Chapter 43
Developing Measurement Instruments for Science Education Research

Xiufeng Liu

Standardized measurement instruments (SMIs) refer to tools that produce valid and reliable quantitative measures about a construct. Development of SMIs in science education has been an active field of research for the past five decades (Doran et al. 1994; Tamir 1998), which is particularly true for large-scale studies in science education (Britton and Schneider 2007). SMIs have been receiving increasing attention over the past decade for a number of reasons. First, there is a growing worldwide trend toward standards-based science education in which standardized testing is used for accountability. Second, there is a growing realization of limitations of qualitative research approaches and a call for randomized experimentation that incorporates standardized measurements (National Research Council [NRC] 2002). Third, the continuing interest in identifying student alternative conceptions has created a demand for more efficient and large-scale survey of student alternative conceptions. Today, SMIs are playing a vital role in various science education research programs and will continue to do so in the future.

This chapter reviews the development of SMIs in refereed science education publications by excluding commercial measurement instruments, those developed for large-scale state, national, and international assessments, and instruments reported in theses, dissertations, and conferences. For a comprehensive review of large-scale standardized measurement in science education, refer to Edward Britton and Steven Schneider (2007); for a comprehensive review of SMIs in science education research over the past 50 years in North America, refer to Xiufeng Liu (2009). This chapter is divided into three sections: an overview of SMIs developed since 1990 in terms of their content, target population, validation, and reliability; approaches to and issues associated with developing SMIs; and desirable future directions for developing SMIs for science education research.

X. Liu (✉)
Department of Learning and Instruction, Graduate School of Education, State University of New York at Buffalo, Buffalo, NY 14260-1000, USA
e-mail: xliu5@buffalo.edu

B.J. Fraser et al. (eds.), *Second International Handbook of Science Education*, Springer International Handbooks of Education 24, DOI 10.1007/978-1-4020-9041-7_43,

Overview of Standardized Measurement Instruments

A search for SMIs reviewed in the *Buros Mental Measurement Yearbooks* (Spies and Plake 2005) database returned only one entry. It is apparent that Buros yearbooks miss most standardized measurement instruments for science education research. A search of the ERIC database from 1990 to the present using *measurement techniques* and *science education* as descriptors returned 229 entries. After going through the abstracts and examining relevant websites cataloguing various measurement instruments, 49 SMIs reported in refereed publications were located (with the others being related to measurement instruments for science laboratories, or measurement instruments for other subjects such as mathematics, computer science, and so on). The above measurement instruments cover the following areas of science education research (the number of instruments is in the parenthesis): conceptual understanding (15), attitudes (11), cognitive reasoning (3), nature of science (5), learning environment (9), and teacher beliefs and practices (6). The list of 49 SMIs organized by the content area and then the publication year is available in the Appendix. Although these SMIs might not be exhaustive of all instruments published in refereed publications, they are likely to represent the SMIs developed in science education in the past 18 years.

Approaches To and Issues Associated with Developing Standardized Measurement Instruments

One central component of developing SMIs is to establish evidence of validity. Conceptions of validity have evolved considerably over the years. Validity used to be solely concerned with prediction. Later on, validity evolved into three types: content, criterion-related (i.e., predictive and concurrent), and construct. Validity is an integrated notion called construct validity. Establishing the construct validity of an instrument is to develop coherent and empirical arguments to support the intended interpretation or use of measurement scores (Kane 2006). Thus, there is no absolute validity; validity is closely tied to the intended interpretations and uses of scores.

Related to validity is the issue of reliability. Similarly, much change has taken place over the years in the conceptualization of reliability. Although the central concern of reliability remains the consistency of scores across repeated applications of a measurement instrument, approaches to establishing evidence of reliability have changed significantly. Generalizability theory is now the overarching conceptual framework for reliability (Haertel 2006); internal consistency as measured by KR–20 and Cronbach's alpha represent only one possible source of inconsistency in scores.

It is apparent that the above conceptual frameworks of validity and reliability have influenced the development of SMIs since 1990. The most important issues when evaluating a measurement instrument are the appropriateness of the defined construct and the intended population of the measurement instrument. An instrument validated for one population might not be valid for a different population. Only after the evaluation of these two issues should the focus of instrument evaluation shift to reported technical properties of items (e.g., item difficulty and discrimination) and the instrument (e.g., content validity, criterion-related validity, and reliability). Given that there can be a variety of different ways of establishing validity and reliability, it is important to examine the relevance of reported validity and reliability evidence to the intended use of the instrument. On the other hand, because statistics based on Classical Test Theory (CTT), which is the foundation of most of the above SMIS, are always sample dependent, and in many cases the samples used for validation are local or convenient samples, it is always necessary to continue validating an instrument.

