Chapter 3
Aligning Mission and Measurement

Steven E. Stemler and Mary DePascale

3.1 Introduction

There is a deep disconnect between what K-12 schools in the USA articulate as their primary mission and how those schools are held accountable for their performance. Specifically, principals, teachers, parents, employers, and policy makers all believe that schools should be doing more than just teaching students to read, write, and do math (Stemler, Bebell, & Sonnabend, 2011). There is a strong consensus that students should also develop emotional skills and learn the skills associated with effective citizenship (Stemler & Bebell, 2012). These broader skills are viewed as fundamental to the core mission of schools and, in some cases, hold an even higher priority than the basic cognitive elements. Yet, the current push for accountability coming from the federal level, in the form of standardized testing (most recently the Common Core State testing), relates exclusively to cognitively oriented, domain-specific knowledge (Common Core State Standards Initiative, 2010a, 2010b).

The most common arguments we have encountered for dismissing the idea of assessing broader skills are as follows: (1) they are nice “add-ons” to a student’s education, but are not fundamental to it; (2) these skills cannot be assessed because they are too subjective. Our primary goal in this chapter will be to address these two arguments directly.

In order to ensure that measurement is aligned with mission, we must first systematically examine the mission of schooling. Therefore, this chapter begins with a review of the literature related to school purpose. We draw on various sources of data (e.g., surveys, legal documents, school mission statements) from a diverse

S.E. Stemler (✉) • M. DePascale
Department of Psychology, Wesleyan University,
207 High Street, Middletown, CT 06437, USA
e-mail: steven.stemler@wesleyan.edu; mdepascale@wesleyan.edu

A.A. Lipnevich et al. (eds.), *Psychosocial Skills and School Systems in the 21st Century*, The Springer Series on Human Exceptionality,
DOI 10.1007/978-3-319-28606-8_3

range of constituents (e.g., courts, businesses, schools) in order to examine empirically the core competencies these constituencies expect schools to develop in students.

In the second part of the chapter, our main objective is to illustrate for the reader how different noncognitive competencies of interest have been measured in the educational and psychological literature. While many instruments that purport to measure these important noncognitive skills have been developed, those included in this chapter have demonstrated strong psychometric evidence, are typically aimed at K-12 students, and represent a diversity of approaches to measuring the core competency. We comment on the nature of the instruments (e.g., self-report vs. performance), provide sample items, highlight information related to their reliability and validity, and refer the reader to where, specifically, the complete instruments may be found. Our review is not intended to be exhaustive; rather, the main point we wish to convey is that a wide variety of psychometrically sound measures currently exist for nearly any competency one wishes to measure. In addition, new measures are being introduced into the literature constantly.

We conclude this chapter by proposing a new approach to accountability that aligns mission and measurement. We argue that because not all schools aim to achieve the same outcomes, an ideal accountability system would be one in which schools are held accountable for those objectives they aim to achieve. We outline one model for what that could look like within the current political context.

3.2 The Mission of Schools

The purpose of schooling is a topic that has been debated by philosophers, politicians, academics, legal courts, businesses, parents, and students since the inception of formal schooling. Philosophers as diverse as Aristotle, Emerson, Plato, Locke, Confucius, Dewey, Counts, and Adler have written about this topic (Noddings, 1995; Reed & Johnson, 1996), and it is a favorite topic of sociologists (deMarrais & LeCompte, 1995; Labaree, 1997) and historians alike (Goodlad, 1979; Tyack, 1988). We will now briefly review some of these perspectives.

Legal Perspectives Despite the growing federal influence in American education over the past 30 years, which has largely emphasized mainly cognitive development and vocational preparation (e.g., A Nation at Risk, 1983; Goals 2000; NCLB, 2001; Common Core State Standards Initiative, 2010a, 2010b), court rulings have largely maintained individual states' rights to self-determine the goals of their public educational systems.

Within the past 25 years, courts in states ranging from Kentucky to Massachusetts have outlined what they believe to be the purposes of schooling from a legal perspective. In 1989 (Rose v. Council for Better Education, 1989), the Kentucky State Supreme Court ordered the General Assembly to reform the property tax system and provide funding "sufficient to provide each child in Kentucky an adequate education." In defining an adequate education, the court enumerated seven learning goals that have been widely cited as precedent and subsequently adopted by other states (e.g., McDuffy v. Secretary, 1993). The seven distinct components of education outlined by the court include the development of (i) sufficient oral and written communication skills to enable a student to function in a complex and readily changing civilization; (ii) sufficient knowledge of economic, social, and political systems to enable students to make informed choices; (iii) sufficient understanding of government processes to enable the student to understand the issues that affect his or her community, state, and nation; (iv) sufficient self-knowledge and knowledge of his or her mental and physical wellness; (v) sufficient grounding in the arts to enable each student to appreciate his or her cultural and historical heritage; (vi) sufficient training or preparation for advanced training in either academic or vocational fields so as to enable each child to choose and pursue life work intelligently; and (vii) sufficient level of academic or vocational skills to enable public school students to compete favorably with their counterparts in surrounding states, in academics, or in the job market.

In recognizing the many goals of public education, the Kentucky precedent emphasizes that public schooling should not simply focus on cognitive outcomes. Specifically, the courts in Kentucky, Massachusetts, and many other states articulate an equal emphasis on a variety of student outcomes including cognitive, civic, and emotional development. As a practical consequence, states have slowly begun to require exit exams in broader domains, such as citizenship, US history, or social studies (Kemler, 2015; McIntosh, 2012), as well as mathematics, science, reading, or writing.

Business Perspectives The perspectives of citizens (Immerwahl, 2000) as well as businesses have been studied via the use of large-scale surveys. One recent large-scale study of employers conducted by the Association of American Colleges and Universities (2010) revealed that the educational outcomes considered important by employers include the ability to communicate effectively, orally, and in writing (89 % of employers surveyed); critical thinking and analytical reasoning skills (81 %); the ability to apply knowledge and skills to real-world settings through internships or other hands-on experiences (79 %); the ability to connect choices and actions to ethical decisions (75 %); the ability to analyze and solve complex problems (75 %); teamwork skills and the ability to collaborate with others in a diverse group setting (71 %); the ability to innovate and be creative (70 %); the ability to locate, organize, and evaluate information from multiple sources (68 %); the ability to work with numbers and understand statistics (63 %); an understanding of the role of the USA in the world (57 %); an appreciation for cultural diversity in America and other countries (57 %); and civic knowledge, civic participation, and community engagement (52 %).

Similarly, an evaluation of Recruiting Trends (Gardner, 2007), a publication based on information supplied by hundreds of companies and organizations concerning the recruitment of recent college graduates, reveals what specific skills employers are seeking in their recruits. In 2002–2003, ethics and integrity were considered the most important competencies. The following year, employers expressed their preference for college students to have better developed skills in communication, personal attributes (work ethic, flexibility, initiative, and motivation), teamwork, interpersonal skills, and learning (willing to learn continuously new skills and ideas). The next year, in the 2005–2006 issue, employers placed a high value on geographic awareness and a global understanding of events as they pertain to the company and industrial sector. Likewise, Casner-Lotto and Barrington (2006) surveyed 400 employers across the USA and found that the most important skills they felt were needed to succeed in the workplace included (1) professionalism/work ethic, (2) oral and written communication, (3) teamwork/collaboration, and (4) critical thinking/problem solving.

School Perspectives The study of school mission statements has emerged as a productive approach to empirically studying what schools themselves articulate as their core objectives (Bebell & Stemler, 2002; Schmitt, 2012; Stemler & Bebell, 1999, 2012; Stemler et al., 2011; Stober, 1997). A wide range of school effectiveness research has consistently shown that commitment to a shared mission statement is one of the leading factors differentiating more effective schools from less effective schools (Kurland, Peretz, & Hertz-Lazarowtiz, 2010; Renchler, 1991; Renihan, Renihan, & Waldron (1986); Rutter & Maughan, 2002; Teddlie & Reynolds, 2000). Although we would not argue that the mission is the only indicator of a school's cultural values, we do argue that it provides a straightforward and accessible indicator.

School mission statements represent a useful source of data for gaining access to the school perspective on matters of purpose for at least four reasons. First, nearly all major school accrediting bodies require a mission statement from schools seeking accreditation (AdvanceEd, 2006). Indeed, the very first standard articulated by the nation's largest secondary school accreditation body requires that:

> Schools establish and communicate a shared purpose and direction for improving the performance of students and the effectiveness of the school. In fulfillment of the standard, the school: (i) establishes a vision for the school in collaboration with its key stakeholders, (ii) communicates the vision and purpose to build stakeholder understanding and support, (iii) identifies goals to advance the vision, (iv) ensures that the school's vision and purpose guide the learning process, and (v) reviews its vision and purpose systematically and revises them when appropriate. (AdvanceEd, 2006, p. 9)

As recognized by most accrediting boards (as well as many business, civic, and private organizations in America), mission statements represent an important summation or distillation of an organizatio's core goals represented by concise and simple statements that communicate broad themes. Furthermore, school mission statements are one of the only written documents outlining purpose that nearly all schools have. The fact that nearly all American schools have a mission statement thus provides a common measure allowing for systematic comparison across diverse institutions. Second, school mission statements tend to be publicly available and easily accessible, making them well suited for study, particularly in the age of online data collection. Third, research has demonstrated that mission statements can be systematically and reliably coded by applying content analysis techniques (Bebell & Stemler, 2004; Berleur & Harvanek, 1997; Schmitt, 2012; Stemler & Bebell, 1999; Stober, 1997; Stemler et al., 2011). In 1999, Stemler and Bebell introduced a coding rubric for school mission statements, using emergent analytic coding, that allowed mission statements to be classified into major thematic categories (e.g., social development, cognitive development, emotional development, civic development, physical development) as well as 33 specific subcategories. Across a series of studies, these authors found their rating system exhibited median consensus estimates of interrater reliability ranging from 0.77 to 0.80 (Bebell & Stemler, 2004; Stemler & Bebell, 1999). Thus, a given school mission statement can be dichotomously coded on 11 independent traits, which then allows for quantitative/statistical analyses of these traits across samples of schools.

In 2012, Stemler and Bebell studied the mission statements of a wide variety of school types in the USA. Schools were drawn from public and private K-12 education, Montessori schools, charter schools, Waldorf schools, Native American schools, and vocational schools. Even across this wide variety of school types, the data suggested a general convergence on the importance of cognitive, emotional, and civic goals; however, there were notable differences by school type. For example, none of the Montessori schools sampled included any mention of any competencies related to citizenship. Further, the Waldorf schools tended to emphasize broad, emotionally based competencies and made little explicit mention of cognitive outcomes. Vocational schools tended to focus not only on job preparation but also on cognitive outcomes and spoke little of emotional or civic elements. Figure 3.1 provides four sample mission statements drawn from different school types. Note the emphasis on individual cognitive and emotional competencies in the Montessori example (and the absence of reference to the civic), the emphasis on a wide range of emotionally oriented competencies in the Waldorf school, the nearly exclusive focus on cognitive outcomes from the charter school, and the civic emphasis of the public high school.

