References

Adams, D. (1978). *Hitchhiker's guide to the galaxy: The primary phase.* BBC. (Audio Recording.)

Adams, R., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Cáki (Eds.), *Proceedings of the 2nd international symposium on information theory* (p. 267-281).

Aleven, V., & Koedinger, K. R. (2002, Mar-Apr). An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*(2), 147-179.

Almond, R. G. (1995). *Graphical belief modeling.* Chapman and Hall. Available from `http://www.crcpress.com/product/isbn/9780412066610`

Almond, R. G. (2007a). Cognitive modeling to represent growth (learning) using Markov decision processes. *Technology, Instruction, Cognition and Learning (TICL)*, *5*, 313–324. Available from `http://www.oldcitypublishing.com/TICL/TICL.html`

Almond, R. G. (2007b). *An illustration of the use of Markov decision processes to represent student growth (learning)* (Tech. Rep. No. RR-07-40). ETS Research Report. Available from `http://www.ets.org/research/researcher/RR-07-40.html`

Almond, R. G. (2010a). 'I can name that Bayesian network in two matrixes'. *International Journal of Approximate Reasoning*, *51*, 167–178. Available from `http://dx.doi.org/10.1016/j.ijar.2009.04.005`

Almond, R. G. (2010b). Using evidence centered design to think about assessments. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs.* (pp. 75–100). Springer.

Almond, R. G., DiBello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., et al. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola & T. Richardson (Eds.), *Artificial intelligence and statistics 2001* (p. 137-143). Morgan Kaufmann.

Almond, R. G., Herskovits, E., Mislevy, R. J., & Steinberg, L. S. (1999). Transfer of information between system and evidence models. In D. Heckerman & J. Whittaker (Eds.), *Artificial intelligence and statistics 99* (p. 181-186).

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223-238.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2006). *Models for local dependence among observable outcome variables* (ETS Research Report No. RR-06-36). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-06-36.html`

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J.-D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning*, *50*, 450–460.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002a). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology , Learning, and Assessment*, *1*, (online). Available from `http://www.jtla.org/`

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002b). A framework for reusing assessment components. In H. Yanai, O. A., K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (p. 281-288). Springer.

Almond, R. G., Yan, D., & Hemat, L. A. (2008). Parameter recovery studies with a diagnostic Bayesian network model. *Behaviormetrika*, *35*(2), 159–185.

Almond, R. G., Yan, D., Matukhin, A., & Chang, D. (2006). *StatShop testing* (Research Memorandum No. RM-06-04). Educational Testing Service.

Andersen, S. A., Madigan, D., & Perlman, M. D. (1996). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, *25*, 505-41.

Anderson, R. D., & Vastag, G. (2004). Causal modeling alternative in operations research: Overview and application. *European Journal of Operational Research*, *156*(1), 92–109.

Andreassen, S., Woldbye, M., Falck, B., & Andersen, S. (1987). Munin—a causal probabilistic network for interpretation of elecromyographic findings. In *Proceedings of the 10th international joint conference on artificial intelligence.*

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2.0. *The Journal of Technology, Learning, and Assessment*, *4*(3), 13–18. Available from `http://escholorship.bc.edu/jtla/vol4/3/`

Bacchetti, P., Segal, M. R., & Jewell, N. P. (1993). Backcalculation of HIV infection rates (with discussion). *Statistical Sciences*, *8*, 82-119.

Baldwin, D., Fowles, M., & Livingston, S. (2008). *Guidelines for constructed-responses and other performance assessments* (Research Report No. RR-07-02). ETS. Available from `http://www.ets.org/Media/About_ETS/pdf/8561_ConstructedResponse_guidelines.pdf`

Barr, A., & Feigenbaum, E. (1982). *Handbook of artificial intelligence* (Vol. 2). Wadsworth International.

Bart, W. M., Post, T., Behr, M. J., & Lesh, R. (1994). A diagnostic analysis of a proportional reasoning test item: An introduction to the properties of a semi-dense item. *Focus on Learning Problems in Mathematics*, *16*(3), 1–11.

Barton, P. E. (2003). *Parsing the achievement gap: Baselines for tracking progress* (Policy Information Center Report). ETS. Available from `http://www.ets.org`

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 192–204.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Measurement*, *4*, 295-301.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, *14*, 237–245.

Bejar, I. I., Braun, H., & Tannenbaum, R. (2007). A prospective, predictive and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1–30). JAM Press.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–82). Lawrence Erlbaum Associates.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis,*. Springer-Verlag, New York.

Berliner, M. (2005). *Physical-statistical modeling and prediction.* Available from `http://www.stat.harvard.edu/Dempster_Symposium/Berliner.pdf` (Paper presented at "Some Challenging Applications of Statistical Modeling and Analysis", special seminar series presented at Harvard University on the occasion of the retirement of Arthur P. Dempster)

Bertelè, U., & Brioschi, F. (1972). *Nonserial dynamic programming.* Academic Press.

Best, N., Cowles, M. K., & Vines, K. (1996). Coda: Convergence diagnosis and output analysis software for Gibbs sampling output version 0.30 [Computer software manual].