A large number of SMIs (15) developed since 1990 are related to assessing student conceptual understanding of science concepts. This is probably due to the continued effect of the worldwide alternative conceptions movement (ACM) from the early 1970s to the l990s (Wandersee et al. 1994). Although ACM was primarily based on qualitative research, the development of many SMIs since 1990 was based on rich findings of qualitative research, which made possible large-scale diagnosis of students' alternative conceptions. Validation of the above conceptual measurement instruments has been typically based on expert content reviews for content validity and student interviews and/or factor analysis for construct validity. Because of the fact that all these instruments use multiple-choice questions, reliability is typically established based on KR–20 or Cronbach's alpha. One important issue related to construct validity is the use of diagnostic instruments for summative purposes. At issue is unidimensionality, which is concerned with the question of whether a set of items measure the same construct so that scores on the items can be summed. Without having established unidimensionality, we cannot add individual item scores to obtain a total score, which makes it impossible to compare the gains in total scores from pretest to posttest, or the difference in total scores between two curriculum innovations. Based on principal component and confirmatory factor analysis, some of the instruments (such as FCI, CSEM, CINS, and DIRECT; see Appendix) were found to be multidimensional. Using these instruments for a summative purpose could potentially undermine the construct validity of the scores.

Eleven SMIs in the Appendix are related to attitudes. The variety of standardized measurement instruments for attitudes reflects diverse theoretical frameworks related to attitude. The diversity in theoretical frameworks requires that an attitude instrument is based on a clearly defined construct. For example, Zacharias Zacharia and Angela Calabrese Barton (2004) differentiated two types of student science attitude: attitude toward progressive school science, and attitude toward critical school science. However, not all attitude instruments in the Appendix have clearly defined attitude constructs.

Six SMIs pertain to teacher beliefs and practices. One instrument made a differentiation between teacher beliefs and teacher practices (Wang and Marsh 2002). This distinction is very important because the two are not necessarily always the same. Identifying the discrepancy between teachers' beliefs and practices can inform ongoing science education reforms so that best practices promoted in university classrooms are actually implemented in K–12 classrooms. This issue also points to the critical importance of assessing actual teaching practices and their direct impact on student learning. With the exception of RTOP (see Appendix), validation of other instruments did not involve evidence of teacher practices for predicting student learning outcomes.

There are five SMIs on nature of science. Nature of science refers to the values and assumptions inherent to science, scientific knowledge, and/or the development of scientific knowledge (Lederman 1992) or, in brief, the epistemology of science as distinct from science process and content (Lederman et al. 1998). All the instruments in this section of the Appendix deal with nature of science with the exception of the subscale in VASS that deals with beliefs about learning science. Many of these instruments also adopt a Likert scale or rating scale that is often accompanied by some kind of scoring (such as scores 1–5 for Strongly Agree to Strongly Disagree). Two potential problems are associated with this practice. One problem is that there is a lack of a clear scale to facilitate qualitative interpretation. That is, what does a higher score mean, an issue pointed out by Glen Aikenhead (1973) a long time ago. Another potential problem is bias or privilege assigned to a particular version of nature of science. This problem is pointed out by Lederman et al. (1998) in their review of measurement instruments of nature of science, which still applies today. Because there is no universally agreed-upon version of nature of science, any selected response or closed-ended response question format, including a Likert scale, is likely to force students to think in terms of one version of nature of science, and it remains unclear what students' true understandings of nature of science are. In order to address the above two problems, VOSTS adopts the no-scoring approach and VNOS adopts the interview and open-ended response question format. However, one problem with this no-scoring and open-response approach is the difficulty in establishing internal consistency reliability. As Lederman et al. (1998) pointed out, a forced response format like a Likert scale can still play a role in assessing a specific version of nature of science, but a more comprehensive and accurate assessment of students' and teachers' understandings of nature of science requires a combination of both quantitative and qualitative methods.