International Perspectives Although large-scale internationally comparative studies of achievement began in the 1970s (Stemler, 2001), the policy discussions invoking the results of large-scale tests of achievement reached a fever pitch in the 1990s. Since that time, there has been growing public and political interest in comparative educational studies (e.g., Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), Program for International Student Assessment (PISA)). Although these international comparative studies historically have focused on traditional cognitive outcomes such as mathematics, science, and writing, the international community has begun to recognize the importance of broader competencies and have recently begun to incorporate them into their assessment battery. PISA 2018, for example, will require an assessment of teamwork skills, self-efficacy, and cultural competence in addition to their more traditional core areas (see Chap. 14).

Summary A wide variety of empirical data has been collected from a variety of constituents, including schools, parents, businesses, and courts. The data comes in a variety of forms (e.g., surveys, legal documents, mission statements), but it is remarkably consistent in pointing to the same conclusion. Constituents believe that

Chesterfield Montessori School; Chesterfield, Missouri

Chesterfield Montessori School offers an authentic Montessori education that honors children's individuality. Our peaceful environment and compassionate staff nurtures respect for self and others, fosters a strong sense of community, and stimulates independent thinking. Students carry with them a solid record of academic achievement, a belief in the dignity of work, and a sense of responsibility for their own development as happy and productive human beings.

The Bay School (Waldorf School); Blue Hill, Maine

The Bay School's mission is to provide an education that engages and nurtures the whole child, inspiring a balanced growth of heart, mind, body, and spirit. We are committed to developing in our students inner confidence, responsibility, self-motivation, a love of learning, imagination, creativity, and intellectual clarity. The educational ideals and values of the school, rooted in the Waldorf tradition, create a community of children, alumni, parents, and faculty imbued with reverence for others and the natural world.

*Jumoke Academy Charter School;*Hartford, Connecticut

The mission of Jumoke Academy Charter School is to prepare children to successfully compete in the global marketplace despite the social and economic challenges they may presently face. The academy is dedicated to rigorous academic and social standards achieved by holding high expectations for all students during challenging instruction.

The concept of "Jumoke" is central to the academy's mission to provide a safe and nurturing environment for its children while providing high quality instruction. Students in PreK-8th grade will be offered a developmentally appropriate curriculum and an enriched program of extended day activities which addresses the unique talents and background of each child in the areas of science, mathematics, language arts, technology, physical education, music and art enrichment.

Chapin High School (Recognized by the US Dept of Ed as a Blue Ribbon Public High School); Chapin, South Carolina

The mission of Chapin High School of Lexington Richland School District Five, in partnership with the community, is to provide challenging curricula with high expectations for learning that develop productive citizens who can solve problems and contribute to a global society.

Fig. 3.1 Example of school mission statements

schools exist to develop a wide variety of cognitive and noncognitive competencies. Furthermore, there is no evidence to suggest that cognitive outcomes are of greater importance to these groups or that noncognitive competencies are somehow secondary or "add-ons" to the core educational experience. To the contrary, if precedence exists, it seems to favor the development of noncognitive competencies. If schools value these skills, then should not these skills also be assessed to determine whether or not students are developing them as a result of their schooling? We now turn to the question of how to measure these important broader skills and competencies.

3.3 Measuring Broader Skills and Competencies

One reason that cognitive skills may be perceived as taking precedence over broader, noncognitive measures is because the measurement of cognitive skills has a long history in the literature, dating back over 100 years with the first standardized IQ tests (Birney & Stemler, 2007; Ciancialo & Sternberg, 2004). In addition, it is cognitive skills, and those skills alone, that are emphasized within the current high-stakes testing policies in the USA associated with federal educational policies such as Race to the Top (http://www2.ed.gov/programs/racetothetop/index.html) and the Common Core State Standards (2010a, 2010b). The reason that cognitive skills such as critical thinking are given precedence in federal policies is likely attributable to the perception that noncognitive skills are subjective and cannot be assessed as reliably as cognitive skills such as critical thinking. However, there has been substantial research in the past half century devoted to the measurement of civic, emotional, and personal skills (see Chap. 2) The argument that there is no way to measure these broader competencies is simply a canard.

The remainder of this chapter will be devoted to debunking this fallacy by citing assessment efforts associated with some of the core skills and competencies schools say they aim to develop. Given the large number of potential competencies that could be included and the limitations of space associated with this chapter, we have chosen to focus on the two core elements that Stemler and Bebell (Stemler et al., 2011; Stemler & Bebell, 2012) have shown to have equal status to the cognitive domain: those of emotional development and citizenship. Note that each of these domains represents a broad category that itself consists of many other competencies. Each of the various competencies has a substantial literature associated with it, and we fully recognize that our efforts here are merely scratching the surface. Furthermore, specialists in these areas may well dispute our grouping of the competencies as falling under the "Emotional" or "Civic" umbrella. We admit that these are rough categorizations; however, we will generally follow the rubric set forth by Stemler et al. (2011) which specifies subdomains associated with each broad category. This rubric has been demonstrated to have strong interrater reliability across a variety of studies. The instruments we have chosen to include in our review were purposefully selected to meet two or more of the following three criteria: (1) they possess strong psychometric properties; (2) they exemplify a variety of different assessment techniques (e.g., self-report, ability test, observations); and (3) they are relevant to a K-12 audience. The instruments we highlight are not the only instruments that potentially meet these criteria, but they do suffice to illustrate our broader point that objective, quantitative, and psychometrically sound measures exist for nearly any competency one wishes to consider.

3.4 Emotional Development

Emotional development was the second most frequently cited theme to emerge from Stemler et al.'s (2011) analysis of a true random sample of 500 US high school mission statements across ten geographically and politically diverse states. Fully 55 %

of the 421 randomly selected public high schools mentioned emotional development, as compared to 58 % who mentioned citizenship and 53 % who mentioned cognitive development. Further, in a subsequent study, Stemler and Bebell (2012) found that across a nonrandom sample of 111 schools from ten different school types (e.g., public, parochial, Montessori, charter, Waldorf, Native American), emotional development was the most frequently cited theme (66 %). Consequently, we being with a review of competencies associated with the domain of emotional development. According to the rubric outlined in the studies by Stemler and colleagues, these include (1) emotional intelligence; (2) empathy; (3) self-esteem, self-confidence, and self-efficacy; (4) motivation; and (5) self-directed learning. At the end of this section, we present a table that summarizes the competency under investigation, the name of the instruments used to assess the competency, example items, their psychometric properties, and information on how to obtain the instrument.

Emotional Intelligence There are two main approaches to measuring emotional intelligence that pervade the educational and psychological literature (Zeidner, Matthews, & Roberts, 2009; see also Chap. 11). One approach is based on self-report assessments of emotional intelligence and follows in the tradition of personality assessment. Perhaps the most popular instrument in this tradition is the Bar-On EQ-i. The instrument contains 133 items, each of which is rated on a 5-point Likert scale ("very seldom or not true of me," "seldom true of me," "sometimes true of me," "often true of me," "very often true of me or true of me"). Prompts ask questions similar to the following: "I can tell when someone is angry," "I know how to get someone into a good mood," and "When I need to get a job done, I can get myself into the right emotional state to get it done." This instrument has been used widely and normed on a large database of participants. The instrument has shown solid levels of internal consistency reliability, ranging from 0.69 to 0.86, and test–retest reliability coefficients ranging from 0.55 to 0.82 across the published technical manuals (Bar-on, 1997). There is a youth version of the EQ-i that is suitable for use with elementary, middle, and high school students. It has been normed against children in North America, and norms are provided separately for boys and girls across four age groups. The EQ-i youth form has been shown to be a significant predictor of academic achievement (Parker et al., 2004)

A second major approach to the measurement of emotional intelligence is through the use of performance-based assessments. In this domain, the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) is the most prominent test in use. In contrast to the Bar-On, this test consists of four dimensions: (1) perceiving emotions, (2) using emotions to facilitate thought, (3) understanding emotions, and (4) managing emotions. Rather than being a self-report measure, the MSCEIT is viewed as an ability test, in which there are correct and incorrect responses to each of the items. The items vary in type of presentation. In one subtest, participants are shown a face and asked to rate the extent to which different emotions are present or absent from the facial expression. In another test, individuals are presented with a particular scenario (e.g., meeting the parents of your new love interest for the first time) and asked what emotions would be most useful in that context. The instrument has shown very high levels of internal consistency reliability (alpha = 0.91 full scale) as well as strong test–retest reliability (alpha = 0.86 full scale) in the published technical manuals (Mayer, Salovey, & Caruso, 2002).

Other tests measure components of emotional intelligence, such as emotional management. MacCann, Fogarty, Zeidner, and Roberts (2011) have developed a Situational Test of Emotional Management for Youth (STEM-Y) aimed at assessing this particular element of emotional intelligence. An example item from the test is: "You and James sometimes help each other with homework. After you help James on a difficult project, the teacher is very critical of this work. James blames you for his bad grade. You respond that James should be grateful, because you were doing him a favor. What would you do in this situation? (a) Tell him from now on he has to do his own homework. (b) Apologize to him. (c) Tell him 'I am happy to help, but you are responsible for what you turn in.' (d) Don't talk to him." They administered the test to 383 eighth grade students and found significant positive associations with GPA and problem-focused coping.

In addition, the Trait Emotional Intelligence Questionnaire (TEIQue) is often used as a measure of emotional intelligence. It is available in both a child form (TEIQue-CF) and an adolescent short form (TEIQue-ASF). The items on this self-report measure include questions such as "I can tell when a friend is sad" which are answered using a Likert scale. The internal consistency reliability of the test is strong, at 0.84. Further, Mavroveli and Sanchez-Ruiz (2011) found that children who scored higher on trait EI were rated by their peers as significantly more socially competent.

Another measure that has been used is the "Guess Who" peer assessment technique. This technique involves giving students descriptions of certain behaviors that other students may exhibit. Students then list other students' names for whom they think the descriptions are appropriate. Teachers may also provide responses. Scales used in the technique may include categories like "cooperative," "disruptive," "shy," "aggressive," "dependent," "a leader," "intimidating," and "is kind" and "is a bully" (Mavroveli, Petrides, Sangareau, & Furnham, 2009). Students rated as "kind" on this technique showed significantly higher levels of emotional intelligence, whereas students rated as "bullies" showed significantly lower levels of trait EI (Mavroveli et al., 2009)

Empathy There are a variety of interesting methods for measuring empathy found in the empirical literature. One common measure of empathy is the Bryant Index of Empathy Measurement for children and adolescents. This index asks students 22 questions about their emotional state in certain situations, to which they may respond yes or no (Leontopoulou, 2010). The index has also been adapted so that responses are made on a 1–5 Likert scale of "I strongly agree" to "I strongly disagree" (Woods, Wolke, Nowicki, & Hall, 2009). Leontopoulou reported a Cronbach's alpha of 0.52 for the Index, and Woods et al. reported a Cronbach's alpha of 0.70 for the Index.

For younger students, empathy can also be measured with the Southampton Test of Empathy for Preschoolers (STEP). STEP is a test that involves showing students video clips of emotional situations and asking students how both they and the people in the video would respond to the situation emotionally. Questions take the format "How does [character's name] feel? How did you feel when [character experienced event]?". Students use images of emotion faces to answer these questions (Howe, Pitten Cate, Brown, & Hadwin, 2008). Howe et al. found a Cronbach's alpha of 0.70 for questions related to understanding how the people in the video would feel (STEP-UND) and 0.86 for sharing an emotional experience with the people in the video (STEP-SHA). In addition, STEP scores were positively correlated with parent evaluations of children's empathy and teacher evaluations of students' prosocial behavior.