Bishop, Y. M., Fienberg, S., & Holland, P. (1975). *Discrete multivariate analysis.* MIT Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, *5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-147. Available from `http://ditc.missouri.edu/docs/blackBox.pdf`

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., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrica*, *35*, 179–197.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, *11*, 1-94. Available from `citeseer.ist.psu.edu/boutilier99decisiontheoretic.html`

Box, G. E. P. (1976). Science and statistics. *JASA*, *71*, 791-799. (Source of the quote "All models are models are false, but some are useful.")

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. John Wiley and Sons. (Reprinted in Wiley Classics Library Edition, 1992.)

Boyen, X., & Koller, D. (1998). Tractable inference for complex stochastic process. In *In proceedings of the fourteenth annual Conference on Uncertainty in Artificial Intelligence* (p. 33-42). Morgan Kaufmann.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Breese, J. S., Goldman, R. P., & Wellman, M. P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on System, Man and Cybernetics*, *24*, 1577-1579.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skills* (Tech. Rep. No. 0-87447-280-6). College Entrance Examinations Board.

Brennan, R. L., & Prediger, D. J. (1977). *Coefficient kappa: Some uses, misuses, and alternatives* (Technical Bulletin No. 29). ACT.

Breyer, F. J., Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). Designing technology-based assessments: It's the evidence for the inferences that are important. In *Annual convention of the society for industrial organizational psychology.*

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (Tech. Rep. No. RR-01-23). Educational Testing Service.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434-55.

Bunt, A., & Conati, C. (2002). Assessing effective exploration in open learning environments using Bayesian networks. In *Intelligent tutoring systems.*

Bunt, A., & Conati, C. (2003). Probabilistic student modelling to improve exploratory behaviour. *User Modeling and User-Adapted Interaction*,

13(3), 269–309.

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, *2*, 159-225.

Buntine, W. L. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Trans. on Knowledge and Data Engineering*, *8*, 195-210.

Cai, L. (2008). Sem of another flavour: Two new applications of the supplemented em algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309-329.

Cannings, C., Thompson, E. A., & Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability*, *10*, 26-61.

Chaloner, K. M., & Duncan, G. T. (1983). Assessment of a beta prior distribution: PM elicitation. *The Statistician*, *32*, 174-180.

Chambers, J. L. (2004). *Programming with data: A guide to the S language*. Springer. ("Green" book.)

Chickering, D. (1996). Learning equivalence classes of Bayesian-network structures. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (p. 87-98). Morgan Kaufmann.

Cobb, B. R., & Shenoy, P. P. (2005). Hybrid Bayesian networks with linear deterministic variables. In F. Bacchus & T. Jaakkola (Eds.), *Proceedings of the twenty-first conference* (p. 136-144). AUAI Press.

Collis, J. M., Tapsfield, P. G. C., Irvine, S. H., Dann, P. L., & Wright, D. (1995). The British Army Recruit Battery goes operational: From theory to practice in computer-based testing using item generation techniques. *International Journal of Selection and Assessment*, *3*, 96-104.

Consortium, I. G. L. (2000). IMS question & test interoperability information model specification (Version 1.0 ed.) [Computer software manual]. Available from `http://www.imsproject.org`. (On-line document, retrieved May, 2000.)

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 309-347.

Council, N. R. (Ed.). (1996). *National science education standards*. National Acadamies Press.

Cowell, R. G., & Dawid, A. P. (1992). Fast retraction of evidence in a probabilistic expert system. *Statistics and Computing*, *2*, 36-41.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer.

Cowell, R. G., Dawid, A. P., & Spiegelhalter, D. J. (1993). Sequential model criticism in probabilistic expert systems. *IEEE Transaction of Pattern Analysis and Machine Intelligence*, *15*, 209-129.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. Chapman and Hall.

Cronbach, L. J. (1989). Intelligence: Measurement, theory, and public policy. In R. L. Linn (Ed.), *Construct validation after thirty years* (p. 147-171). University of Illinois Press.

Crowley, R., & Medvedeva, O. (2006, Jan). An intelligent tutoring system for visual classification problem solving. *Artificial Intelligence In Medicine*, *36*(1), 85-117.

Daniel, B., Zapata-Rivera, J.-D., & McCalla, G. (2003). A Bayesian computational model of social capital in virtual communities. In *Proceedings of the first international conference on communities and technologies: C & t 2003.* Kluwer Academic Publishers.

Darroch, J. N. S. L. L., & Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, *8*, 522-539.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B*, *41*, 1-31.

Dayton, C. M. (1999). *Latent class scaling analysis.* Sage.

Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computer Intelligence*, *5*, 142–150.

Deane, P., & Quinlan, T. (In press). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research.*

De Finetti, B. (1970). *Teoria delle probabilità.* Wiley classics Library. (Republished as in English as Theory of Probability, 1990)

DeGroot, M. H. (1970). *Optimal statistical decisions.* McGraw–Hill.

Dekhtyar, A., Finkel, R., Goldsmith, J., Goldstein, B., & Isenhour, C. (2005). Adaptive decision support for planning under hard and soft constraints. In M. J. Druzdzel & T.-Y. Leong (Eds.), *Challenges to decision support in a changing world: Papers from 2005 AAAI spring symposium.* AAAI Press. Available from `http://www.cs.engr.uky.edu/~goldsmit/papers/wtw1.pdf`

Dekhtyar, A., Goldsmith, J., Goldstein, B., Mathias, K. K., & Isenhour, C. (2010). Planning for success: The interdisciplinary approach to building Bayesian models. *International Journal of Approximate Reasoning*, *50*, 416–426.