Developing standardized measurement instruments to assess classroom and school learning environments has been very active and productive over the past four decades (Fraser 1994, 1998). This trend has certainly been continuing since 1990 (Fraser 2007). The nine SMIs included in the Appendix represent a typical approach to establishing validity and reliability of learning environment measurement instruments based on multifaceted (i.e., content, criterion-related, and construct) and multistage processes (i.e., pilot, revision, further testing, expanded testing). One trend in developing standardized measurement instruments related to learning environments is to develop various forms of a same instrument pertaining to different constructs such as personal versus class forms, preferred versus actual form, short versus long form, and so on. Another trend is that many of the instruments have been translated by or adapted to other countries or cultures, which adds to cross-cultural validation. Indeed, "few fields of educational research can boast the existence of such a rich array of validated and robust instruments" (Fraser 2007, p. 105). This wide array of SMIs has supported many productive research programs related to learning environments (Fraser 1994, 1998).

It is common to adopt the Likert scale (Likert 1932) when developing measurement instruments related to attitudes, learning environments, teacher beliefs and practices, and nature of science. The Likert scale is a "softer form of data collection" (Bond and Fox 2007, p. 101) because of the subjectivity in responding to the statements. A more serious issue associated with the Likert scale is the use of a total scale score by adding individual item scores. Values such as 1–5 assigned to five choices of a statement do not have the same origin and interval unit because they are not on a ratio or interval scale. Also, different Likert scale items have different degrees of likelihood for being endorsed. The consequence of being non-interval and having varying likelihood of being endorsed is that we cannot meaningfully add individual item scores into a total score. In order to address this issue, ways of analyzing Likert scale data that are different from using total scores should be adopted. The best way currently available is to use Rasch modeling to convert raw scores into latent scores so that respondents' attitudes or beliefs can be measured on a latent scale, which was the case in the development of CARS (Siegel and Ranney 2003). Without using Rasch modeling, data analysis might have to stay at the individual item level. For example, responses to different items in an attitude scale can be represented by a profile and the difference in profiles between different groups or between two time points can be meaningfully compared. Because of the above potential issues with the Likert scale, alternatives to the Likert scale can be considered. Examples of such alternatives are the Thurston scale (Thurston 1925), Guttman scale (Guttman 1944), semantic differential (Osgood et al. 1971), and checklist.

Although there was a major interest in developing SMIs on student cognitive reasoning (Liu 2009) during the 1960s and 1970s, only three SMIs related to cognitive reasoning were found since 1990. The current interest seems to have shifted to metacognition (e.g., Anderson and Nashon 2007). Given Rosalind Driver and Jack Easley's (1978) seminal review summarizing the limitations of Piagetian content-free logical reasoning in explaining students' understanding in science, there has been less interest in measuring students' content-free cognitive reasoning during the 1990s and 2000s. However, there is currently a demand for the development of measurement instruments that reflect both the domain-specific and development-dependent nature of children's concept development. The development of WPSPI and IPSPI (see Appendix; Shin et al. 2003) in astronomy is consistent with this demand.

Desirable Future Directions for Developing Standardized Measurement Instruments

Developing SMIs involves three components: observation, interpretation, and cognition (NRC 2001). Observation refers to measurement tasks through which a construct is probed; interpretation refers to measurement models through which the measurement data are interpreted; and cognition refers to theories about the construct. Significant advances in all three components have taken place over the years as reviewed in this handbook. For example, new theories on student learning progression (e.g., NRC 2007a) probably will create a demand for SMIs for measuring student long-term concept development. One example of this type of instruments for measuring students' long-term concept development is PUM (Progression of Understanding Matter; Liu 2007). In terms of measurement task formats, standardized measurement instruments reviewed in this chapter have almost exclusively relied on the paper-and-pencil format. With today's technology capability, observations for measurement instruments can now be in multimedia formats or in computer modeling. In addition, many advanced measurement models are now available and already being applied in the testing industry (NRC 2001). Development of a new generation of measurement instruments in science education should take full advantage of advances in all the above three areas.

In today's context of worldwide standards-based science education reforms, there is a demand for a coherent system of assessment in which testing using standardized measurement instruments plays an important role (NRC 2007b). A coherent system of standards-based science assessment needs to be demonstrated in multiple dimensions: horizontally among various curriculum, instruction and assessment forms, vertically among different grade levels (e.g., K–12) and educational organizations (e.g., classroom, school, school district, state/provincial), and developmentally (e.g., cognitive, affective, and so on). For example, a standardized measurement instrument can be developed for both formative and summative purposes or for both classroom and large-scale state/provincial assessments. New measurement models and techniques (NRC 2001) have made it possible for students of different populations, or the same group of students at different times, to be assessed and directly compared even though they answer different sets of questions of a same standardized measurements (Bond and Fox 2007).