Another approach to measuring empathy is based on the self-report methodology. A good example of this approach is the Basic Empathy Scale (BES; Jolliffe & Farrington, 2006). This 20-item scale assesses both cognitive and affective empathy and was designed to measure the degree to which a person understands and shares the emotions of another. Cognitive empathy is tapped by items such as "It is hard for me to understand when my friends are sad," whereas affective empathy is tapped by items such as "I usually feel calm when other people are scared." The BES has been shown to have strong internal consistency reliability (alpha = 0.87 full scale), and confirmatory factor analyses have shown a good fit of the two-factor model. The BES has demonstrated validity by showing positive associations with prosocial behavior and negative correlations with bullying.

Another popular self-report measure related to empathy is the Interpersonal Reactivity Index (IRI) developed by Davis (1983). The IRI includes 28 items that are intended to tap four domains: (1) perspective-taking, (2) fantasy, (3) empathetic concern, and (4) personal distress. Examples of items are "Before criticizing somebody, I try to imagine how I would feel if I were in their place" and "When I see someone being taken advantage of, I feel kind of protective toward them." Participants are to rate each item on a 5-point scale ranging from "Does not describe me well" to "Describes me very well." The instrument has demonstrated strong internal consistency reliability (alpha = 0.80 and higher) as well as strong test–retest reliability (Batanova & Loukas, 2013). Mestre, Frias, and Samper (2004) replicated the four-factor structure of the instrument, thereby providing further evidence in support of its construct validity.

A fourth approach to the measurement of empathy comes from the performance-based tradition. In that regard, the multifaceted empathy test (MET) is exemplary (Dziobek et al., 2008). In this test, 40 photographs depicting different people in positive or negative emotional situations are presented. Cognitive empathy is assessed by asking participants to choose one of four adjectives that best describes the emotional state of the depicted person. For the explicit assessment of affective empathy, participants are asked to rate how strongly they feel for the person on a 9-point rating scale (1 "not at all" and 9 "very strongly"). Participants are also asked to rate on the same 9-point rating scale how strongly they feel affected by the presented photograph in order to assess implicit affective empathy. The MET has strong internal consistency reliability, with the Cronbach's alpha ranging between 0.71 for the cognitive and 0.91 for the explicit emotional empathy scale and 0.92 for the implicit emotional empathy scale. The MET has demonstrated evidence of construct validity by discriminating between healthy individuals and patients with Asperger's autism (Dziobek et al., 2008). Evidence for the convergent validity of the instrument comes from the fact that scores on the MET show significant correlations with the IRI.

In an approach similar to the MET, Rae Westbury and Neumann (2008) developed a video-based assessment in which they measured empathy physiologically. Participants were shown 10 s video vignettes of five animal groups (humans, primates, companion mammals, utilitarian mammals, and chickens) in distressing situations. Physiological recording sensors for corrugator EMG, skin conductance, and respiration were attached to each participant. After each film clip, participants were prompted on the screen with the written instruction "Please make rating now" to rate their level of empathic feeling using a scale of 0–9 (where 0 = none and 9 = maximal response) using a computer keyboard. Participants were instructed to separate any feelings of disgust from their ratings. Their approach found strong convergent validity with the BEES (Mehrabian, 1996), another psychometrically validated measure of the affective element of empathy.

Finally, as a behavioral measure of empathy, one can measure daily helping. For example, Rameson, Morelli, and Lieberman (2012) used an assessment called the Daily Experience Survey. They asked participants in their study to complete an end-of-day online survey for 14 consecutive days. Two forms of daily helping were measured: stranger–acquaintance helping (e.g., picking up dropped objects and holding a door open; $\alpha = 0.82$) and friend helping (e.g., lending money and giving a ride; $\alpha = 0.73$).

Self-esteem By far the most frequently used measure of self-esteem is Rosenberg's (1965) Self-Esteem Scale. This scale measures general self-esteem and includes ten items capturing a continuum of self-worth statements. The scale has been used extensively in samples with a variety of ages, nationalities, and socioeconomic levels. The internal consistency reliability is strong across studies (alphas range from 0.72 to 0.87). Examples of items are "I feel that I am a person of worth, at least on an equal plane with others" and "I wish I could have more respect for myself" with the latter item being reverse scored.

The Coopersmith Self-Esteem Inventory is widely used to measure self-esteem and is relevant for K-12 students. It has 50 yes/no questions as items, which relate to global self-esteem and relationships with parents and friends. The inventory includes questions like "I am a lot of fun to be with," "I have a low opinion of myself," "I often wish I were someone else," and "Kids usually follow my ideas." Hills, Francis, and Jennings (2011) found that it could be reduced to 19 items while simultaneously yielding improved psychometric properties. Specifically, the internal consistency reliability of the shortened scale is strong (Cronbach's alpha = 0.86), and confirmatory factor analyses validated the three key sources of self-esteem (personal self-esteem, self-esteem derived from peers, and self-esteem derived from parents).

The Piers-Harris Children's Self-Concept Scale (CSCS) has also been used to measure students' self-esteem. The two CSC scales most related to self-esteem are the happiness and satisfaction scale and the popularity scale. The happiness and satisfaction scale contains 10 items, and the popularity scale contains 12 items. Rousseau, Drapeau, Lacroix, Bagilishya, and Heusch (2005) report that the Cronbach's alpha ranged from 0.57 to 0.71

Whereas the instruments just mentioned attempt to measure global (i.e., domain-general) self-esteem, there are many contexts in which it is more appropriate to measure self-confidence within a particular domain. For this purpose, one psychometrically strong assessment is the Personal Evaluation Inventory (PEI) (Shrauger & Schohm, 1995). This 54-item instrument has eight domain-specific subscales: "academic performance," "physical appearance," "athletics," "romantic relationships," "social interactions," "speaking before others," "general confidence," and "mood" state. The subscales contain between 5 and 7 items each. All items are presented as Likert scales that scored 1–4 (negative items reversed), with 4 indicating strong agreement with items reflecting self-confidence. Examples of items are "I am pleased with my physical appearance" and "I have no difficulty maintaining a satisfying romantic relationship." Evidence for the competency validity of the PEI scores comes from correlational studies showing that PEI scores are significantly correlated in expected directions with other independent measures of anxiety, hopelessness, depression, and optimism (Shrauger & Schohm, 1995). The Cronbach's alpha for the PEI instrument in the original development of this questionnaire ranged from 0.77 for the academic subscale to 0.91 for the athletics subscale. One-month test–retest reliabilities ranged from 0.73 for the academic subscale to 0.90 for the athletics subscale, indicating good stability.

Motivation As with self-esteem, motivation can be thought of and measured in ways that are domain general or ways that are domain specific (see Chap. 10). One of the most prominent measures of domain-general motivation is the Intrinsic Motivation Inventory (IMI) – a self-report inventory designed to assess the level of intrinsic motivation experienced by an individual engaged in an achievement-oriented task (Ryan, Mims, & Koestner, 1983). Six subscales measure various underlying dimensions or indices of intrinsic motivation: (1) interest/enjoyment, (2) perceived competence, (3) effort/importance, (4) pressure-tension, (5) perceived choice, and (6) value/usefulness. All items are scored on a 7-point Likert-type scale ranging from 1 (*not at all true*) to 7 (*very true*). Example items include "I enjoyed doing this activity very much" and "I tried very hard on this activity." The interesting part about the IMI is that it can be adapted to almost any activity. Because motivation is typically believed to be domain specific (i.e., related to particular activity), however, the IMI provides a framework for assessing specific activities. For example, Amorose and Horn (2001) evaluated the psychometric properties of a sport version of the IMI and found that the internal consistency reliability estimates ranged from 0.70 to 0.80. McAuley, Duncan, and Tammen (1987) showed evidence for the construct validity of a higher-order factor of intrinsic motivation with four second-order factors related to specific elements of sport.

In addition, the PISA study measures student motivation within the context of four domains: (1) perseverance (an index derived from items asking students about their willingness to continue working on difficult problems, even when they encounter problems), (2) openness to problem solving, (3) locus of control (the extent to which they see outcomes as being associated with their own effort), and (4) intrinsic vs. instrumental motivation to learn mathematics (http://www.oecd.org/pisa/keyfindings/PISA2012-Vol3-Chap3.pdf).

There are a variety of assessments used to measure student motivation at different levels of their education and in specific domains. For example, the Preschool Reading Attitude Scale (PRAS) and the Emergent Readers Motivation and Reading Scale (ERMAS) tap preschoolers' motivation for reading. The PRAS contains 34 items which students rate with a 1–3 Likert scale of emotion faces (happy, neutral, and sad) (Sperling, Sherwood, & Hood, 2013).

For middle school students, Brookhart, Walsh, and Zientarski (2006) used a collection of scales to measure motivation in social studies and science. These scales included "perceived task characteristics," "perceived self-efficacy," "mastery goal orientations," "performance goal orientations," "amount of invested mental effort," "active learning strategy use," and "superficial learning strategy use." A 1–5 Likert scale ranging from "Yes!" to "No!" was used for student responses.

An assessment of motivation for high school students is the Student Motivation and Engagement Scale-High School (MES-HS). This scale has 44 items which are rated with a 1–7 Likert scale of "strongly disagree " to "strongly agree" (Plenty & Heubeck, 2011). The assessment is typically used to assess motivation in general and, however, can be adapted to assess motivation in a particular subject area, such as math, by rewording questions to pertain to this subject. A similar questionnaire developed by Yin et al. (2008) focuses on motivation in science.

Because motivation is measured in a domain-specific way, there have been other efforts to assess it using specific instruments in the domains of reading, motivation for studying, and even friendship. Logan, Medford, and Hughes (2011) have developed a 15-item questionnaire called the Motivation for Reading Questionnaire (MRQ) that has yielded strong internal consistency (alpha = 0.75). Another interesting measure of motivation comes from Ojanen, Sijtsema, Hawley, and Little (2010). They state that their measure of intrinsic and extrinsic motivation was partly adapted from the Reasons Survey (Ryan & Connell, 1989) and the Achievement Goal Questionnaire (Elliott & Sheldon, 1997) and partly developed by their research team. As Ojanen et al. note, "All motivation items were assessed with respect to three frames: 'Why do you make new friends', 'Why do you get a friend to do something together with you', and 'Why do you keep a good friend'. The participants provided answers in a Likert-scale, ranging from 1 to 5 (1 = I disagree, 5 = I agree). Two items (with respect to the three frames, i.e., six questions altogether) were used to measure intrinsic motivation (alpha = 0.73) and four items (with respect to the three frames, i.e., 12 questions altogether) were used to measure extrinsic motivation (alpha = 0.92). An example of a frame/item combination from the intrinsic scale reads: [Frame] 'Why do you make new friends?' [item] 'Is it because you enjoy doing it?' The other intrinsic item was: 'Is it because you like to do it?' The extrinsic items were: 'Is it because you want to make your parents happy? Is it because you want to get praise from your teachers? Is it because you don't want your teachers to think you are no good?; Is it because you don't want your parents to be angry with you?'"