Dempster, A. P. (1968). A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B*, *30*, 205-247.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157-175.

Dempster, A. P. (1990). Bayes, Fisher and belief functions. In S. Geisser, J. S. Hodges, S. J. Press, & A. Zellner (Eds.), *Bayesian likelihood methods in statistics and econometrics* (pp. 35–47). Amsterdam: Elsevier Science Publications.

Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JRSS B*, *39*, 1-38.

DeVito, P. J., & Koenig, J. A. (2000). *Investigating district-level and market-basket reporting.* National Research Council. Available from `http://www.nap.edu/catalog/9891.html`

Díez, F. J. (1993). Parameter adjustment in Bayes networks. the generalized noisy or-gate. In D. Heckerman & A. Mamdani (Eds.), *In uncertainty in artificial intelligence 93* (pp. 99–105). Morgan-Kaufmann.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer.

Draper, D. (1995). Assessment and propagation of model uncertainty. *JRSS B*, *57*, 45-98.

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N., & Rubin, D. B. (1987). *A research agenda for assessment and propagation of model uncertainty* (Rand Note No. N-2683-RC). The RAND Corporation.

Edwards, D. (1990). Hierarchical interaction models. *Journal of the Royal Statistical Society, Series B*, *52*, 3-20.

Edwards, D. (1995). *Introduction to graphical modelling*. Springer-Verlag.

Eid, M. (2002). A closer look at the measurement of change: Integrating latent state-trait models into the general framework of latent mixed Markov modeling. *Methods of Psychological Research Online*. Available from `http://www.mpr-online.de`

El Saadawi, G. M., Tseytlin, E., Legowski, E., Jukic, D., Castine, M., Fine, J., et al. (2008, Dec). A natural language intelligent tutoring system for training pathologists: implementation and evaluation. *Advances In Health Sciences Education*, *13*(5), 709-722.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380-396.

Enright, M. K., Moreley, M., & Sheehan, K. M. (1999). *Items by design: The impact of systematic feature variation on item statistical characteristics* (Research Report Nos. RR-99-20, GREB-95-15R). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-99-20.html`

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*, 985-987.

Feller, W. (1968). *An introduction to probability theory and its applications (3rd ed)*. Wiley.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Wiley.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, *38*, 87–111.

Freedman, D., Pisani, R., & Purves, R. (1980). *Statistics*. W. W. Norton and Company.

Gamboa, H. (2001). Designing intelligent tutoring systems: a Bayesian approach. In *Proceedings of the ana fred 3rd international conference on enterprise information systems ICEIS* (pp. 452–458).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995/2003). *Bayesian data analysis* (2nd. ed.). Chapman and Hall.

Gelman, A., Meng, X. L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733-807.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions,and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Gertner, A., Conati, C., & VanLehn, K. (1998). Procedural help in andes: Generating hints using a Bayesian network student model. In *Proceedings of the fifteenth national conference on artificial intelligence AAAI-98* (p. 106-111). The MIT Press.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment: Theories and applications* (pp. 242–274). Cambridge University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall.

Gitomer, D. H., & Steinberg, L. S. (1999). Representational issues in assessment design. In I. E. Sigel (Ed.), *Development of mental representation* (p. 351-370). Erlbaum.

Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. D. Nichols, S. F. Chipman, & R. L. Brennen (Eds.), *Cognitively diagnostic assessment* (pp. 73–101). Lawrence Erlbaum.

Glas, C. A. W., & van der Linden, W. J. (2001, July). Modeling variability in item parameters in cat. In *International meeting of the psychometrics society.*

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247–261.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J. C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement:the Buros-Nebraska symposium on measurement and testing* (Vol. 3, p. 41-85). Erlbaum.

Glasziou, P., & Hilden, J. (1989). Test selection measures. *Medical Decision Making*, *9*, 133-141.

Glück, J., & Spiel, C. (2007). Studying development via item response models: A wide range of potential uses. In M. von Davier & C. H. Carstense (Eds.), *Multivariate and mixture distribution rasch models: Extensions and applications* (pp. 281–292). Springer.

Good, I. J. (1952). Rational decisions. *JRSS B*, *14*, 104-114.

Good, I. J. (1971). 46656 varieties of Bayesian. *American Statistician*, *25*, 62-63. (Reprinted in Good Thinking University of Minnesota Press, 1983, 20–21.)

Good, I. J. (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. In C. A. Hooker & W. Harper (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science, vol. ii* (Vol. 2, p. 125-174). Dordrecht Reidel Publishing Company. (Reprinted in Good Thinking, University of Minnesota Press, 1983, 22–55.)

Good, I. J. (1983). *Good thinking*. U. Minnesota Press.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. Bernardo, M. DeGroot, D. Lindley, & A. Smith (Eds.), *Bayesian statistics 2* (p. 249-269). North Holland.

Good, I. J., & Card, W. (1971). The diagnostic process with special reference to errors. *Methods of Information in Medicine*, *10*, 176-188.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764. Available from `http://www.jstor.org/stable/2281536`

Graesser, A. C., VanLehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*(22), 39–52.