The ultimate goal of developing a measurement instrument is to construct a meaningful measure so that quantitative comparisons can be made. Ben Wright (1999) succinctly summarized characteristics of measures to be: (1) linear, (2) on abstract units (i.e., inferences by stochastic approximations), (3) of unidimensional quantities, and (4) impervious to extraneous factors. Developing instruments that produce measures requires new approaches. Mark Wilson (2005) proposes one such approach involving four cyclic stages: (1) defining the construct and making a hypothesis, (2) designing tasks to solicit student responses, (3) defining the outcome space in which the measured construct is demonstrated, and (4) applying a measurement model to map the observed scores into latent scores (i.e., measures) and testing the hypothesis. The above process continues until no evidence is present to reject the hypothesis. Development of the majority of the instruments reviewed in this chapter followed the classical test theory, which relies on means and standard deviations of raw scores to establish validity and reliability evidence, which would not be sufficient to produce scores as measures. Developing the next generation of measurement instruments needs to involve new measurement models such as the Rasch models (Bond and Fox 2007; Wilson 2005), or other models discussed in a national research council committee report (NRC 2001). Examples of applications of Rasch models in developing measurement instruments are available in Xiufeng Liu and William Boone (2006).

Appendix

Standardized measurement instruments reported in refereed publications since 1990

Instrument	Content	Population	Validation	Reliability	Source
Conceptual understanding					
Physical Changes Concepts Test (PCCT)	Conceptual: chemistry	High school	Content, criterion-related, and construct	n/a	Haidar and Abraham (1991)
General Science Literacy	Conceptual: General	University	Content, criterion-related	KR–20	Cannon and Jinks (1992)
Test of Understanding Graphs in Kinematics (TUG–K)	Conceptual: Physics	High school to university	Content, construct	KR–20	Beichner (1994)
Force Concept Inventory (FCI)	Conceptual: Physics	9th grade to university	Content, construct	n/a	Hestenes et al. (1992) and Hestenes and Halloun (1995)
Diffusion and Osmosis Test (DOT)	Conceptual: Biology	University	Construct	Split-half internal	Odom and Barrow (1995)
Force and Motion Conceptual Evaluation (FMCE)	Conceptual: Physics	University	Content, construct	n/a	Thornton and Sokoloff (1998)
Test to Identify Student Conceptualizations (TISC)	Conceptual: Physics	University	Content, construct	KR–20	Voska and Keikkinen (2000)
Conceptual Survey of Electricity and Magnetism (CSEM)	Conceptual: Physics	College	Content, construct	KR–20	Maloney et al. (2001)
Conceptual Inventory of Natural Selection (CINS)	Conceptual: Biology	University	Content, criterion-related, construct	KR–20	Anderson et al. (2002)
Chemistry Concept Inventory (CCI)	Conceptual: Chemistry	College	Content, criterion-related, construct	Cronbach's alpha	Mulford and Robinson (2002)

(continued)

(continued)

Instrument	Content	Population	Validation	Reliability	Source
Conceptual understanding					
Testing Students' Use of the Particulate Theory (TSUPT)	Conceptual: Chemistry	University	Content, construct	Inter-rater	Williamson et al. (2004)
Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT)	Conceptual: Physics	High school to university	Content, construct	KR–20	Engelhardt and Beichner (2004)
Brief Electricity and Magnetism Assessment (BEMA)	Conceptual: Physics	College	Content, construct	KR–20	Ding et al. (2006)
Geoscience Concept Inventory (GCI)	Conceptual: Earth science	College	Construct	Rasch index	Libarkin and Anderson (2006)
Progression of Understanding Matter (PUM)	Conceptual: Chemistry	Grades 3–12	Construct	Rasch index	Liu (2007)
Attitudes					
Attitude to Science Instrument (ASI) (Short Version)	Science	Elementary school (Grs. 5–6)	Concurrent	Cronbach's alpha	Caleon and Subramaniam (2008)
Attitudes toward Science Inventory (ATSI)	Science	College	Construct	Construct	Gogolin and Swartz (1992)
Attitude toward Science Questionnaire (ASQ)	Science	Upper, middle, and lower high school	Construct	Cronbach's alpha	Parkinson et al. (1998)
Secondary School Students' Attitude toward Science	Science	Secondary school	Content, criterion-related, construct	Cronbach's alpha	Francis and Greer (1999)
Attitude toward Science	Science	Elementary school	Criterion-related	Cronbach's alpha	Pell and Jarvis (2001)
Attitude Scale (AS)	Science	Junior high school	Construct	Split-half	Kesamang and Taiwo (2002)
Chemistry Attitudes and Experiences Questionnaire (CAEQ)	Chemistry	First year university	Content, criterion-related, construct	Cronbach's alpha	Dalgety et al. (2003)

