262.

Follman, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53,* 553-562.

Freeman, M.F., & Tukey, J.W. (1950). Transformations related to the angular and square root. *The Annals of Mathematical Statistics, 21,* 607-611.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measure- ment: Issues and practice, 7,* 53-63.

Glas, C.A.W. (1981). *Het Raschmodel bij data in een onvolledig design*. (PSM-Progress reports, 81-1). Utrecht: Vakgroep PSM van de subfaculteit Psychologie.

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* Arnhem: Cito.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In: M. Wilson (red.). *Objective measurement: Theory into practice: Vol. 1* (pp. 236-258). Norwood: Ablex.

Glas, C.A.W., & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Glas, C.A.W., & Verhelst, N.D. (in voorbereiding). Testing the Rasch model. In: G.H.Fischer, & I.W.Molenaar (red.). *Rasch models: Their foundations, recent developments and applications.*

Green, S.B., & Lissitz, R.W. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37,* 827-838.

Groot, A.D. de (1966). *Vijven en zessen.* Groningen: Wolters.

Groot, A.D. de, & Naerssen, R.F. (1973). *Studietoetsen, construeren, afnemen, analyseren: Deel I en II.* Den Haag: Mouton.

Gruijter, D.N.M. de (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22,* 263-269.

Guilford, J.P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education.* Tokyo: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gustafsson, J.E. (1979). *PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items.* (Reports from the Institute of Education, nr. 63). Göteborg: University of Göteborg.

Guttman, L. A. (1950). The Basis of Scalogram Analysis. In: S.A. Stouffer, L.A. Gutmann, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (red.). *Measurement and prediction: Studies in social psychology in World War II: Vol. 4.* Princeton: Princeton University Press.

Guttman, L. A. (1954). A new approach to factor analysis: The radex. In: P.F. Lazersfeld (red.). *Mathematical thinking in the social sciences* (pp. 258-348). New York: Colombia University Press.

Haggard, E.A. (1958). *Intraclass correlation and the analysis of variance.* New York: The Dryden Press.

Hambleton, R.K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159-170.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Compa- rison of IRT area and Mantel-Haenszel methods. *Applied Psychological Measurement, 2,* 313-334.

Harris, D.H., & Crouse, J.D. (1992). *A study of criteria used in equating.* Paper presented at the annual meeting of the National Council on Measurement in Education.

Heinen, T. (1993). *Discrete latent variable models.* Proefschrift, Katholieke Universiteit Brabant.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28,* 211-218.

Hofstee, W.K.B. (1977). Cesuurprobleem opgelost. *Onderzoek van Onderwijs, 6/2,* 6-7.

Hofstee, W.K.B. (1981). *Psychologische uitspraken over personen.* Deventer: Van Loghum Slaterus.

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In Anderson, S.B., & Helmick, J.S. (red.). *On educational testing.* San Francisco: Jossey-Bass.

Hoijtink, H., & Boomsma, A. (1991). *Statistical inference with latent ability estimates.* (Prepublication Department of Statistics and Measurement Theory). Groningen: University of Groningen.

Hoijtink, H. (red.). (1993). *Kwantitatieve Methoden nr. 42.*

Holland, P.W., & Rubin, D.B. (1982). *Test equating.* New York: Academic Press.

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In: H. Wainer, & H.I. Braun (red.). *Test validity* (pp.129-145). Hillsdale: Lawrence Erlbaum.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal, 25,* 423-430.

Houston, W.M., Raymond, M.R., & Svec, J.C. (1991). Adjustments for rater effects in performance assessment. *Applied Psychological Measurement, 15,* 409-421.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Applications to psychological measurement.* Homewood: Dow-Jones Irwin.

Iker, H.P., & Perry, N.C.A. (1960). A further note concerning the reliability of the point-biserial correlation. *Educational and Psychological Measurement, 20,* 505-507.

Imbos, Tj. (1989). *Het gebruik van einddoel toetsen bij aanvang van de studie.* Proefschrift, Rijksuniversiteit Limburg.

Inspectierapport. (1992). *Examens op punten getoetst: Onderzoek naar de ontwikkeling van de normen bij de centrale examens in het voortgezet onderwijs.*

James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater relia- bility with and without response bias. *Journal of Applied Psychology, 69,* 85-98.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357-373.

Jansen, G.G.H. (1979). *Het meten van veranderingen in de klassieke testtheorie.* (Bulletinreeks nr. 2). Arnhem: Cito.