Graf, E. A. (2003, September). *Designing a proficiency model and associated item models for a mathematics unit on sequences.* Princeton, NJ. (Paper presented at the Cross Division Math Forum)

Graf, E. A. (2008). *Approaches to the design of diagnostic item models* (Research Report No. RR-08-07). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-08-07.html`

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B*, *29*, 83-100.

Haberman, S. J. (1972). Log-linear fit for contingency tables — algorithm as51. *Applied Statistics*, *21*, 218-225.

Haberman, S. J. (2005). *Latent-class item response models* (Research Report No. RR-05-28). ETS.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, *8*, 333-346.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement test items. *Journal of Educational Measurement*, *26*, 301-321.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hansen, E. G., & Mislevy, R. J. (2004). Toward a unified validity framework for ensuring access to assessments by individuals with disabilities and english language learners. In *Annual meeting of the national council on*

measurement in education (NCME). Available from `http://www.ets.org/research/dload/NCME2004-Hansen.pdf`

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2003). Evidence-centered assessment design and individuals with disabilities. In *Annual meeting of the national council on measurement in education.* Available from `http://www.ets.org/research/dload/ncme03-hansen.pdf`

Hartz, S., Roussos, L., & Stout, W. (Submitted). The fusion model for cognitive diagnosis: blending theory with practicality. *Psychometrika*, *TBA*, TBA.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Heckerman, D. (1991). *Probabilistic similarity networks.* ACM Press.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (p. 301-354). Kluwer Academic Publishers. (Also Technical Report MSR-TR-95-06, Microsoft Research, March, 1995 (revised November, 1996).)

Heckerman, D., Gieger, D., & Chikering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*, 197-243.

Heckerman, D., Horvitz, E., & Middleton, B. (1993). An approximate nonmyopic computation for value of information. *IEEE Transaction of Pattern Analysis and Machine Intelligence*, *15*, 292-298.

Heidelberger, P., & Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, *24*, 233-245.

Henrion, M., & Druzdzel, M. (1990). Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. In *Proceedings of the 6th conference on uncertainty in artificial intelligence* (p. 10-20).

Hensen, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262–277.

Henze, N., & Nejdl, W. (1999). Student modeling in an active learning environment using bayesian networks. In *In proceedings of the seventh international conference on user modeling, UM99.* Springer.

Hilden, J. (1970). GENEXX — an algebraic approach to pedigree probability calculus. *Clinical Genetics*, *1*, 319-348.

Hoey, J., St-aubin, R., Hu, A., & Boutilier, C. (2001). SPUDD: Stochastic planning using decision diagrams. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence.*

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Lawrence Erlbaum Associates.

Howard, R. A., & Matheson, J. E. (1981). Influence diagrams. In *Principles and applications of decision analysis.* Menlo Park, CA: Strategic Decisions Group.

Hrycej, T. (1990). Gibbs sampling in Bayesian networks. *Artificial Intelligence*, *46*, 351-363.

Irvine, S. H., Dann, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology*, *81*, 173-195.

Irvine, S. H., & Kyllonen, P. (Eds.). (2002). *Generating items for cognitive tests: Theory and practice*. Erlbaum.

Jaakkola, T. S. (2001). Tutorial on variational approximation methods. In M. Opper & D. Saad (Eds.), *Advanced mean field methods: Theory and practice* (p. 129-159). MIT Press.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *SSC-4*, 227-241.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Jensen, F. V. (1988). *Junction trees and decomposable hypergraphs* (Tech. Rep.). JUDEX Research Report, Aalborg, Denmark.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. Springer-Verlag.

Johnson, M. S., & Sinharay, S. (2003). *Calibration of polytomous item families using Bayesian hierarchical modeling* (Research Report No. RR-03-23). ETS.

Jordan, M. I. (Ed.). (1998). *Learning in graphical models*. Kluwer Academic Publishers. (Reprinted by MIT Press.)

Joreskog, K. G., & Sorbom, D. (1979). *Advances in factor analysis and structural equation models*. Abt Books.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kadane, J. B. (1980). Predictive and structural methods for eliciting prior distributions. In A. Zellner (Ed.), *Bayesian analysis and statistics*. North-Holland.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., & Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *JASA*, *75*, 845-854.

Kadie, C. M., Hovel, D., & Horvitz, E. (2001). MSBNx: A component-centric toolkit for modeling and inference with Bayesian networks [Computer software manual]. Available from `http://www.research.microsoft.com/adapt/MSBNx/` (Microsoft Research Technical Report MSR-TR-2001-67.)

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 17–64). American Council on Education/Praeger.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Sage Publications.

Katz, I. R., Williamson, D. M., Nadelman, H. L., Kirsch, I., Almond, R. G., Cooper, P. L., et al. (2004). *Assessing information and communications technology literacy for higher education.* (Paper presented at the 30th annual conference of the International Association for Educational Assessment)

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2006). *ConceptMap.* Computer Program, Bear Center: UC Berkeley, CA. Available from `http://bearcenter.berkeley.edu/GradeMap` (Previous version was known as GradeMap.)