(continued)

(continued)

Instrument	Content	Population	Validation	Reliability	Source
Attitudes					
Changes in Attitudes about the Relevance of Science (CARS)	Affective: Attitude	Middle and high school	Construct	Rasch index, Cronbach's alpha	Siegel and Ranney (2003)
Attitude toward Critical School Science Activity (ATCSSA) and Attitude toward Progressive School Science Activity (ATPSSA)	Affective: Attitude	Middle school	Construct	Inter-rater, Cronbach's alpha	Zacharia and Calabrese Barton (2004)
Colorado Learning Attitude about Science Survey (CLASS)	Affective: Attitude	High school and college physics	Construct	Test-retest	Adams et al. (2006)
Attitude toward Science Measures (ATSM)	Affective: Attitude	Secondary school	Content, construct	Cronbach's alpha	Kind et al. (2007)
Cognitive reasoning					
A Test of Scientific Creativity	Cognitive: Creativity	Secondary school	Content, construct	Cronbach's alpha, inter-rater	Hu and Adey (2002)
Well-Structured Problem-Solving Process Inventory (WPSPI) and Ill-Structured Problem-Solving Process Inventory (IPSPI)	Cognitive: Problem-solving	High school	Content, construct	Inter-rater	Shin et al. (2003)
Metacognition Baseline Questionnaire (MBQ)	Metacognition	High school	Content, criterion-related, construct	Crobach's alpha	Anderson and Nashon (2007)

(continued)

(continued)

Instrument	Content	Population	Validation	Reliability	Source
Nature of science					
Views on Science–Technology–Society (VOSTS)	Nature of science	High school	Content, construct	n/a	Aikenhead and Ryan (1992)
Views about Sciences Survey (VASS)	Nature of science	High school and college	Content, construct	n/a	Halloun and Hestenes (1998)
Views of Nature of Science Questionnaire Form B and Form C (VNOS–B and VNOS–C)	Nature of science	Preservice and in-service science teachers	Content, construct	Inter-rater	Lederman et al. (2002)
Thinking about Science Instrument (TSI)	Nature of science	Preservice elementary teachers	Content, construct	Cronbach's alpha	Cobern and Loving (2002)
Views on Science and Education Questionnaire (VOSE)	Nature of science	Preservice science teacher	Content, construct	Cronbach's alpha	Chen (2006)
Learning environments					
Science Laboratory Environment Inventory (SLEI)	Learning environment: laboratory setting	High school and university teachers	Content, criterion-related, construct	Cronbach's alpha	Fraser et al. (1993)
Questionnaire on Teacher Interaction (QTI)	Learning environment: Teacher–student relationship	Elementary to high school	Content, criterion-related, construct	Cronbach's alpha	Wubbels et al. (1991, 1993)
Constructivist Learning Environment Survey (CLES)	Learning environment: Constructivist	Elementary to high school	Content, criterion-related, construct	Cronbach's alpha	Taylor et al. (1997)
Cultural Learning Environment Questionnaire (CLEQ)	Culturally sensitive classroom instruction	Secondary school	Content, criterion-related, construct	Cronbach's alpha	Fisher and Waldrip (1997)
What Is Happening In this Class? (WIHIC)	Learning environment: Comprehensive	Elementary to high school to university	Content, criterion-related, construct	Cronbach's alpha	Aldridge et al. (1999)

(continued)

(continued)