Jarjoura, D. (1983). Best linear prediction of composite universe scores. *Psychometrika, 48,* 525-539.

Jazwinsky, A.H. (1970). *Stochastic processes and filtering theory.* New York: Academic Press.

Johnson, H.M. (1935). Some neglected principles in aptitude testing. *American Journal of Psychology, 47* 159-165.

Jonge, H. de (1963). *Inleiding tot de medische statistiek: Deel I.* Groningen: Wolters-Noordhoff.

Jöreskog, K.G. (1970). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology, 23,* 121-145.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7, user's reference guide.* Mooresville: Scientific Software.

Kamphuis, F.H., & Engelen, R.J.H. (in voorbereiding). Estimation and testing of structured latent ability covariance matrices in IRT models.

Kane, M.T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

Kelderman, H. (1988). *Loglinear multidimensional IRT model for polytomously scored items.* (Research Report 88-17). Enschede: Universiteit Twente.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Kelderman, H., & Steen, R. (1988). *LOGIMO I: Loglinear item response theory modeling.* (Computer Program). Enschede: University of Twente, Department of Educational Technology.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-327.

Kelley, T.L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Kendall, M., & Stuart, A. (1973). *The advanced theory of statistics: Vol. 2.* Londen: Griffin.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213-228.

Kolen, M.J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25,* 97-110.

Koppen, M.G.M. (1987). On finding the bidimension of a relation. *Journal of Mathematical Psychology, 3*1, 155-178.

Knol, D.L. (1986). *Een overzicht van meerdimensionale itemresponsmodellen*. (Rapport R-86-5). Enschede: Univeriteit Twente, Faculteit TO, vakgroep OMD.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30,* 61-70.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills: Sage Publications.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lahey, M.A., Downey, R.G., & Saal, F.E. (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin, 93,* 586-595.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Laros, J.A., & Tellegen, P.J. (1991). *Construction and validation of the SON-R 5½-17, the Snijders-Oomen non-verbal intelligence test.* Groningen: Wolters-Noordhoff.

Lazarsfeld, P.F. (1950). Logical and mathematical foundations of latent structure analysis. In: S.A. Stouffer. *Studies in social psychology in World War II, IV*. Princeton, NJ: Princeton University Press.

LBR (1988). *Psychologische tests en allochtonen.* Symposiumverslag 1987, LBR-Reeks nr. 6.

LBR (1990). *Toepasbaarheid van psychologische tests bij allochtonen.* Rapport van de testscreeningscommissie ingesteld door het LBR in overleg met het NIP, LBR-Reeks nr. 11.

Leeuw, J. de, & Verhelst, N.D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11*, 183-196.

Leeuwe, J.F.J. van (1990). *Probabilistic conjunctive models.* Proefschrift. Nijmegen: NICI.

Linden, W.J. van der (red.). (1982). Aspects of criterion-referenced measurement. *Evalua- tion in Education: An International Review Series, 5.*

Linden, W.J. van der (1983). *Van standaardtest naar itembank*. Universiteit Twente (oratie).

Linden, W.J. van der (1984). Some thoughts on the use of decision theory to set cutoff scores: Comment on De Gruijter and Hambleton. *Applied Psychological Measurement, 8,* 9-17.

Linden, W.J. van der (1985). Decision theory in educational research and testing. In: T. Husén, & T.N. Postlethwaite (red.). *International encyclopedia of education: Research and studies.* Oxford: Pergamon Press.

Linden, W.J. van der, & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement, 12,* 201-209.

Linden, W.J. van der, & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237-247.

Lindsay, B., Clifford, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96-107.

Linn, R.L. (red.). (1989). *Intelligence: Measurement, theory, and public policy*. Chicago: University of Illinois Press.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and performance tests.* Princeton, NJ: Educational Testing Service.

Lord, F.M. (1950). *Notes on comparable scales for test scores* (Research Bulletin 50-48). Princeton, NJ: Educational Testing Service.

Lord, F.M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*, 181-194.

Lord, F.M. (1953). On the statistical treatment of football numbers. *The American Psycholo- gist, 8,* 750-751.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Erlbaum.

Lord, F.M. (1983a). Unbiased estimators of ability parameters, their variance and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Lord, F.M. (1983b). *Estimating the imputed social cost of errors of measurement.* (Report RR-83-33-ONR). Princeton, NJ: Educational Testing Service.