Kim, J. H., & Pearl, J. (1983). A computational model for causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th international joint conference on artificial intelligence* (p. 190-193). William Kaufmann.

Klein, M. F., Birenbaum, M., Standiford, S. N., & Tatsuoka, K. K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions* (Research Report No. 81-6). Computer-based Education Research Laboratory, University of Illinois.

Koedinger, K. R., & Aleven, V. (2007, Sep). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, *19*(3), 239-264.

Koller, D., & Learner, U. (2001). Sampling in factored dynamic systems. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential Monte Carlo methods in practice* (p. 445-464). Springer.

Koller, D., & Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the thirteenth conference on uncertainty in artificial intelligence (UAI-97)* (p. 302-313). Available from `http://citeseer.nj.nec.com/koller97objectoriented.html`

Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, *18*, 416-441.

Laskey, K. B., & Mahoney, S. M. (2000). Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, *12*, 481-486.

Lauritzen, S. L. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, *87*, 1098-1108.

Lauritzen, S. L. (1996). *Graphical models oxford statistical science series.* Claredon Press.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computation with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B*, *50*, 205-247. (Reprinted in Shafer and Pearl (1990))

Lee, P. M. (1989). *Bayesian statistics: An introduction.* Oxford University Press.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment: Theories and applications*. Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*, 205–236.

Ley, T., Kump, B., & Albert, D. (2010, APR). A methodology for eliciting, modelling, and evaluating expert knowledge for an adaptive work-integrated learning system. *International Journal Of Human-Computer Studies*, *68*(4), 185-208.

Li, Z., & D'Ambrosio, B. (1994). Efficient inference in Bayes nets as a combinatorial optimization problem. *Intl Journal of Approximate Reasoning*, *11*, 55-81.

Linden, W. J. van der. (2005). *Linear models for optimal test design*. Springer-Verlag.

Linden, W. J. van der, & Glas, C. A. W. . (2010). *Elements of adaptive testing*. Springer.

Little, R., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer.

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

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B*, *44*, 226-233.

Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, *28*, 3049–3082.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Lynch, S. M. (2007). *Introduction to applied bayesian statsticis and estimation for social sciences*. Springer.

Madigan, D., & Almond, R. G. (1995). Test selection strategies for belief networks. In D. Fisher & H. J. Lenz (Eds.), *Learning from data: AI and Statistics V* (p. 89-98). Springer-Verlag.

Madigan, D., Gavrin, J., & Raftery, A. E. (1995). Enhancing the predictive performance of Bayesian graphical models. *Communications in Statistics: Theory and Methods*, *24*, 2271-2292.

Madigan, D., Hunt, E., & Levidow, B. (1995). *Bayesian graphical modeling for intelligent tutoring systems* (Tech. Rep.).

Madigan, D., Mosurski, K., & Almond, R. G. (1997). Graphical explanation in belief networks. *Journal of Computational Graphics and Statistics*, *6*(2), 160-181. Available from `http://www.amstat.org/publications/jcgs/index.cfm?fuseaction=madiganjun`

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *JASA*, *89*, 1535-1546.

Madigan, D., Raftery, C., A. E.and Volinsky, & Hoeting, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models.*

Mahoney, S. M., & Laskey, K. B. (1996). Network engineering for complex belief networks. In E. Horvitz & F. Jensen (Eds.), *In proceedings of the 12th conference on uncertainty in artificial intelligence (UAI-96)* (pp. 389–396). Morgan Kaufmann.

Martin, J., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. D. Nichols, S. F. Chipman, & R. L. Brennen (Eds.), *Cognitively diagnostic assessment* (p. 141 - 165). Lawrence Erlbaum Associates, Inc.

Matheson, J. E. (1990). Using influence diagrams to value information and control. In R. M. Oliver & S. J. Q. (Eds.), *Influence diagrams, belief nets and decision analysis* (pp. 25–48). John Wiley & Sons.

Mathias, K. K., Isenhour, C., Dekhtyar, A., Goldsmith, J., & Goldstein, B. (2006). *Eliciting and combining influence diagrams: Tying many bowties together* (Technical Report No. TR453-06). University of Kentucky, Department of Computer Science. Available from `http://www.cs.uky.edu/~dekhtyar/publications/TR453-06.pdf`

Matsuda, N., & VanLehn, K. (2000). Decision theoretic instructional planner for intelligent tutoring systems. In *Proceedings workshop on modeling human teaching tactics and strategies, its2000.*

Mayo, M., & Mitrovic, A. (2001). Optimising its behaviour with bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, *12*(2), 124–153.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models. (2nd edition).* Chapman and Hall.

Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensiona, and mixutre distribution rasch models for the analysis of repeated observations. *Methods of Psychological Research Online*, *3*(2). Available from `http://www.mpr-online.de//issue5/art6/article.html`

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (p. 13-103). American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23(2)*, 13-23.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335-341.

Millan, E., Agosta, J., & Cruz, J. de la. (2001, MAR). Bayesian student modeling and the problem of parameter specification. *British Journal of Educational Technology*, *32*(2), 171-181.

Millan, E., & Cruz, J. Perez-de-la. (2002). A bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, *12*(2-3), 281-330.

Miller, P. (1983). Attending: Critiquing a physician's management plan. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *5*, 449-461.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *12*, 341-369.