Instrument	Content	Population	Validation	Reliability	Source
Learning environments					
Learning Environment Scales (LES)	Teacher goals and climate of cooperation	High school	Content, criterion-related, construct	Cronbach's alpha	Nolen (2003)
Outcome-Based Learning Environment Questionnaire (OBLEQ)	Outcome-based learning	Secondary school	Content, criterion-related, construct	Cronbach's alpha	Aldridge et al. (2006)
Science Teacher School Environment Questionnaire (STSEQ)	School culture	Secondary school	Content, criterion-related, construct	Cronbach's alpha	Huang (2006)
Students' Perception of Assessment Questionnaire (SPAQ)	Classroom assessment	Secondary school	Content, criterion-related, construct	Cronbach's alpha	Dhindsa et al. (2007)
Teacher beliefs and practices					
Science Teacher Self-efficacy Instrument	Teacher Beliefs and practices: Efficacy	Preservice elementary science teachers	Content, construct	Cronbach's alpha	Czerniak and Schriver (1994)
Attitudes toward Teaching of Environmental Risk (ATER)	Attitude	Science teachers	Construct	Construct	Zint (2002)
The Attitudes and Beliefs about the Nature and the Teaching of Mathematics and Science	Teacher beliefs and practices	Preservice teachers	Content, construct	Cronbach's alpha	McGinnis et al. (2002)
Teacher Perceptions and Practices Regarding the Use of the History of Science in their Classrooms	Teacher beliefs and practices	Elementary and secondary science teachers	Content	Cronbach's alpha	Wang and Marsh (2002)

(continued)

(continued)

Instrument	Content	Population	Validation	Reliability	Source
Teacher beliefs and practices					
Reformed Teaching Observation Protocol (RTOP)	Teacher beliefs and practices	Science teachers	Criterion-related, construct	n/a	Admson et al. (2003)
Survey of Instructional and Assessment Strategies (SIAS)	Teacher beliefs and practices	College teachers	Content, construct	Cronbach's alpha	Walczyk and Ramsey (2003)
Science Lesson Plan Analysis Instrument (SLPAI)	Lesson planning	Elementary and secondary	Content, criterion-related	Inter-rater reliability	Jacobs et al. (2008)

References

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics – Physics Education Research*, *2*, 010101.

Admson, S. L., Banks, D., Burtch, M., Cox, F., Judson, E., Turley, J. B., et al. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching*, *40*, 939–957.

Aikenhead, G. (1973). The measurement of high school students' knowledge about science and scientists. *Science Education*, *57*, 539–549.

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument: Views on science–technology–society (VOSTS). *Science Education*, *76*, 477–491.

Aldridge, J. M., Fraser, B. J., & Huang, T.–C. I. (1999). Investigating classroom environments in Taiwan and Australia with multiple research methods. *Journal of Educational Research*, *93*, 48–62.

Aldridge, J. M., Laugksch, R. C., Seopa, M. A., & Fraser, B. J. (2006). Development and validation of an instrument to monitor the implementation of outcomes-based learning environments in science classrooms in South Africa. *International Journal of Science Education*, *28*, 45–70.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in science Teaching*, *39*, 952–978.

Anderson, D., & Nashon, S. (2007). Predators of knowledge construction: Interpreting students' metacognition in an amusement park physics program. *Science Education*, *91*, 298–320.

Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics*, *62*, 750–762.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Britton, E. D., & Schneider, S. A. (2007). Large-scale assessments in science education. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 1007–1040). Mahwah, NJ: Lawrence Erlbaum.

Caleon, I., & Subramaniam, R. (2008). Attitudes towards science of intellectually gifted and mainstream upper primary students in Singapore. *Journal of Research in Science Teaching*, *45*, 940–954.

Cannon, J. R., & Jinks, J. (1992). A cultural literacy approach to assessing general science literacy. *School Science and Mathematics*, *92*, 196–200.

Chen, S. (2006). Development of an instrument to assess views on nature of science and attitudes toward teaching science. *Science Education*, *90*, 803–819.

Cobern, W. W., & Loving, C. C. (2002). Investigation of preservice elementary teachers' thinking about science. *Journal of Research in Science Teaching*, *39*, 1016–1031.

Czerniak, C., & Schriver, M. (1994). An examination of preservice science teachers' beliefs and behaviours as related to self-efficacy. *Journal of Science Teacher Education*, *5*, 77–86.

Dalgety, J., Coll, R. K., & Jones, A. (2003). Development of chemistry attitudes and experiences questionnaire (CAEQ). *Journal of Research in Science Teaching*, *40*, 649–668.

Dhindsa, H. S., Omar, K., & Waldrip, B. (2007). Upper secondary Bruneian science students' perceptions of assessment. *International Journal of Science Education*, *29*, 1261–1280.