Self-directed Learning Historically speaking, the most widely used measure of self-directed learning is Guglielmino's Self-Directed Learning Readiness Scale (1978). The original scale consisted of 41 items that exhibited high internal consistency reliability overall (alpha = 0.81); however, the items were thought to constitute eight dimensions of self-directed learning. Guglielmino reported that principal component analysis (PCA) with varimax rotation yielded an eight-factor structure. She labeled these factors (1) openness to learning opportunities, (2) self-concept as an effective learner, (3) initiative and independence in learning, (4) informed acceptance or responsibility for one's own learning, (5) love of learning, (6) creativity, (7) future orientation, and (8) ability to use basic study skills and problem-solving skills. Response options are on a 5-point Likert scale ranging from "Almost never true of me; I hardly ever feel this way" to "Almost always true of me; there are very few times when I don't feel this way." Example items include "I know what I want to learn," "I don't work very well on my own," and "No one but me is truly responsible for what I learn."

Although the SDLRS is widely used in practice, there is controversy surrounding its psychometric properties. Specifically, advanced data analyses call into question the eight-factor model and appear to demonstrate six highly correlated factors which could be subsumed under a single higher-order factor (West & Bentley, 1990). Other authors have pointed out that the psychometric properties are inconsistent across race and class populations (Straka, 1995). As a result, other researchers (e.g., Abd-El-Fattah, 2010) have created new scales of self-directed learning aptitude. The full scale consists of 40 items and has confirmatory factor evidence for three subscales.

Another sound instrument for measuring self-directed learning is the Self-Directed Learning Scale, which consists of ten items with responses made on a 5-point Likert scale ranging from strongly disagree to strongly agree. Examples of items include "I am good at finding the right resources to help me do well in school" and "If there is something I need to learn, I find a way to do so right away." The SDLS has been found to be an internally consistent measure with Cronbach's alpha values in the mid to high 0.80s (Lounsbury, Saudargas, Gibson, & Leong, 2005). Evidence for the validity of the instrument comes from its positive association to college student life satisfaction and negative association with intention to withdraw from college.

Finally, another scale used to measure students' self-directed learning is the Self-Directed Learning with Technology Scale (SDLTS) developed by (Timothy et al., 2010). This scale includes seven items such as "I go online to ask my teachers questions on my lessons when I am not in school" and "I use the computer to work with information for my learning", and is intended for use with elementary school students. The results of the SDLTS scale were tested with 398 middle school students, 568 high school students, and 1159 college students. SDLTS scores were significantly predictive of cumulative GPA along with college and life satisfaction. It also exhibited convergent validity with Guglielmino's Self-Directed Learning Readiness Scale.

Summary The competencies chosen to represent the domain of emotional development are certainly not exhaustive, nor are the instruments discussed in this section. However, we have tried to highlight three important features. First, there exist many psychometrically sound instruments for measuring each of these competencies. Second, many of these instruments are specifically tailored to K-12 populations. And third, there are a variety of interesting methods used for assessment. The particular assessments mentioned in this section include some self-report measures, but also many other innovative approaches to measuring emotional development. Table 3.1 presents a summary that summarizes the measures discussed in this section, highlighting the competency they are intended to measure, their psychometric properties, example items, and how to acquire these instruments.

3.5 Assessing Effective Citizenship

Of the 421 randomly sampled public high school mission statements analyzed by Stemler et al. (2011), a total of 58 % of the mission statements endorsed civic development, making it the most highly ranked theme out of 11 major themes identified in their study. In the context of higher education, citizenship was emphasized in 68 % of the statements of essential capabilities of national US universities, making it the third most highly cited essential capability out of ten total capabilities (Stemler, 2012a). Thus, we next review competencies associated with the competency of citizenship. These competencies include (1) civic knowledge, (2) leadership, (3) teamwork, and (4) ethics. At the end of this section, we present a table that summarizes the competency under investigation, the name of the instruments used to assess the competency, example items, their psychometric properties, and information on how to obtain the instrument.

Civic Knowledge Without question, the most widely used and high-stakes test of citizenship is the United States Citizenship and Immigration Services (USCIS) naturalization test. Indeed, this test is a useful indication of what our government considers important in its citizens. In 2007, the assessment was revised and shifted its emphasis more toward larger concepts of American democracy and the rights and responsibilities of American citizens rather than focusing on particular historical facts. For example, a question that is formerly asked, "What country did we fight during the Revolutionary War?", has been revised to read "Why did the colonists fight the British?". Because this test represents the American government's definition of effective citizenship, its structure and content are important to

Table 3.1 Selected measures of emotional intelligence, empathy, self-esteem, motivation, and self-directed learning

Measure	Sample item/item types	Psychometrics	Availability
Emotional intelligence			
Bar-On EQ-i youth version	"When I need to get a job done, I can get myself into the right emotional state to get it done"	Solid levels of internal consistency reliability ranging from 0.69 to 0.86 and test–retest reliability coefficients ranging from 0.55 to 0.82. Shown to predict academic success	Purchase through MHS (http://www.mhs.com)
MSCEIT	Some items have participants view a face and rate the extent to which different emotions are present or absent from the facial expression. Other items are scenario based and ask participants what emotions would be most useful in the given scenarios	High levels of internal consistency reliability (alpha=0.91 full scale) as well as strong test–retest reliability (alpha=0.86 full scale) Has been shown to predict a variety of important outcomes, including job performance and academic success	Purchase through MHS (http://www.mhs.com)
STEM-Y	"You and James sometimes help each other with homework. After you help James on a difficult project, the teacher is very critical of this work. James blames you for his bad grade. You respond that James should be grateful, because you were doing him a favor. What would you do in this situation?"	Significant predictor of GPA and problem-focused coping	Contact: C. MacCann carolyn.maccann@sydney.edu.au
TEIQue-CF	"I can tell when a friend is sad"	Internal consistency of 0.84 Convergent validity evidence with peer ratings of social competence	Contact: K. V. Petrides k.petrides@ucl.ac.uk
"Guess Who" peer assessment	Students nominate peers who fit certain descriptions (e.g., cooperative, kind, a bully)	Convergent validity evidence between trait EI and ratings of kindness (positive) and bullying (negative)	Available in Mavroveli et al. (2009)
Empathy			
Bryant Index of Empathy Measurement for children and adolescents	Asks questions about their emotional state in certain situations to which they respond Yes or No. The scale has also been modified to a Likert scale (strongly agree to strongly disagree)	Reported Cronbach's alphas range from 0.52 to 0.70 for the Index	de Wied et al. (2012)

Southampton Test of Empathy for Preschoolers (STEP)	"How does [character's name] feel?" "How did you feel when [character experienced event]?"	Cronbach's alpha of 0.70 for questions related to understanding how the people in the video would feel and 0.86 for sharing an emotional experience with the people in the video STEP scores were positively correlated with parent evaluations of children's empathy and teacher evaluations of students' prosocial behavior	Howe et al. (2008)
Basic Empathy Scale	"It is hard for me to understand when my friends are sad" "I usually feel calm when other people are scared"	Cronbach's alpha of 0.86 (full scale); confirmatory factor analyses have shown a good fit of the two-factor model Results have shown convergent validity with prosocial behavior and negative correlations with bullying	Free of charge for research purposes by writing to D. Jolliffe at dj211@cam.ac.uk
Interpersonal Reactivity Index (IRI)	"Before criticizing somebody, I try to imagine how I would feel if I were in their place" "When I see someone being taken advantage of, I feel kind of protective toward them"	Strong internal consistency reliability (alpha = 0.80 and higher) as well as strong test–retest reliability	Available from the original publications (Davis, 1980, 1983)
Multifaceted Empathy Test (MET)	Participants are shown photographs depicting people in positive or negative emotional situations. Participants are asked to rate how strongly they feel for the person in the picture	Cronbach's alpha ranging between 0.71 for the cognitive and 0.91 for the explicit emotional empathy scale and 0.92 for the implicit emotional empathy scale Scores on the MET show convergent validity with the IRI	Available from the first author, I. Dziobek, at: dziobek@mpib-berlin.mpg.de
Daily Experience Survey	Stranger–acquaintance helping (e.g., picking up dropped objects and holding a door open) and friend helping (e.g., lending money and giving a ride)	Stranger–acquaintance helping Cronbach's alpha 0.82; friend helping Cronbach's alpha 0.73	Rameson, Morelli, and Lieberman (2012)
Self-esteem			
Rosenberg's (1965) Self-Esteem Scale	"I feel that I am a person of worth, at least on an equal plane with others" "I wish I could have more respect for myself"	Cronbach's alpha ranges from 0.72 to 0.87 across studies	Available from the original publication (Rosenberg, 1965) and easily found online

(continued)

Table 3.1 (continued)

Measure	Sample item/item types	Psychometrics	Availability
Coopersmith Self-Esteem Inventory	"I am a lot of fun to be with" "I have a low opinion of myself" "I often wish I were someone else" "Kids usually follow my ideas"	Cronbach's alphas of 0.86 for the reduced, 19-item scale Construct validity evidence for three-factor solution (personal self-esteem, self-esteem derived from peers, self-esteem derived from parents)	Available in Hills, Francis, and Jennings (2011) article
Piers-Harris Children's Self-Concept Scale (CSCS)	The happiness and satisfaction scale and the popularity scale are most closely aligned with self-esteem	Cronbach's alpha ranges from 0.57 to 0.71 across studies	Contact: K. V. Petrides k.petrides@ucl.ac.uk
Personal Evaluation Inventory (PEI)	"I am pleased with my physical appearance" "I have no difficulty maintaining a satisfying romantic relationship"	Cronbach's alpha in the original development of this questionnaire ranged from 0.77 for the academic subscale to 0.91 for the athletics subscale. One-month test–retest reliabilities ranged from 0.73 for the academic subscale to 0.90 for the athletics subscale PEI scores are significantly correlated in expected directions with other independent measures of anxiety, hopelessness, depression, and optimism	Available in Robinson, Shaver, and Wrightsman (1991) and by searching online
Motivation			
Intrinsic Motivation Inventory (IMI)	"I enjoyed doing this activity very much" "I tried very hard on this activity"	Internal consistency reliability estimates ranged from 0.70 to 0.80	Freely available for research purposes from: www.ravansanji.ir/?std1019IMIfull
PISA Drive and Motivation scale	"I put off difficult problems" "I like to solve complex problems" "If I put in enough effort, I could succeed in mathematics"	No validity or reliability information is reported; however, the items are derived from well-known scales	Available from PISA report at: www.oecd.org/pisa/keyfindings/PISA2012-Vol3-Chap3.pdf

Motivation for Reading Questionnaire (MRQ)	"If a book is interesting, I don't care how difficult it is to read" "I like reading so that I can learn more about things" "Time goes very slowly when I am reading"	Cronbach's alpha 0.75	Available from the Appendix of Logan, Medford, and Hughes (2011)
Ojanen, Sijtsema, Hawley, and Little's (2010) measure of intrinsic and extrinsic motivation	"Why do you make new friends?" Is it because you enjoy doing it? Is it because you like to do it? Is it because you want to make your parents happy? Is it because you want to get praise from your teachers? Is it because you don't want your teachers to think you are no good? Is it because you don't want your parents to be angry with you?	Cronbach's alphas range from 0.73 to 0.92 by scale (intrinsic vs. instrumental)	Available from Tiina Ojanen, at: tojanen@usf.edu
Self-directed learning			
Guglielmino's Self-Directed Learning Readiness Scale	"I know what I want to learn" "I don't work very well on my own" "No one but me is truly responsible for what I learn"	High internal consistency reliability overall (Cronbach's alpha 0.81). Principal component analysis with varimax rotation yielded an eight-factor structure	Available from Learning Preference Assessment, at: http://www.lpasdlrs.com
Self-Directed Learning Scale	"I am good at finding the right resources to help me do well in school" "If there is something I need to learn, I find a way to do so right away"	Cronbach's alpha values in the mid- to high 0.80s Positive association to college student life satisfaction and negative association with intention to withdraw from college	Available in Lounsbury, Levy, Park, Gibson, and Smith (2009)
Self-Directed Learning with Technology Scale (SDLTS)	"I go online to ask my teachers questions on my lessons when I am not in school" "I use the computer to work with information for my learning"		Available in Timothy et al. (2010) article

keep in mind when thinking about what civic skills should be expected of students and how best to measure them. Curiously, no psychometric information on the test is publicly available; however, the test is used for granting naturalization to citizens. Importantly, the state of Arizona has just passed legislation requiring high school students to pass the US naturalization test in order to receive a high school diploma (Rojas & Rich, 2015). Several other states currently have similar legislation pending.