Mislevy, R. J. (1995a). *Information-decay pursuit of dynamic parameters in student models* (Research Report No. RM-94-14-onr). ETS. Available from `http://www.ets.org/research/policy_research_reports/rm-94-14-onr`

Mislevy, R. J. (1995b). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennen (Eds.), *Cognitively diagnostic assessment* (p. 43-71). Lawrence Erlbaum.

Mislevy, R. J. (1995c). Test theory and language learning in assessment. *Language Testing*, *12*, 341-369.

Mislevy, R. J. (1996). *Bayes modal estimation of item parameters.*

Mislevy, R. J. (2013). Missing responses in item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory (2nd edition).* Chapman & Hall.

Mislevy, R. J., Almond, R. G., & Steinberg, L. S. (1998). *A note on knowledge-based model construction in educational assessment* (CSE Technical Report No. 480). The National Center for Research on Evaluation, Standards, Student Testing (CRESST). Available from `http://www.cresst.org/reports/TECH480.pdf` (Numerical example for testing Ergo.)

Mislevy, R. J., Almond, R. G., & Steinberg, L. S. (2002). Design and analysis in a task-based language assessment. *Language Testing*, *19*(4), 477-496.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In K. B. Laskey & H. Prade (Eds.), *Uncertainty in artificial intelligence '99* (p. 437-446). Morgan-Kaufmann.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, *5*, 253-282.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*, 55-78.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (p. 97-128). Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisci-*

plinary Research and Perspective, *1*(1), 3-62.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G., & Penuel, W. (2003). Leverage points for improving educational assessment. In B. Means & G. Haertel (Eds.), *Evaluating the effects of technology in education* (p. 149-180). Erlbaum.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–47). Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior*, *15*, 29-42.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2001). *Making sense of data from complex assessments* (CSE Technical Report No. 538). The National Center for Research on Evaluation, Standards, Student Testing (CRESST). Available from `http://www.cresst.org/reports/TECH538.pdf`

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (Research Report No. RR-94-28-ONR). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-94-28-ONR.html`

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and Bayesian irt ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30-ONR). ETS.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press.

Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, *10*, 11-33.

Murphy, K., & Russell, S. (2001). Rao-blackwellised particle filtering for dynamic Bayesian networks. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential Monte Carlo methods in practice* (p. 499-515). Springer.

Murphy, K. P., Weiss, Y., & Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In K. B. Laskey & H. Prade (Eds.), *Uncertainty in artificial intelligence, proceedings of the fifteenth conference (UAI 99)* (p. 467-475). Morgan Kaufmann.

Murray, R., VanLehn, K., & Mostow, J. (2004). Looking ahead to select tutorial actions: A decision-theoretic approach. *International Journal of Artificial Intelligence in Education*, *14*(3, 4), 235–278.

Neal, R. M. (2003). Slice sampling (with discussion). *Annals of Statistics*, *31*, 705-767.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems:theory and algorithms*. Wiley.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Prentice Hall.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Prentice-Hall, Englewood Cliffs, NJ.

Neyman, J., & Scott, E. L. (1948). Consistent estimators based on partially consistent observations. *Econometrika*, *16*, 1-32.

Nichols, P. D., Chipman, S. F., & Brennen, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. Lawrence Erlbaum.

Nicholson, A. E., & Jitnah, N. (1998). Using mutual information to determine relevance in Bayesian networks. In *Pacific rim international conference on artificial intelligence* (p. 399-410). Available from `http://citeseer.ist.psu.edu/nicholson98using.html`

Nikovski, D., & Brand, M. (2003). Model minimization of dynamic belief networks for group elevator control. In *Proceedings of the 1st Bayesian modeling application workshop of the 19th conference on uncertainty in artificial intelligence.*

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). Springer-Verlag.

Norsys, Inc. (2004). *Netica.* Available from `http://www.norsys.com` (Bayesian network Computer Software)

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika*, *63*, 329-333.

Oliver, R. M., & Smith, J. Q. (1990). *Influence diagrams,belief nets and decision analysis*. John Wiley and Sons.

Patz, R. J., & Junker, B. W. (1999). A straight forward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, *27*(2), 226–284. Available from `http://smr.sagepub.com/content/27/2/226`

Pelligrino, J., Glaser, R., & Chudowsky, N. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Research Council.

Plummer, M. (2012, May). JAGS version 3.2.0 user manual (3.2.0 ed.) [Computer software manual]. Available from `http://mcmc-jags.sourceforge.net/`

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). coda: Output analysis and diagnostics for MCMC [Computer software manual]. (R package version 0.10-7)

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Ramsey, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika*, *40*, 337–360.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* The University of Chicago Press.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 35–64). American Psychological Association.

Reye, J. (2004). Student modelling based on belief networks. *Internationl Journal of Artificial Intelligence in Education*, *14*, 63–96.