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

Lord, F.M. & Wingerskey, M.S. (1983). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8, 453-461.*

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

Maris, E. (1992). *Psychometric models for psychological processes and structures.* Proefschrift, Universiteit Leuven.

Martin-Löf, P. (1973). *Statistika Modeller: Anteckningar från seminarier Lasåret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973.* Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure if the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*, 3-18.

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

Masters, G.N., & Wright, B.D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544.

Maxwell, A.E., & Pilliner, A.E.G. (1968). Deriving coefficients of reliability and agreement. *The British Journal of Mathematical and Statistical Psychology, 21,* 105-116.

McKinley, R.L., & Reckase, M.D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15,* 389-390.

Meerling (1981). *Methoden en technieken van psychologisch onderzoek: Deel 1.* Meppel: Boom.

Mellenbergh, G.J. (1977). The replicability of measures. *Psychological Bulletin, 84,* 378-384.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Mellenbergh, G.J. (1983). Conditional item bias methods. In: S.H. Irvine, & W.J. Berry (red.). *Human assessment and cultural factors* (pp. 293-302). New York: Plenum Press.

Mellenbergh, G.J. (1985). Vraag-onzuiverheid: definitie, detectie en onderzoek. *Nederlands Tijdschrift voor Psychologie, 40,* 425-435.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In: H. Wainer, & H.I. Braun (red.). *Test validity* (pp.33-45). Hillsdale: Lawrence Erlbaum.

Messick, S. (1989). Validity. In: R.L. Linn (red.). *Educational Measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In: R.L. Linn (red.). *Educational Measurement* (3rd ed., pp. 335-366). Washington, DC: American Council on Education.

Mills, C.N., & Melican, G.J. (1987). *A preliminary investigation of three compromise methods for establishing cut-off scores.* (Report RR-87-14). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R.J., & Bock, R.D. (1986). *PC-BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items.* Mooresville: Scientific Software.

Mislevy, R.J., & Wu, P.K. (1988). *Inferring examinee ability when some item responses are missing.* (Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Sheenan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54,* 661-680.

Moelands, A.H.J. (1988). *Entreetoets: Basisvaardigheden taal, rekenen en informatieverwerking (Verantwoording)*. Arnhem: Cito.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* Den Haag: Mouton.

Molenaar, I.W. (1981). *Programmabeschrijving van PML (versie 3.1) voor het Raschmodel.* (Heymans Bulletins Psychologische Instituten R.U.Groningen, nr. HB-81-538-RP). Groningen: Rijksuniversiteit Groningen.

Molenaar, I.W. (1983). *Item steps.* (Heymans Bulletins Psychologische Instituten R.U Groningen, nr. HB-83-630-EX). Groningen: Rijksuniversiteit Groningen.

Molenaar I.W., & Hoijtink, H (1990). The many null-distributions of person fit indices. *Psychometrika, 55,* 75-106.

Muskens, G.J. (1980). *Frames of meaning - are they measurable?* Proefschrift, Katholieke Universiteit Nijmegen.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, B. (1989). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.* Mooresville: Scientific Software.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Nederlands Instituut van Psychologen. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen.* Amsterdam: Nederlands Instituut van Psychologen.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

Oud, J.H.L., & Mommers (1988). Longitudinale computerondersteunende ondersteuning van lees- en spellingsmoeilijkheden: Een toepassing van het Kalmanfilter in de onderwijs- praktijk. *Tijdschrift voor Onderwijsresearch, 13,* 31-50.

Pennings, A.H. (1988). The development of strategies in embedded figure tasks. *International Journal of Psychology, 23,* 65-78.

Pennings, A.H. (1991). *Individual differences in the development of the restructuring ability in children.* Proefschrift, Rijksuniversiteit Utrecht.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (red.). *Eductional Measurement* (3rd ed., pp. 221-262). Washington, DC: American Council on Education.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data.* Proefschrift, Rijksuniversi- teit Groningen.

Popping, R. (1989). *AGREE: Computing agreement on nominal data, version 5.* (User's manual) Groningen: IEC ProGamma.

Popping, R. (1992). *Taxonomy on nominal scale agreement 1945 - 1990.* Groningen: IEC ProGamma.

Rao, C.R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50-57.

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

Rasch, G. (1961). On the general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* 321-333. Berkeley: University of California Press.

Rasch, G. (1977). *On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements.* Berkeley: University of California Press.

Read, T.R.C., & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer.

Reckase, M.D., & Mckinley, R.L. (1985). Some latent trait theory in a multidimensional latent space. In: D.I. Weiss (red.). *Proceedings of the 1982 computerized adaptive testing conference* (pp. 151-177). Minneapolis: University of Minnesota.