From a scientific perspective, the gold standard for assessing citizenship comes from the National Assessment of Educational Progress (NAEP). NAEP tests students in grades 4, 8, and 10 in three areas: (1) civic knowledge, (2) intellectual and participatory skills, and (3) civic dispositions. The following are some example items: “Identify a civic responsibility (e.g., voting)”; “What is the main source of government funding (taxes)?”; “Why does the US Constitution limit the powers of government?”; “Identify one way to express an opinion on a public policy issue”; “Identify the appropriate official to contact to solve a problem.” Not all items are released to the public; however, subsets of items can be found within their technical manual (Carr, 2014). The test is scored using item response theory and scale anchoring (one assumes from the scant reporting in the technical manuals); however, no psychometric data regarding fit indices for items are available publicly, nor are any validation data. Remarkably, however, the NAEP civic assessment has been suspended indefinitely for fourth and twelfth graders due to sequestration (Klein, 2013) – only eighth graders will continue to take the test. Notably, no cuts were made to cognitively oriented assessments.

The International Association for the Evaluation of Educational Achievement (IEA) Civic Education Study (International Association for the Evaluation of Educational Achievement, 1999) was an international assessment of the civic knowledge and skills of youths around the world, and its goal was to investigate, using a comparative framework, the ways in which youths are prepared for their roles as citizens in democracies and aspiring democracies. In the late 1990s, the test was administered to a group of about 90,000 14-year-olds from 28 countries and a group of about 60,000 16–18-year-olds from 16 countries. The test contained three sections: (i) a multiple-choice section that tested civic knowledge and skills in interpreting civic information, (ii) a section asking for background information, and (iii) a section on conceptual understanding, attitudes, behavior, and actions. This last section was further divided to include civic knowledge, interpretation skills, economic literacy, conventional citizenship, social movement-related citizenship, trust in governmental institutions, economy-related government responsibilities, society-related government responsibilities, positive attitude toward one’s nation, positive attitude toward immigrants, confidence in participation in school, expected participation in political activities, and open climate for classroom discussion. Cronbach’s alphas ranged from 0.57 to 0.86 for the various subscales, but no validity information has been published.

Stemler (2015) has recently developed a new test of citizenship that is targeted at K-12 students. The test is designed to measure civic attitudes, behaviors, and cognitions (knowledge). The test was administered to 442 K-12 students, and the test demonstrated strong construct validity as well as predictive validity with teachers' ratings of the students' levels of civic engagement.

Leadership One of the most widely used measures of leadership is the Multifaceted Leadership Questionnaire (MLQ) developed by Bass and Avolio (1994). The instrument can be used as a self-report questionnaire, but can also be used in the form of a 360° rating scale assessment by asking others to evaluate a person of interest with regard to each of the items on the rating scale. In this way, multiple perspectives on the individual's leadership style emerge. Bass and Avolio have conceptualized leadership as falling into one of three main forms: (1) *transformational* (in which leaders inspire others, motivate action, and challenge those around them intellectually), (2) *transactional* (in which leaders manage those around them by contingent rewards), and (3) *passive/avoidant* (in which leaders manage by exception, i.e., pointing out mistakes of the subordinates, and generally take a passive role. The MLQ rating scale consists of 45 questions that are to be responded to on a Likert scale where 0 = "Not at all," 1 = "Once in a while," 2 = "Sometimes," 3 = "Fairly often," and 4 = "Frequently, if not always." Examples of questions include "Talks optimistically about the future," "Avoids making decisions," and "Spends time teaching and coaching." Various forms of the MLQ have been developed, including a shorter form. The internal consistency reliability of the instrument is excellent, with Tejeda, Scandura, and Pillai (2001) reporting values of the subscales ranging from 0.69 to 0.92, with a median value of 0.88 across four different samples. In addition, the instrument has been shown, from studies reported in the user manual as well as independent investigations, to have strong competency validity.

The Roets Rating Scale for Leadership (RRSL) has also been used to measure student leadership. It includes 26 items, which students use to provide a self-report rating of their leadership. Examples of items include "Listen to both sides," "Think one can do well as a leader," "Can work with different person types," and "Can say opinions in public" (Chan, 2000). Similarly, the leadership subscale of the Scales for Rating the Behavioral Characteristics of Superior Students (SRBCSS) has also been used to measure student leadership. This subscale includes seven items, and parent and teacher ratings on these items are used to evaluate students' leadership. He found that RRSL and SRBCSS leadership scores were significantly correlated with each other. Teacher ratings of student leadership on the Behavioral Assessment System for Children (BASC) have also been used as a measurement of student leadership with a reliability of 0.83 (Tolan & Larsen, 2014).

There also exist measures of leadership for teachers and principals. One of these measures is the Vanderbilt Assessment for Leadership in Education (VAL-ED). The VAL-ED is considered a "360° assessment" (Covay Minor et al., 2014) because it has teachers and supervisors rate their school principal in 36 domains of leadership. Ratings are done on a 1 to 5 Likert scale, ranging from "ineffective" to "outstandingly effective" (Covay Minor et al., 2014). They reported that the VAL-ED produces ratings of principals' effectiveness that exhibit 70 % agreement with ratings of the same principals' performance provided by school superintendents. Other measurements that can be used are measures of principal instructional leadership and teacher–principal trust (Sebastian & Allensworth, 2012). These measures involve ratings from teachers and principals and have been shown to have a reliability of 0.91 and 0.89, respectively.

Teamwork and Cooperation As with many of the instruments we have reviewed, we begin here by summarizing an instrument that historically has been used quite frequently to measure teamwork. The Teamwork Knowledge Skill and Ability (TKSA) Test is one measure that has been widely used to assess participants' interpersonal and self-management knowledge (Stevens & Campion, 1999). The instrument includes 35 multiple-choice items. An example item is as follows: "Suppose that you find yourself in an argument with several coworkers about who should do a very disagreeable but routine task. Which of the following would likely be the most effective way to resolve this situation? The four response options for this question are: (A) Have your supervisor decide, because this would avoid any personal bias. (B) Arrange for a rotating schedule so everyone shares the chore. (C) Let the workers who show up earliest choose on a first-come, first-served basis. (D) Randomly assign a person to do the task and don't change it."

Although the test authors (Stevens & Campion, 1999) have reported internal consistency reliability as high as 0.80, several other authors have reported lower consistency reliability estimates (e.g., alpha = 0.37–0.59). Furthermore, the primary authors contend that the instrument measures five dimensions of teamwork; however, confirmatory factor analyses suggest that single overarching factor fits the data best.

Consequently, Aguado, Rico, Sanchez-Manzanares, and Salas (2014) recently developed the Teamwork Competency Test (TCT), which was inspired by TWKSAT but that is designed to improve upon its psychometric properties. The TCT consists of 36 items that describe different situations that may arise within a work team. Respondents are asked to use a 4-point Likert scale, where "0" = never/almost never and "4" = always/almost always. Example items are "I often get involved in monitoring the task performance of other team members," "I care and act to make team conflicts explicit in a way that they can be solved," and "I provide my peers with relevant information on how well I think the team tasks are progressing." The TCT has excellent evidence supporting internal consistency reliability (alpha = 0.84 full scale). Furthermore, the instrument has strong predictive validity evidence associated with supervisor and self-evaluations of team performance.

The Individual Performance in Teams Scale (IPIT) is another popular measure of teamwork. It uses a rating-scale approach to the assessment of performance in teams. The IPIT consists of 33 items depicting various aspects of team member behavior related to conflict resolution, collaborative problem solving, communication, performance management, and task coordination. The items reflect issues/topics that previous literature has found to be important for team functioning. Sample behavior items include statements such as "tried to keep group aware of time issues," "responded calmly to others," and "helped resolve any conflicts." The IPIT was designed to be used by raters viewing videotaped team interactions. Raters are asked to indicate on a 7-point scale of behavior frequency with anchors ranging from 1 (to no extent) to 7 (to a great extent) the extent to which the behavior was observed. A "not applicable" was also an option. Items thought to impede team performance (e.g., rudely interrupted other members) were reverse coded. The scale also contains one general item concerning the team member's effectiveness during the team task. While interrater agreement should be assessed with each new study, McClough and Rogelberg (2003) found in their study that the three raters they used achieved strong levels of interrater reliability (intraclass correlation = 0.73). Peer ratings of individual teamwork performance were also gathered. Team members were asked to rate each other on five dimensions (e.g., participation in the group, interpersonal skills). However, the scale did not correlate, in that study, with peer ratings of team performance. This is a common danger associated with 360° feedback approaches – self- and other reports are rarely in alignment (Atwater, Ostroff, Yammarino, & Fleenor, 1998).

Similarly, in the classroom, situational judgment tests (SJTs) have been used to assess teamwork, in addition to self-report and teacher rating scales (Wang, MacCann, Zhuang, Liu, & Roberts, 2009). The SJT assessment gives students a scenario involving a group situation and asks them to rate how effective four possible responses to the given scenario would be. Students' effectiveness ratings are then used to calculate their SJT score. SJT scores were shown to correlate with teacher ratings of teamwork as well as self-report measures. In the Wang et al. study, all three methods showed convergent validity, but only teacher ratings of teamwork predicted composite course grades.

Ethics Perhaps the most widely used domain-general test of moral reasoning is the *Defining Issues Test* (Rest, 1979; Rest, Narvaez, Bebeau, & Thoma, 1999), which is based on Kohlberg's (1969) theory of moral development. Building on Piaget's work on moral judgment, Kohlberg advanced a theory of moral development comprising six stages, divided into three levels: (i) pre-conventional morality, (ii) conventional morality, and (iii) post-conventional morality. The *Defining Issues Test* consists of five situational dilemmas to which respondents are asked to respond. The classic example is the Heinz dilemma in which Heinz's wife is dying of cancer and a chemist has discovered a cure but is charging ten times more than Heinz can afford to pay and he will not alter the price. Without the drug, the wife will die. Participants are asked not only to select what they think Heinz should do (i.e., steal the drug or not) but also to rate the importance of various reasons for their decision (e.g., because it is against the law to steal; because he loves his wife and she will die without the medicine, so it is worth breaking the law). Kohlberg's theory has been criticized for its lack of applicability across gender (Gilligan, 1982) and cultures (Tietjen & Walker, 1985) and for the fact that it does not strongly correlate with actual behavior (Blasi, 1980; Gibbs et al., 1986). Item total correlations range from 0.48 to 0.77 with the internal consistency reliability for the overall scale reported to be 0.65 (Xu, Iran-Nejad, & Thoma, 2007).