Ding, L, Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism tool: Bried electricity and magnetism assessment. *Physical Review Special Topics – Physics Education Research*, *2*, 010105.

Doran, R. L, Lawrenz, F., & Helgeson, S. (1994). Research on assessment in science. In D. L. Gabel (Ed.), *Handbook of Research on science teaching and learning* (pp. 388–442). New York: Macmillan Publishing Company.

Driver, R., & Easley, J., Jr. (1978). Pupils and paradigms: A review of the literature related to concept development in adolescent science students. *Studies in Science Education*, *5*, 61–84.

Engelhardt, P. V., & Beichner, R. J. (2004). Students' understanding of direct current resistive electric circuits. *American Journal of Physics*, *72*, 98–115.

Fisher, D. L., & Waldrip, B. G. (1997). Assessing culturally sensitive factors in the learning environment of science classrooms. *Research in Science Education*, *27*, 41–49.

Francis, L. J., & Greer, J. E. (1999). Measuring attitudes toward science among secondary school students: The affective domain. *Research in Science & Technological Education*, *17*, 219–226.

Fraser, B. J. (1994). Research on classroom and school climate. In D. L. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 493–541). New York: Macmillan.

Fraser, B. J. (1998). Science learning environment: Assessment, effects and determinants. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 527–564). Dordrecht, The Netherlands: Kluwer Academic.

Fraser, B. J. (2007). Classroom learning environments. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 103–124). Mahwah, NJ: Lawrence Erlbaum.

Fraser, B. J., McRobbie, C. J., & Giddings, G. J. (1993). Development and cross-national validation of a laboratory classroom environment instrument for senior high school science. *Science Education*, *77*, 1–24.

Gogolin, L., & Swartz, F. (1992). A quantitative and qualitative inquiry into the attitudes toward science of nonscience college majors. *Journal of Research in Science Teaching*, *29*, 487–504.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Haidar, A. H., & Abraham, M. R. (1991). A comparison of applied and theoretical knowledge of concepts based on the particulate nature of matter. *Journal of Research in Science Teaching*, *28*, 919–938.

Halloun, I., & Hestenes, D. (1998). Interpreting VASS dimensions and profiles for physics students. *Science & Education*, *7*, 533–577.

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to Huffman and Heller. *The Physics Teacher*, *33*, 502–506.

Hestenes, D., Wells, M., & Swackmaher, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141–158.

Hu, W., & Adey, P. (2002). A scientific creativity test for secondary school students. *International Journal of Science Education*, *24*(4), 389–403.

Huang, S. L. (2006). An assessment of science teachers' perceptions of secondary school environments in Taiwan. *International Journal of Science Education*, *8*(1), 25–44.

Jacobs, C. L., Martin, S. N., & Otieno, T. C. (2008). A science lesson plan analysis instrument for formative and summative program evaluation of a teacher education program. *Science Education*, *92*(6), 1096–1126.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kesamang, M. E. E., & Taiwo, A. A. (2002). The correlates of the socio-cultural background of Botswana junior secondary school students with their attitudes towards and achievements in science. *International Journal of Science Education*, *24*(9), 919–940.

Kind, P., Jones, K., Barmby, P. (2007). Developing attitudes toward science measures. *International Journal of Science Education*, *29*, 871–893.

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, *29*, 331–359.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, *39*, 497–521.

Lederman, N. G., Wade, P., & Bell, R. (1998). Assessing understanding of the nature of science: A historical perspective. In W. F. McComas (Ed.), *The nature of science in science education* (pp. 331–350). Dordrecht, The Netherlands: Kluwer Academic.

Libarkin, J. C., & Anderson, S. W. (2006). The geoscience concept inventory: Application of Rasch analysis to concept inventory development in higher education. In X. Liu & W. J. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 45–73). Maple Grove, MN: JAM Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Achieves of Psychology*, *22*, 5–53.

Liu, X. (2007). Growth in students' understanding of matter during an academic year and from elementary through high school. *Journal of Chemical Education*, *84*, 1853–1856.

Liu, X. (2009). Standardized measurement instruments in science education. In W.-M. Roth & K. Tobin (Eds.), *The world of science education: Handbook of research in North America* (pp. 649–677). Rotterdam, The Netherlands: Sense.

Liu, X., & Boone, B. J. (Eds.). (2006). *Applications of Rasch measurement in science education*. Maple Grove, MN: JAM Press.