Rigdon S.E., & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567-574.

Rigdon S.E., & Tsutakawa, R.K. (1986). Estimation for the Rasch model when both ability and difficulty parameters are random. *Journal of Educational Statistics, 12,* 76-86.

Roskam, E.E. (1982). Hypotheses non fingo, een methodologische gevalstudie over onderzoek van intelligentietests. *Nederlands Tijdschrift voor de Psychologie, 37,* 331-359.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D.B. (1980). Using empirical Bayes techniques in law school validity studies. *Journal of the American Statistical Association, 75,* 801-816.

Saal, F.E., Downey, R.G., & Lahey, M. (1980). Rating the ratings: Assessing the psycho- metric quality of rating data. *Psychological Bulletin, 88,* 413-428.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *(Psycho- metric Monograph No. 17).* Psychometric Society.

Samejima, F. (1972). A general model for free response data. *(Psychometric Monograph No. 18)*. Psychometric Society.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42,* 193-198.

Sanders, P.F., Hendrix, A.C., & Luijten, A.J.M. (1984). De beoordeling van de samenvatting Nederlands. *Tijdschrift voor Taalbeheersing, 6,* 241-251.

Sanders, P.F., Theunissen, T.J.J.M., & Baas, S.M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika, 54,* 587-598.

Schouten, H.J.A. (1985). *Statistical measurement of interobserver agreement: Analysis of agreement and disagreement between observers.* Proefschrift, Rijksuniversiteit Utrecht.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park: Sage Publications.

Shepard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammond (red.). *Review of research in education: Vol. 19* (pp.405-450). Washington, DC: American Educational Research Association.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliabi-
lity. *Psychological Bulletin, 86,* 420-428.

Shumway, R.H., & Stoffer, D.S. (1982). An approach to time series smoothing and forcasting using EM algorithm. *Journal of Time Series Analysis,* 3, 253-264.

Siegel, S., & Castellan, N.J.Jr. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Sijtsma, K., & Molenaar, I.W. (1987). Reliability of test scores in non-parametric item response theory. *Psychometrika, 52,* 79-97.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, series B, 13,* 238-241.

Sirotnik, K. (1970). An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement, 30,* 891-908.

Sluijter, C., Boertien, H., de Klijn, W., & van Roosmalen, W. (1991). *De constructie van plaatsingstoetsen.* (Onderzoeksrapporten beginfase voortgezet onderwijs nr. 6). Arnhem: Cito.

Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics, 3,* 319-346.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Staphorsius, G. (1992a). Welk boek is gemakkelijk, mijnheer ? *RAIN informatiebulletin, 2,* 7-10.

Staphorsius, G. (1992b). *Clib-toetsen.* Arnhem: Cito.

Staphorsius, G., & Krom, R.S.H. (1985a). *Leesbaarheidsindex voor het basisonderwijs.* (Bulletin nr. 36). Arnhem: Cito.

Staphorsius, G., & Krom, R.S.H. (1985b). Predictie van leesbaarheid. *Tijdschrift voor Taal- beheersing, 7,* 192-211.

Stine, W.W. (1989). Interobserver relational agreement. *Psychological Bulletin, 106,* 341 -347.

Suen, H.K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale: Lawrence Erlbaum.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika, 50,* 411-420.

Theunissen, T.J.J.M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381-389.

Theunissen, T.J.J.M. (1987). Text banking and test design. *Language Testing, 4,* 1-8.

Thissen, D. (1988). *MULTILOG: Multiple categorical item analysis and test scoring using item response theory*. Mooresville: Scientific Software.

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

Thorndike, R.L. (1951). Reliability. In: E.F. Lindquist (red.). *Educational Measurement* (pp. 560-620). Washington, DC: American Council on Education.

Thorndike, R.L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Tinsley, H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 23,* 358-376.

Uebersax, J.S. (1984). *Reliablility, validity and the kappa coefficient.* (Technical Report No. 12). Austin: University of Texas.

Uebersax, J.S. (1991). *Quantitative methods for the analysis of observer agreement: Towards a unifying model.* Santa Monica: RAND Corporation.

Uiterwijk, J.H. (1990). Verschillen tussen autochtonen en allochtonen bij de overgang van basisonderwijs naar voortgezet onderwijs. In: C.A.C. Klaassen, & P.L.M.