An alternative conception of ethical thinking that overcomes some of the criticisms of the Kohlberg model has been advanced by Schwartz (1992). According to

Schwartz, there are ten universal ethical values (power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity, and security) that have been demonstrated to hold up empirically across more than 50 cultures (Schwartz & Bardi, 2001) and that relate, at least modestly, to relevant behavioral outcomes (Schwartz, 2007). The *Schwartz Value Scale (SVS)* is a 56-item self-report instrument that measures the extent to which individuals endorse different ethical values as guiding principles in their lives. Participants are asked to rate, on a scale ranging from 0 (opposed to my principles), 1 (not important), 4 (important), to 8 (of supreme importance), the importance as a life-guiding principle for them terms such as "power, that is, social power, authority, and wealth," and "achievement, that is, success, capability, ambition, and influence on people and events." Internal consistency reliabilities for subscales have been reported on the order of 0.58 and 0.60.

Recently, Stemler (2012b) has proposed an alternative instrument called the *Ethical Priority Test* that represents a compromised position between the universalism advocated by Kohlberg/Rest and the relativism advocated for by Schwartz. Specifically, like Schwartz, Stemler proposes a finite set of universal human values (e.g., honesty, kindness, responsibility, justice). These are basic values that are articulated in almost all religions and are found pervasively in the ethics and philosophy literature. Most people use these values as guiding principles in their lives. When given a choice between honesty and non-honesty, for example, most rational people will choose to be honest. However, Stemler argues that the reason we observe individual differences in behaviors is because reasonable people may prioritize each of the values differently. The test is designed for use with adolescents and college-aged students. An example item follows: "You have recently been appointed captain of your dance team. Earlier today, you fell down the stairs and broke your ankle. When you went to see your doctor, she said that you should rest as much as you can and not participate in the competition tomorrow. If you do not compete, not only will your team be losing one of its best dancers, but they also will need to change all of their formations. Do you decide to compete? (A) Yes; (B) No." A response of "Yes" corresponds to the theoretical value of "responsibility," whereas a response of "No" corresponds to a theoretical value of "safety."

The novelty of this assessment is that it does not yield a single score, but rather a profile of ethical values. Individual profiles (patterns of values) are generated using Latent Class Analysis. The instrument has shown sufficient evidence of test–retest reliability (ranging from 0.47 to 0.69 across scales). In addition, the EPT has demonstrated excellent convergent validity with predicted elements of the Schwartz and Kohlberg scales as well as discriminant validity with personality traits.

A few measures of ethics have been used specifically with students in school settings. One of these measures is the Ethical Sensitivity Scale Questionnaire. The questionnaire has 28 items that encompass the following seven categories: "(1) reading and expressing emotions, (2) taking the perspectives of others, (3) caring by connecting to others, (4) working with interpersonal and group differences, (5) preventing social bias, (6) generating interpretations and options, and (7) identifying the consequences of actions and options." Within these categories, each item is rated on a 1–5 Likert scale, “totally disagree” to “totally agree” (Tirri & Nokelainen, 2007). Examples of items include “I think it is good that my closest friends think in different ways,” “I believe there are several right solutions to ethical problems,” “I notice if someone working with me is offended by me,” and “I try to consider other people’s needs even in situations concerning my own benefits” (Tirri & Nokelainen, 2007). Internal consistency reliability estimates ranged from 0.50 to 0.78 for the seven scales, which is moderate.

An interesting self-report scale of ethical behavior in leisure is the Aristotelian Ethical Behavior in Leisure Scale (AEBLS) (Widmer, Ellis, & Trunnell, 1996). The AEBLS is a 62-item summative scale derived from research with adolescents. Each item represents one of four domains of ethical behavior that are derived in an interpretation of Aristotle’s Nicomachean ethics philosophy: intellectual activity, creative activity, moral behavior, and meaningful relationships. Widmer, Ellis, and Munson (2003) created a short form of the scale in which 26 items were selected for inclusion (AEBLS-S). The response format is scaled from one to five: one = never, two = seldom, three = sometimes, four = often, and five = always. Six items are reverse coded. Examples of items include “I think about world problems in my free time,” “I spend my holidays with people who are important to me,” and “I am fair when I play games.” The instrument has strong internal consistency reliability (0.98) and some evidence to support its validity.

Summary As was the case with emotional development, the competencies chosen to represent the citizenship are certainly not exhaustive, nor are the instruments discussed in this section that are used to measure these competencies. However, we have again tried to highlight instruments that are (1) psychometrically sound and (2) specifically tailored to K-12 populations and (3) represent a variety of interesting methods used for assessment. Table 3.2 presents a summary that summarizes the measures discussed in this section, highlighting the competency they are intended to measure, their psychometric properties, example items, and how to acquire these instruments.

3.6 Policy Implications of Aligning Mission and Measurement

Our fundamental thesis in this chapter is that mission and measurement should be aligned. Thus, our first task was to examine the mission of schools. A review of the empirical research drawn from a wide variety of data sources and a broad range of constituents reveals that emotional development and civic development are at least on equal footing with the cognitive element as core purposes of schooling (Stemler et al., 2011; Stemler & Bebell, 2012; Stemler, 2012a). Indeed, the data clearly show that broader competencies are not “add-ons” that are secondary to the cognitive purpose of schooling, but rather that these broader competencies tend to be mentioned as the core purposes of schooling with even greater frequency than the cognitive elements.

Table 3.2 Selected measures of civic knowledge, leadership, teamwork, and ethics

Measure	Sample item/item types	Psychometrics	Availability
Civic knowledge			
United States Citizenship and Immigration Services (USCIS) naturalization test	"Why did the colonists fight the British?"	Not found	Freely available from the United States Citizenship and Immigration Services website (http://www.uscis.gov)
National Assessment of Educational Progress (NAEP)	"Identify a civic responsibility (voting)" "What is the main source of government funding (taxes)?" "Why does the US Constitution limit the powers of government?" "Identify the appropriate official to contact to solve a problem"	Not found	Freely available from: http://nces.ed.gov
IEA Civic Education Test	Civic knowledge Civic behaviors	Cronbach's alpha ranges from 0.57 to 0.86 for various subtests	Sample items available from: http://www.iea.nl/cived.html
Stemler's ABC Citizenship Assessment	"What does being a good citizen of your school look like?" "Draw a picture and describe it" "If you disagreed with an action taken by your class council, what would you do?" "How important to you is it to help make decisions for your neighborhood or town?" "Who is the governor of your state?"	Convergent validity with teachers' ratings of students' civic engagement	Available from steven.stemler@wesleyan.edu
Leadership			
Multifaceted Leadership Questionnaire (MLQ)	"Talks optimistically about the future" "Avoids making decisions" "Spends time teaching and coaching"	Cronbach's alpha of the subscales ranges from 0.69 to 0.92, with a median value of 0.88 across four different samples	Available for purchase through: www.mindgarden.com

Roets Rating Scale for Leadership (RRSL)	"Listen to both sides" "Think one can do well as a leader" "Can work with different person types" "Can say opinions in public"	Cronbach's alpha of 0.83 for the BASC Convergent validity with scales for rating the behavioral characteristics of superior students	Available from Roets (1986)
Vanderbilt Assessment for Leadership in Education (VAL-ED)	A 360° feedback rating scale system where principals, teachers, and superintendents are rated	Internal consistency reliability reported at 0.89–0.91 Convergent validity with superintendent's ratings of teacher leadership	Available for purchase from: http://www.valed.com
Teamwork			
Teamwork, Knowledge, Skill, and Ability (TKSA) Test	Suppose that you find yourself in an argument with several coworkers about who should do a very disagreeable but routine task. Which of the following would likely be the most effective way to resolve this situation? (A) have your supervisor decide, because this would avoid any personal bias; (B) arrange for a rotating schedule so everyone shares the chore; (C) let the workers who show up earliest choose on a first-come, first-served basis; (D) randomly assign a person to do the task and don't change it	Internal consistency reliability as high as 0.80; although several others have reported lower consistency reliability estimates (e.g., alpha = 0.37–0.59)	Available by contacting Michael J. Stevens at: http://kozaigroup.com/who-we-are/michael-j-stevens-ph-d/
Teamwork Competency Test (TCT)	"I often get involved in monitoring the task performance of other team members" "I care and act to make team conflicts explicit in a way that they can be solved" "I provide my peers with relevant information on how well I think the team tasks are progressing"	Cronbach's alpha of 0.84 (full scale)	Available in the Appendix of the Aguado, Rico, Sánchez-Manzanares, and Salas (2014) paper

Table 3.2 (continued)

Measure	Sample item/item types	Psychometrics	Availability
Individual Performance in Teams Scale (IPIT)	"tried to keep group aware of time issues" "responded calmly to others" "helped resolve any conflicts"	Interrater reliability was acceptable (intraclass correlation = 0.73)	Available in the paper by McClough and Rogelberg (2003)
The multimethod assessment of teamwork	Includes self-report teamwork items (e.g., I enjoy sharing ideas) Situational judgment test items Behaviorally anchored teacher rating scale	Factor analyses confirmed the three-factor structure of the self-report measure with Cronbach's alphas ranging from 0.78 to 0.88. Alpha for the SJT was 0.71, while alpha for the teacher scale was 0.98	Contact Richard D. Roberts: rroberts@proexam.org
Ethics			
Defining Issues Test (DIT)	Heinz's wife is dying of cancer and a chemist has discovered a cure but is charging ten times more than Heinz can afford to pay and he will not alter the price. Without the drug, the wife will die. Participants are asked not only to select what they think Heinz should do but also to rate the importance of various reasons for their decision	Item total correlations range from 0.48 to 0.77 with an internal consistency reliability for the overall scale of 0.65	Available from the University of Alabama's Center for the Study of Ethical Development (http://ethicaldevelopment.ua.edu). Costs: approximately $2 per individual to administer and score

Schwartz Value Scale (SVS)	Power, that is, social power, authority, and wealth Achievement, that is, success, capability, ambition, and influence on people and events	Internal consistency reliabilities for subscales on the order of 0.58 and 0.60	Contact: crossculturalcentre@yahoo.com
Ethical Priority Test	Your little sister spent a great deal of time fixing her hair and makeup before her first school dance, and she asks you how she looks. You think she looks ridiculous but you realize that her date will be arriving any minute and telling her what you think might undermine her self-confidence. Do you tell her? (A) Yes; (B) No	Test–retest reliability ranging from 0.47 to 0.69 across scales. Convergent validity evidence with the DIT and SVS and discriminant validity evidence with personality measures	Freely available from: steven.stemler@wesleyan.edu
Ethical Sensitivity Scale Questionnaire	"I think it is good that my closest friends think in different ways" "I believe there are several right solutions to ethical problems" "I notice if someone working with me is offended by me"	Internal consistency estimates range from 0.50 to 0.78 for the seven scales	Contact Kirsi Tirri: Kirsi.Tirri@helsinki.fi
Aristotelian Ethical Behavior in Leisure Scale (AEBLS)	"I think about world problems in my free time" "I spend my holidays with people who are important to me" "I am fair when I play games"	Internal consistency reliability of 0.98	Freely available in the publication (Widmer et al., 1996)

This then led us to question why accountability systems tend to rely almost exclusively on measures of the cognitive domain. We believe that the reason for this is because there is a common misperception that there exist no objective, quantitative ways to measure broader, noncognitive competencies. Thus, the bulk of this chapter was devoted to reviewing psychometrically sound instruments measuring a wide variety of noncognitive competencies that are explicitly valued by businesses, courts, citizens, and schools. Our review was not intended to be exhaustive and our listings could easily be expanded. However, our main objective was simply to illustrate that instruments meeting strong psychometric standards do exist for the broader competencies schools care about. Furthermore, there are many innovative methods for assessment that go well beyond simple self-report measures. Further, a few larger efforts have recently emerged to help fill the gap of providing novel measures of key elements of school mission. Roberts and colleagues have developed what they call the Mission Skills Assessment (http://indexgroups.org/msa/). Stemler and Bebell have developed a number of new measures of broader outcomes aligned with school mission and have collected information on other measures that can be used to measure school mission (see http://www.purposeofschool.com). Consequently, we see no reason why measures of broader, noncognitive skills should be entirely ignored by accountability systems.