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *Physics Education Review*, *69*(7), S12–S23.

McGinnis, J. R., Kramer, S., Shama, G., Graeber, A. O., Parker, C. A., & Watanabe, T. (2002). *Journal of Research in Science Teaching*, *39*, 713–737.

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, *79*, 739–744.

National Research Council (NRC). (2001). *Knowing what students know*. Washington, DC: National Academic Press.

National Research Council (NRC). (2002). *Scientific research in education*. Washington, DC: National Academic Press.

National Research Council (NRC). (2007a). *Taking science to school: Learning and teaching science in grades K–8*. Washington, DC: National Academic Press.

National Research Council (NRC). (2007b). *Systems for state science assessment*. Washington, DC: National Academic Press.

Nolen, S. B. (2003). Learning environment, motivation, and achievement in high school science. *Journal of Research in Science Teaching*, *40*, 347–368.

Odom, A. L., & Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. *Journal of Research in Science Teaching*, *32*, 45–61.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1971). *The measurement of meaning*. Chicago: University of Illinois Press.

Parkinson, J., Hendley, D., Tanner, H., & Stables, A. (1998). Pupils' attitudes to science in key stage 3 of the national curriculum: A study of pupils in South Wales. *Research in Science and Technological Education*, *16*, 165–177.
Pell, T., & Jarvis, T. (2001). Developing attitude to science scales for use with children of ages five to eleven years. *International Journal of Science Education*, *23*, 847–862.
Shin, N., Jonassen, D. H., & McGee, S. (2003). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of Research in Science Teaching*, *40*, 6–33.
Siegel, M. A., & Ranney, M. A. (2003). Developing the changes in attitude about the relevance of science (CARS) questionnaire and assessing two high school science classes. *Journal of Research in Science Teaching*, *40*, 757–775.
Spies, R. A., & Plake B. S. (2005). (Eds.). *The sixteenth mental measurements yearbook*. Lincoln, NE: The Buros Institute of Mental Measurement, University of Nebraska Press.
Tamir, P. (1998). Assessment and evaluation in science education: Opportunities to learn and outcomes. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 761–789). Dordrecht, The Netherlands: Kluwer Academic.
Taylor, P. C., Fraser, B. J., & Fisher, D. L. (1997). Monitoring constructivist classroom learning environment. *International Journal of Educational Research*, *27*, 293–302.
Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and learning curricula. *American Journal of Physics*, *66*, 338–352.
Thurston, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433–451.
Voska, K. W., & Heikkinen, H. W. (2000). Identification and analysis of student conceptions used to solve chemical equilibrium problems. *Journal of Research in Science Teaching*, *37*, 160–176.
Walczyk, J. J., & Ramsey, L. L. (2003). Use of learner-centered instruction in college science and mathematics classrooms. *Journal of Research in Science Teaching*, *40*, 566–584.
Wandersee, J. H., Mintzes, J., & Novak, J. (1994). Research on alternative conceptions in science. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 177–210). New York: Macmillan.
Wang, H. A., & Marsh, D. D. (2002). Science instruction with a humanistic twist: Teachers' perception and practice in using the history of science in their classrooms. *Science & Education*, *11*, 169–189.
Williamson, V., Huffman, J., & Peck, L. (2004). Testing students' use of the particulate theory. *Journal of Chemical Education*, *81*, 891–896.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Hillsdale, NJ: Lawrence Erlbaum.
Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know* (pp. 65–104). Hillsdale, NJ: Lawrence Erlbaum.
Wubbels, Th., Brekelmans, M., & Hooymayers, H. (1991). Interpersonal teacher behavior in the classroom. In B. J. Fraser & H. J. Walberg (Eds.), *Educational environments: Evaluation, antecedents and consequences* (pp. 141–160). London: Pergamon.
Wubbels, Th., Creton, H., Levy, J., & Hooymayers, H. (1993). The model for interpersonal teacher behavior. In Th. Wubbels & J. Levy (Eds.), *Do you know what you look like: Interpersonal relationships in education* (pp. 13–28). London: Falmer.
Zacharia, Z., & Calabrese Barton, A. C. (2004). Urban middle-school students' attitudes toward a defined science. *Science Education*, *88*, 197–222.
Zint, M. (2002). Comparing three attitude-behaviour theories for predicting science teachers' intentions. *Journal of Research in science Teaching*, *39*, 819–844.