Jungbluth (red.). *Onderwijs researchdagen 1990, onderwijs en samenleving.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Uiterwijk, J.H., & Engelen, R.J.H. (1993). *Verantwoording eindtoets basisonderwijs 1990.* Arnhem: Cito.

Umesh, U.N., Peterson, R.A., & Sauber, M.H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49,* 835-850.

Vale, C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Verhelst, N.D. (1989). Informatiewinst bij vertakt toetsen. In: W.J. van der Linden, & L.J.Th. van der Kamp (red.). *Meetmethoden en data-analyse* (pp. 89-96). Lisse: Swets en Zeitlinger.

Verhelst, N.D. (1993). *On the standard errors of parameter estimates in the Rasch model.* (Measurement and Research Department Reports 93-1). Arnhem: Cito.

Verhelst, N.D., Glas, C.A.W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly, 1,* 245-262.

Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. (PPON-rapport, nr. 4). Arnhem: Cito.

Verhelst, N.D., & Kamphuis, F.H. (1989). *Statistiek met* $\hat{\theta}$.(Bulletinreeks nr. 77). Arnhem: Cito.

Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.J.H.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model.* (Measurement and Research Department Reports 91-10). Arnhem: Cito.

Verhelst, N.D., & Veldhuijzen, N.H. (1991). *A new algorithm for computing elementary symmetric functions and their first and second derivatives.* (Measurement and Research Department Reports 91-1). Arnhem: Cito.

Verhelst, N.D., & Verstralen, H.H.F.M. (1991). *The partial credit model with non-sequential solution strategies.* (Measurement and Research Department Reports 91-5). Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (in druk). A dynamic generalization of the Rasch model. *Psychometrika, 58.*

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1993). *OPLM: One parameter logistic model.* Computer program and manual. Arnhem: Cito.

Verhelst, N.D., Verstralen.H.H.F.M., & Jansen, M.G.H. (1993) *A logistic model for time limit tests.* (Measurement and Research Department Reports 92-1). Arnhem: Cito.

Verschoor, A.J. (1991). *Optimal test design.* (Computer progamm and manual). Arnhem: Cito.

Verschoor. A.J., & Sanders, P.F. (1993). *Parallel test construction using the framework of classical test theory*. (Measurement and Research Department Reports 93-2). Arnhem: Cito.

Verstralen, H.H.F.M., & Verhelst, N.D. (1992). *The sample strategy of a test information function in computerized test design.* (Measurement and Research Department Reports 91-6). Arnhem: Cito.

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *Elementary School Journal, 28,* 373-381.

Wainer, H., & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. In: H. Wainer (red.). *Computerized adaptive testing: A primer* (pp. 65-101). Hillsdale: Lawrence Erlbaum.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426-482.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Weiss, D.J. (red.). (1983). *New horizons in testing.* New York: Academic Press.

Wijnstra, J.M. (1988). *Balans van het rekenonderwijs in de basisschool.* Arnhem: Cito.

Wilson, D.T., Wood, R., & Gibbons, R.T. (1991). *TESTFACT.* Chicago: Scientific Software.

Wilson, M., & G.N. Masters, (1993). The partial credit model and null categories. *Psycho- metrika, 58,* 87-99.

Witkin, H.A. (1950). Individual differences in ease of perception of embedded figures. *Jour- nal of Personality, 19,* 1-15.

Witkin, H.A., & Goodenough, D.R. (1981). Cognitive styles: Essence and origins. *Psychological Issues* (Monograph 51). New York: International Universities Press.

Wollenberg, A.L. van den (1979). *The Rasch model and time limit tests.* Nijmegen: Studentenpers.

Wollenberg, A.L. van den (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-140.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B.D., & Mead, R.J. (1977). *BICAL: Calibrating items and scales with the Rasch model.* (Research Memorandum 23). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

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

Yen, W.M. (1981). Using simultaneous results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yen, W.M. (1984). Tau-equivalence and equipercentile equating. *Psychometrika, 48,* 353- 369.

Zegers, F.E. (1989). Het meten van overeenstemming. *Nederlands Tijdschrift voor de Psychologie, 44,* 145-156.

Zegers, F.E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement, 15,* 321-333.

Zieky, M.J. (1987). *Methods of setting standards of performance on criterion referenced tests.* Paper presented at the 13th International Conference of the IAEA, Bangkok.

Zwinderman, A.H. (1991). *Studies of estimating and testing Rasch models.* (NICI Technical Report 91-02). Nijmegen: NICI.