Accountability needs to start with the vision of the school – the mission set by the leadership in conjunction with input from the community and especially the teachers working there. Schools should be given the power to determine for themselves what they believe are their most important aims. They should then be given the tools to observe for themselves whether they are making progress toward the aims they hold most dear. When people are allowed to pursue with vigor those things about which they are truly passionate, and when they are supported in their quest, then we will truly see schools that are preparing students to be successful.

We propose that one approach for realizing this vision would be to have instruments, such as those listed in this chapter, included in an "approved list" by the US Department of Education. Instruments that are candidates for inclusion could be submitted to a special committee of the Department of Education for review. This committee would consist of experts in education and psychometrics, and their responsibility would be to determine whether the submitted instruments meet the federal seal of approval for use in measuring progress toward the measurement of a particular skill or outcome. Thus, the federal government would have a list of "approved" measures for a wide variety of important competencies that schools may wish to measure. Schools would then be responsible for articulating the aims they wish to pursue, and the state and federal government would assess progress toward achieving those aims by using one of the "approved" measures. This approach would allow schools to have the autonomy to set their own objectives that are responsive to local needs, while at the same time preserving the external desires for accountability.

The focus would not be exclusively on the extent to which schools had achieved their aims but also on the quality of their instructional plan for making progress. Schools would be required to select aims they care about, collect data on them, and evaluate how they are working and what they might want to do differently or whether they feel a change is needed. These reports would be evaluated every 3–4 years for coherence and accountability purposes. This model of accountability is similar to an accreditation approach and is much more appropriate for school level accountability than simply aggregating individual student scores on tests of cognitive achievement. In the end, our argument is simple. Give schools the choice to determine their goals, which they already do via their mission statement. Let schools develop a plan for what they will do and how they will know that they are accomplishing it. Assist them in this effort by providing a federally reviewed and approved list of instruments they can use to measure whatever competencies they care about. Then, give them time to execute their plan and take a closer look to see whether they have accomplished their aims. If they are making progress, then accolades would follow. If they are not, then further support and/or accountability recommendations would be triggered.

3.7 Conclusion

Aligning measurement with mission is an intuitive concept that is easily grasped by educators, students, parents, businesses, and policy makers. Indeed, it is remarkable that accountability systems could be based on any other premise! We believe that the misalignment of accountability systems and school purpose has been the chief source of resistance to the accountability movement. Most teachers we have met are in favor of the concept of accountability; but they want systems that are aligned with the goals they value. As we have demonstrated previously, a wide variety of sources converge on the importance of emotional development, civic development, and cognitive development as key outcomes of schooling. And there is abundant evidence to suggest that these three competencies are of equal value – in no source of evidence do we find any suggestion that one of the three competencies is of greater worth than the other two. We are not suggesting that an emphasis on the cognitive domain be abandoned. Rather, our goal in this chapter was to point out that (1) there is widespread consensus across a broad range of constituencies about the goals/competencies that form the core purposes of schooling and (2) there is a solid tradition of objective, quantitative, and psychometrically sound assessment of a wide variety of noncognitive competencies. The alignment of mission and measurement need not be a fantasy. It is a realistic option that is well within our grasp. It can and should be the basis for a new sort of accountability system – one in which schools have agency and the federal and state role is to support and monitor progress toward those goals.

References

Abd-El-Fattah, S. M. (2010). Garrison's model of self-directed learning: Preliminary validation and relationship to achievement. *The Spanish Journal of Psychology, 13*(2), 586–596.

AdvanceEd. (2006). *Accreditation standards for quality schools* (No. LC# 2006905052). Decateur: AdvanceEd.

Aguado, D., Rico, R., Sánchez-Manzanares, M., & Salas, E. (2014). Teamwork competency test (TWCT): A step forward on measuring teamwork competencies. *Group Dynamics: Theory, Research, and Practice, 18*(2), 101–121.

Amorose, A. J., & Horn, T. S. (2001). Pre-to post-season changes in the intrinsic motivation of first year college athletes: Relationships with coaching behavior and scholarship status. *Journal of Applied Sport Psychology, 13*(4), 355–373.

Association of American Colleges and Universities. (2010). *Raising the bar: Employer's views on college learning in the wake of the economic downturn*. Washington, DC: Hart Research Associates.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*, 577–598.

Bar-on, R. (1997). *EQ-I technical manual*. Toronto, ON: Multi-Health Systems.

Bass, B. M., & Avolio, B. J. (Eds.). (1994). *Improving organizational effectiveness through transformational leadership*. Thousand Oaks, CA: Sage.

Batanova, M., & Loukas, A. (2014). Unique and interactive effects of empathy, family, and school factors on early adolescents' aggression. *Journal of Youth and Adolescence, 43*(11), 1890–1902.

Bebell, D., & Stemler, S. (2002, April). *MCAS and the purpose of school*. Paper presented at the annual meeting of the New England Educational Research Organization, Northampton, MA.

Bebell, D., & Stemler, S. E. (2004, April). *Reassessing the objectives of educational accountability in Massachusetts: The mismatch between Massachusetts and the MCAS*. Paper presented at the American Educational Research Association, San Diego, CA.

Berleur, J., & Harvanek, R. F. (1997). *Analysis of mission statements or similar documents of Jesuit universities and higher education institutions*. Retrieved October 17, 2002, from http://www.info.fundp.ac.be/~jbl/mis-stat/index.htm

Birney, D., & Stemler, S. E. (2007). Intelligence quotient. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 2, pp. 473–476). Thousand Oaks, CA: Sage.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin, 88*, 1–45.

Brookhart, S. M., Walsh, J. M., & Zientarski, W. A. (2006). The dynamics of motivation and effort for classroom assessments in middle school science and social studies. *Applied Measurement in Education, 19*(2), 151–184.

Carr, P. (2014). *Grade 8 sample questions: Civics, geography, US history, science*. National Assessment of Educational Progress. Downloaded May 28, 2015 from: https://nces.ed.gov/nationsreportcard/subject/field_pubs/sqb/pdf/2014_sq_cghs_g8.pdf

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. Retrieved from http://www.conference-board.org/pdf_free/BED-06-Workforce.pdf

Chan, D. W. (2000). Assessing leadership among chinese secondary students in Hong Kong: The use of the roets rating scale for leadership. *Gifted Child Quarterly, 44*(2), 115–122.

Cianciolo, A., & Sternberg, R. J. (2004). *Intelligence: A brief history*. Boston: Wiley-Blackwell.

Common Core State Standards Initiative (2010a). *Common core state standards for english language arts & literacy in history/social studies, science, and technology subjects*. Downloaded November 30, 2014, from http://www.corestandards.org/wp-content/uploads/ELA_Standards.pdf

Common Core State Standards Initiative (2010b). *Common core state standards for mathematics*. Downloaded November 30, 2014, from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Covay Minor, E., Porter, A. C., Murphy, J., Goldring, E. B., Cravens, X., & Elloitt, S. N. (2014). A known group analysis validity study of the vanderbilt assessment of leadership in education in US elementary and secondary schools. *Educational Assessment, Evaluation and Accountability, 26*(1), 29–48.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology, 1980, 10, 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113–126.

de Wied, M., Maas, C., van Goozen, S., Vermande, M., Engels, R., Meeus, W., Matthys, W., & Goudena, P. (2012). Bryant's empathy index: A closer examination of its internal structure. *European Journal of Psychological Assessment, 23*(2), 99–104. Downloaded August 10, 2015, from http://devpsychopathologyru.nl/wp-content/uploads/2012/11/2007-Bryant%E2%80%99s-Empathy-Index.pdf

deMarrais, K. B., & LeCompte, M. D. (1995). *The way schools work: A sociological analysis of education* (2nd ed.). White Plains, NY: Longman Publishers.

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., et al. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism Developmental Disorder, 38*(3), 464–473.

Elliott, A. J., & Sheldon, K. M. (1997). Avoidance achievement motivation: A personal goals analysis. *Journal of Personality and Social Psychology, 73*(1), 171–185.

Gardner, P. D. (2007). *Recruiting trends: 2006–2007*. East Lansing, MI: Collegiate Employment Research Institute, Michigan State University.

Gibbs, J. C., Clark, P. M., Joseph, J. A., Green, J. L., Goodrick, T. S., & Makowski, D. G. (1986). Relations between moral judgment, moral courage, and filed independence. *Child Development, 57*, 1040–1043.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Goals 2000: Educate America Act, H.R. 1804 (1994).

Goodlad, J. I. (1979). *What schools are for*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Guglielmino, L. M. (1978). Development of the self-directed learning readiness scale. (Doctoral dissertation, University of Georgia, 1977). *Dissertation Abstracts International, 38,* 6467A.

Hills, P. R., Francis, L. J., & Jennings, P. (2011). The school short-form coopersmith self-esteem inventory: Revised and improved. *Canadian Journal of School Psychology, 26*(1), 62–71.

Howe, A., Pitten Cate, I. M., Brown, A., & Hadwin, J. A. (2008). Empathy in preschool children: The development of the southampton test of empathy for preschoolers (STEP). *Psychological Assessment, 20*(3), 305–309.

Immerwahl, J. (2000). *Great expectations: How Californians view higher education*. San Jose, CA: National Center for Public Policy and Higher Education and Public Agenda.

International Association for the Evaluation of Educational Achievement. (1999). *IEA Civic education study* (No. ICPSR 3892). Ann Arbor, MI: Interuniversity consortium for political and social research.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the basic empathy scale. *Journal of Adolescence, 29*, 589–611.

Kemler, E. R. (2015). *Defining and measuring responsible citizenship in higher education contexts in the US and Argentina*. Undergraduate thesis, Wesleyan University.

Klein, A. (2013, May 14). Sequestration forces cuts to national social studies tests. *Education Week*. Downloaded May 28, 2015, from http://blogs.edweek.org/edweek/campaign-k-12/2013/05/social_studies_naep_tests_post.html

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of socialization theory and research*. New York: Rand McNally.