Rijmen, F. (2008). Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, *48*, 659–666.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2007). Latent class models for diary method data: parameter estimation by local computations. *Psychometrica*. (To Appear)

Rissanen, J. (1987). Stochastic complexity (with discussion). *JRSS B*, *49*, 223–265.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007, APR). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin and Review*, *14*(2), 249-255.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). The asymptotic distribution of p-values in composite null models (with discussion). *Journal of the American Statistical Association*, *95*(422), 1143–1172. Available from `http://www.hsph.harvard.edu/causal/publications/p-values.pdf`

Ross, S. M. (1988). *A first course in probability.* Macmillan.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment: Theories and applications* (pp. 281–292). Cambridge University Press.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using irt-based latent class models. *Journal of Educational Measurement*, *44*(4), 293–311.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondentsin sample surveys. *Journal of the American Statistical Association*, *72*, 538-543.

Rubin, D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rupp, A. A., Templin, J., & Hensen, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications.* Guilford Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, *34*(4), (Part 2).

Savage, L. J. (1972). *The foundations of statistics (second edition).* Dover. ((First edition, 1954, Wiley.))

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning.* Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Science, A. A. for the Advancement of. (1994). *Benchmarks for scientific literacy.*

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research*, *34*, 871–82. (Reprinted in Shafer and Pearl (1990).)

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton University Press.

Shafer, G. (1996). *The art of causal conjecture.* MIT Press.

Shenoy, P. P. (1991). A fusion algorithm for solving Bayesian decision problems. In *Uncertainty in artificial intelligence, proceedings of the seventh conference* (pp. 361–369).

Shenoy, P. P., & Shafer, G. (1990). Axioms for probability and belief-function propagation. In *Uncertainty in artificial intelligence 4* (p. 169-198). (Reprinted in Shafer and Pearl[1990])

Shute, V. J. (2004). Towards automating ecd-based diagnostic assessments. *Technology, Instruction, Cognition, and Learning*, *2*(1-2), 1-18.

Shute, V. J. (2006). *Assessments for learning: Great idea, but do they work?* (Paper presented at the annual meeting of the American Educational Research Association (AERA))

Shute, V. J., Graf, E. A., & Hansen, E. G. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. M. Pytlikzillig, R. H. Bruning, & M. Bodvarsson (Eds.), *Technology-based education; bringing researchers and practitioners together* (p. 169-202). Greenwich, CT: Information Age Publishing.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: The impact of feedback and adaptivity on learning.* (Research Report No. RR-07-26). ETS. Available from `http://www.ets.org/research/researcher/RR-07-26.html`

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it - or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education*, *18*(4), 289–316. Available from `http://www.ijaied.org/iaied/ijaied/abstract/Vol_18/Shute08.html`

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, &

P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Routledge, Taylor and Francis.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence* (1st ed.). Oxford University Press, USA.

Sinharay, S. (2003). *Assessing convergence of the Markov chain Monte Carlo algorithms: A review* (Research Report No. RR-03-07). ETS.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, *42*(4), 375–394.

Sinharay, S. (2006, SPR). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*(1), 1-33.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitively diagnostic models—a case study. *Educational and Psychological Measurement*, *67*(2), 239–257.

Sinharay, S., Almond, R. G., & Yan, D. (2004). *Assessing fit of models with discrete proficiency variables in educational assessment* (Research Report No. RR-04-07). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-04-07.html`

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 197–219.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, *22*(4), 26–33.

Spandel, V., & Stiggins, R. L. (1990). *Creating writers: Linking assessment and writing instruction* (2nd ed.). Longman.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, *64*, 583–639.

Spiegelhalter, D. J., & Knill-Jones, R. (1984). Statistical and knowledge-based approaches to clinical decision support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society, (Series A)*, *147*, 35–77.

Spiegelhalter, D. J., & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, *20*, 579–605.

Spiegelhalter, D. J., Thomas, A., Best, N., & Gilks, W. (n.d.). Bugs 0.5 examples volume 1 (version i) [Computer software manual]. Available from `http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml`

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1995). Bugs: Bayesian inference using Gibbs sampling, version 0.50 [Computer software manual]. Cambridge. Available from `http://www.mrc-bsu.cam.ac.uk/bugs/`

Spirtes, P., Meek, C., & Richardson, T. S. (1997). A polynomial-time algorithm for determining dag equivalence in the presence of latent variables

and selection bias. In D. Madigan & P. Smythe (Eds.), *Preliminary papers of the sixth international workshop on AI and statistics* (pp. 489–501).

Srinivas, S. (1993). A generalization of the noisy-or model, the generalized noisy or-gate. In D. Heckerman & A. Mamdani (Eds.), *Uncertainty in artificial intelligence '93* (pp. 208–215). Morgan-Kaufmann.

Steinberg, L. S., Almond, R. G., Baird, A. B., Cahallan, C., Chernick, H., Dibello, L. V., et al. (2003). *Introduction to the Biomass project: An illustration of evidence-centered assessment design and delivery capability* (CSE Report No. 609). National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Available from `http://www.cse.ucla.edu/reports/R609.pdf`

Steinberg, L. S., & Gitomer, D. H. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science*, *24*, 223–258.

Stevens, R. H., & Thadani, V. (2007). Quantifying student's scientific problem solving efficiency and effectiveness. *Technology, Instruction, Cognition, and Learning*, *5*(4), 325–338.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (p. 284-300). Macmillan.

Suermondt, H. (1992). *Explanation in Bayesian belief networks.* Unpublished doctoral dissertation, Departments of Computer Science and Medicine, Stanford University.