Kurland, H., Peretz, H., & Hertz-Lazarowtiz, R. (2010). Leadership style and organizational learning: The mediate effect of school vision. *Journal of Educational Administration, 48*(1), 7–30.

Labaree, D. F. (1997). *How to succeed in school without really learning*. New Haven, CT: Yale University Press.

Leontopoulou, S. (2010). An exploratory study of altruism in greek children: Relations with empathy, resilience and classroom climate. *Psychology, 1*(5), 377–385.

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. *Learning and Individual Differences, 21*(1), 124–128.

Lounsbury, J. W., Saudargas, R. A., Gibson, L. W., & Leong, F. T. (2005). An investigation of broad and narrow personality traits in relation to general and domain-specific life satisfaction of college students. *Research in Higher Education, 46*(6), 707–729.

Lounsbury, J. W., Levy, J. J., Park, S. H., Gibson, L. W., & Smith, R. (2009). An investigation of the construct validity of the personality trait of self-directed learning. Learning and Individual Differences, 19(4), 411–418. doi:10.1016/j.lindif.2009.03.001.

MacCann, C., Fogarty, G. J., Zeidner, M., & Roberts, R. (2011). How emotional intelligence and coping affect academic success. *Contemporary Educational Psychology, 36*, 60–70.

Mavroveli, S., Petrides, K. V., Sangareau, Y., & Furnham, A. (2009). Exploring the relationships between trait emotional intelligence and objective socio-emotional outcomes in childhood. *British Journal of Educational Psychology, 79*(2), 259–272.

Mavroveli, S., & Sanchez-Ruiz, M. J. (2011). Trait emotional intelligence influences on academic achievement and school behaviour. *British Journal of Educational Psychology, 81*, 112–134.

Mayer, J., Salovey, P., & Caruso, D. (2002). *Mayer-Salovey-Caruso emotional intelligence test manual*. Toronto, ON: Multi-Health Systems.

McAuley, E., Duncan, T., & Tammen, V. V. (1987). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport, 60*, 48–58.

McClough, A. C., & Rogelberg, S. G. (2003). Selection in teams: An exploration of the teamwork knowledge, skills, and ability test. *International Journal of Selection and Assessment, 11*(1), 56–66.

McIntosh, S. (2012). *State high school exit exams: A policy in transition*. Washington, DC: Center on Educational Policy.

Mehrabian, A. (1996). *Manual for the balanced emotional empathy scale (BEES)*. Monterey, CA: Albert Mehrabian.

Mestre, V., Frias, M. D., & Samper, P. (2004). La medida de la empatia: Analysis del interpersonal reactivity index [Measuring empathy: Analysis of the interpersonal reactivity index]. *Psicothema, 16*, 255–260.

National Commission on Excellence in Education. (1983). A Nation at Risk: The Imperative for Educational Reform. Washington, DC: US Department of Education.

Noddings, N. (1995). *Philosophy of education*. Boulder, CO: Westview Press.

Ojanen, T., Sijtsema, J. J., Hawley, P. H., & Little, T. D. (2010). Intrinsic and extrinsic motivation in early adolescents' friendship development: Friendship selection, influence, and prospective friendship quality. *Journal of Adolescence, 33*(6), 837–851.

Parker, J. D. A., Creque, R. E., Barnhart, D. L., Harris, J. I., Majeski, S. A., Wood, L. M., et al. (2004). Academic achievement in high school: Does emotional intelligence matter? *Personality and Individual Differences, 37*(7), 1321–1330.

PISA (2012). *PISA 2012 Results: What students know and can do*. Paris: OECD.

Plenty, S., & Heubeck, B. G. (2011). Mathematics motivation and engagement: An independent evaluation of a complex model with Australian rural high school students. *Educational Research and Evaluation, 17*(4), 283–299.

Rae Westbury, H., & Neumann, D. L. (2008). Empathy-related responses to moving film stimuli depicting human and non-human animal targets in negative circumstances. *Biological Psychology, 78*(1), 66–74.

Rameson, L. T., Morelli, S. A., & Lieberman, M. D. (2012). The neural correlates of empathy: Experience, automaticity, and prosocial behavior. *Journal of Cognitive Neuroscience, 24*(1), 235–245.

Reed, R. F., & Johnson, T. W. (Eds.). (1996). *Philosophical documents in education*. White Plains, NY: Longman Publishers.

Renchler, R. (1991). Leadership with a vision: How principals develop and implement their visions for school success. *OSSC Bulletin, 34*(5), 1–29.

Renihan, P. J., Renihan, F. I., & Waldron, P. (1986). The common ingredients of successful school effectiveness projects. *Education Canada, 26*(3), 16–21.

Rest, J. (1979). *Development in judging moral issues*. Minneapolis, MN: University of Minnesota Press.

Rest, J., Narvaez, D., Bebeau, M., & Thoma, S. (1999). DIT-2: Devising and testing a new instrument of moral judgment. *Journal of Educational Psychology, 91*(4), 644–659.

Robinson, J.P., Shaver, P.R., & Wrightsman, L.S. (1991). (Eds). Measures of Personality and Social Psychological Attitudes. San Diego: Elsevier.

Roets, L. (1986). *Roets rating scale for leadership*. Des Moines, IA: Leadership Publishers.

Rojas, R., & Rich, M. (2015, January 27). States move to make citizenship exams a classroom aid. *New York Times*. Accessed May 28, 2015, from http://www.nytimes.com/2015/01/28/us/states-move-to-make-citizenship-exams-a-classroom-aid.html

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rousseau, C., Drapeau, A., Lacroix, L., Bagilishya, D., & Heusch, N. (2005). Evaluation of a classroom program of creative expression workshops for refugee and immigrant children. *Journal of Child Psychology and Psychiatry, 46*(2), 180–185.

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology, 40*(6), 451–475.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*(5), 749–761.

Ryan, R. M., Mims, B., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology, 45*, 736–750.

Schmitt, N. (2012). Development of rationale and measures of noncognitive college student potential. *Educational Psychologist, 47*(1), 18–29.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). New York: Academic.

Schwartz, S. (2007). Value priorties and behavior: Applying a theory of integrated value systems. In C. Seligman, J.M. Olson, and M.P. Zanna (2013). The Psychology of Values: The Ontario Symposium, Vol. 8, pp. 1–24.

Schwartz, S., & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross-Cultural Psychology, 32*(3), 268–290.

Sebastian, J., & Allensworth, E. (2012). The influence of principal leadership on classroom instruction and student learning: A study of mediated pathways to learning. *Educational Administration Quarterly, 48*(4), 626–663.

Shrauger, J., & Schohm, M. (1995). Self-confidence in college students: Conceptualization, measurement, and behavioral implications. *Assessment, 2*(3), 255–278.

Sperling, R. A., Sherwood, T. P., & Hood, A. M. (2013). Relating motivation to read and emergent reading skills: A measurement validation study. *Reading Psychology, 34*(5), 461–485.

Stemler, S. E. (2001). Examining school effectiveness at the fourth grade: A hierarchical analysis of the third international mathematics and science study (TIMSS). (Doctoral dissertation, Boston College, 2001). *Dissertation Abstracts International, 62(03A), 919*.

Stemler, S. E. (2012a). What should university admissions tests predict? *Educational Psychologist, 47*(1), 5–17.

Stemler, S. E. (2012b). *Using situational judgment testing to measure ethical values among college students.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Stemler, S. E. (2015). *Measuring creativity and citizenship in K-12 students*. Final report of the evaluation of the Higher Order Thinking Skills (HOTS) schools in connecticut.

Stemler, S. E., & Bebell, D. (1999, April). *An empirical approach to understanding and analyzing the mission statements of selected educational institutions*. Paper presented at the New England Educational Research Organization (NEERO), Portsmouth, NH.

Stemler, S. E., & Bebell, D. (2012). *The school mission statement: Values, goals, and identities in American education*. New York: Routledge.

Stemler, S. E., Bebell, D., & Sonnabend, L. (2011). Using mission statements for reflection and research. *Educational Administration Quarterly, 47*(2), 383–420.

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25*, 207–228.

Stober, S. S. (1997). *A content analysis of college and university mission statements*. Unpublished doctoral dissertation, Temple University, Philadelphia, PA.

Straka, G. A. (1995). *Problems of measuring self-directed learning readiness*. Paper presented at the Asia-Pacific seminar on self-directed learning, Seoul, Korea.

Teddlie, C., & Reynolds, D. (Eds.). (2000). *The international handbook of school effectiveness research*. New York: Falmer Press.

Tejeda, M. J., Scandura, T. A., & Pillai, R. (2001). The MLQ revisited: Psychometric properties and recommendations. *The Leadership Quarterly, 12*, 31–52.

Tietjen, A. M., & Walker, L. J. (1985). Moral reasoning and leadership among men in a Papua New Guinea society. *Developmental Psychology, 21*, 982–989.

Timothy, T., Chee, T. S., Beng, L. C., Sing, C. C., Ling, K. J. H., Li, C. W., et al. (2010). The self-directed learning with technology scale (SDLTS) for young students: An initial development and validation. *Computers & Education, 55*(4), 1764–1771.

Tirri, K., & Nokelainen, P. (2007). Comparison of academically average and gifted students' self-rated ethical sensitivity. *Educational Research and Evaluation, 13*(6), 587–601.

Tolan, P. H., & Larsen, R. (2014). Trajectories of life satisfaction during middle school: Relations to developmental-ecological microsystems and student functioning. *Journal of Research on Adolescence, 24*(3), 497–511.

Tyack, D. B. (1988). Ways of seeing: An essay on the history of compulsory schooling. In R. M. Jaeger (Ed.), *Complementary methods for research in education* (pp. 24–59). Washington, DC: American Educational Research Association.

United States Citizenship and Immigration Services. (2014). *Civics test.* Retrieved November 25, 2014, from http://www.uscis.gov

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. *Canadian Journal of School Psychology, 24*(2), 108–124.

West, R. F., & Bentley, E. I. (1990). Structural analysis of the self-directed learning readiness scale: A confirmatory factor analysis using LISREL modeling. In H. B. Long (Ed.), *Advances in research and practice in self-directed learning* (pp. 157–180). Norman, OK: Oklahoma Research Center for Continuing Professional and Higher Education of the University of Oklahoma.

Widmer, M. A., Ellis, G. D., & Munson, W. W. (2003). Development of the Aristotelian ethical behavior in leisure scale short form. *Therapeutic Recreation Journal, 37*(3), 256–274.

Widmer, M. A., Ellis, G. D., & Trunnell, E. P. (1996). Measurement of ethical behavior in leisure among high- and low-risk adolescents. *Adolescence, 31*(122), 397–408.

Woods, S., Wolke, D., Nowicki, S., & Hall, L. (2009). Emotion recognition abilities and empathy of victims of bullying. *Child Abuse and Neglect, 33*(5), 307–311.

Xu, Y., Iran-Nejad, A., & Thoma, S. J. (2007). Administering defining issues test online: Do response modes matter? *Journal of Interactive Online Learning, 6*(1), 10–27.

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M., Brandon, P. R., Furtak, E. M., et al. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335–359.

Zeidner, M., Matthews, G., & Roberts, R. D. (2009). *What we know about emotional intelligence: How it affects learning, work, relationships, and our mental health*. Cambridge, MA: MIT Press.