Suppes, P. (1969). Stimulus response theory of finite automata. *Journal of Mathematical Psychology*, *6*, 327–355.

Takikawa, M., D'Ambrosia, B., & Wright, E. (2002). Real-time inference with large-scale temporal bayes nets. In J. Breese & D. Koller (Eds.), *Proceedings of the 18th UAI conference.* Morgan Kaufmann.

Tanimoto, S. (2001). Distributed transcripts for online learning: Design issues. *Journal of Interactive Media in Education*, *2001*(2). Available from `http://www-jime.open.ac.uk/2001/2/`

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems* (Vol. 20; NIE Final report No. NIE-G-81-002). University of Illinois, Computer-based Education Research.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennen (Eds.), *Cognitively diagnostic assess-*

ment (pp. 327–359). Lawrence Erlbaum.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement*, *25*(4), 301-319.

Tatsuoka, M. M., & Tatsuoka, K. K. (1989). Rule space. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 217–220). Wiley.

te Marvelde, J. M., Glas, C. A. W., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*(1), 5–34. Available from `http://epm.sagepub.com/cgi/content/abstract/66/1/5`

Thissen, D., & Wainer, H. (2001). *Test scoring*. Lawrence Erlbaum Associates.

Thomas, A., Spiegelhalter, D. J., & Gilks, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4*. (pp. 837–842). Clarendon Press.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

van de Pol, F., Langeheine, R., & de Jong, W. (1996). *Panmark 3: User's manual* (Computer Software Manual). Statistics Netherlands. Available from `http://www.assess.com/xcart/product.php?productid=245&cat=32&page=1`

van der Gaag, L. C., Bodlaender, H. L., & Feelders, A. (2004). Monotonicity in Bayesian networks. In M. Chickering & J. Halpern (Eds.), *Proceedings of the twentieth conference on uncertainty in artificial intelligence* (pp. 569–576). AUAI.

VanLehn, K. (n.d.). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (p. 113-138). Erlbaum.

VanLehn, K., & Martin, J. (1997). Evaluation on an assessment system based on bayesian student modeling. *International Journal of Artificial Intelligence in Education*(8), 179–221.

Vermunt, J. K. (1993). *Lem: Log-linear and event history analysis with missing data using the em algorithm* (Computer Software Manual). Tillburg University. (WORC paper 93.09.015/7)

Vomlel, J. (2002). Exploiting functional dependence in Bayesian network inference. In *Proceedings of the eightteenth conference on uncertainty in artificial intelligence (uai)* (pp. 528–535). AUAI.

Vomlel, J. (2003). Two applications of bayesian networks. In *Proceedings of conference znalosti 2003* (pp. 73–82).

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, *12*, 83–100.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: ETS. Available from `http://www.ets.org/research/researcher/RR-05-16.html`

Vygotsky, L. (1978). *Mind in society: The development of higher mental processes*. Harvard University Press.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applicaitons*. Cambridge University Press.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., et al. (2000/2001). *Computerized adaptive testing: A primer (second edition)*. Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall.

Weaver, W. (1948). Probability, rarity, interest, and surprise. *Scientific Monthly*, *67*, 390–392.

Weiss, Y. (2000). Correctness of local probability propagation in graphical models with loops. *Neural Computation*, *12*, 1–41.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, *16*, 3-118.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Williamson, D. M. (2000). *Utility of model criticism indices for Bayesian inference networks in cognitive assessment*. Unpublished doctoral dissertation, Fordham University.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. *International Journal of Testing*, *4*, 303–332.

Williamson, D. M., Mislevy, R. J., & Almond, R. G. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainty in artificial intelligence 16* (p. 634-643). Morgan Kaufmann.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Lawrence Erlbaum Associates.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Psychology Press.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2002). *Testfact: Test scoring, item statstics, and item factor analysis* (Computer Software Manual). Scientific Software International.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–85.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215.

Yan, D., Almond, R. G., & Mislevy, R. J. (2004). *Comparison of two models for cognitive diagnosis* (Research Report No. RR-04-02). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-04-02.html`

Yan, D., Lewis, C., & Stocking, M. (2002). *Adaptive tesing without irt in the presence of multidimensionality* (Research Report No. RR-02-09). Educational Testing Service.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (Research Report No. RR-03-32). Educational Testing Service. Available from `http://www.ets.org/research/researcher/RR-03-32.html`

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

York, J. (1992). Use of the Gibbs sampler in expert systems. *Artificial Intelligence*, *56*, 115–130.

Zapata-Rivera, J., & Greer, J. (2004, OCT). Inspectable bayesian student modelling servers in multi-agent tutoring systems. *International Journal of Human-Computer Studies*, *61*(4), 535-563.

Zapata-Rivera, J. D., & Greer, J. E. (2004). Interacting with inspectable Bayesian student models. *International Journal of Artificial Intelligence in Education*, *14*(2), 127-163.

Zapata-Rivera, J. D., Neufeld, E., & Greer, J. (1999). Visualization of bayesian belief networks. In *Ieee visualization 1999 late breaking hot topics proceedings* (pp. 85–88). Press.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG): Multiple-group irt analysis and test maintenance for binary items [Computer software manual]. (Scientific Software.)