



Assessing 21st Century Skills: Summary of a Workshop

ISBN
978-0-309-21790-3

158 pages
6 x 9
PAPERBACK (2011)

Judith Anderson Koenig, Rapporteur; National Research Council

 Add book to cart

 Find similar titles

 Share this PDF



Visit the National Academies Press online and register for...

- ✓ Instant access to free PDF downloads of titles from the
 - NATIONAL ACADEMY OF SCIENCES
 - NATIONAL ACADEMY OF ENGINEERING
 - INSTITUTE OF MEDICINE
 - NATIONAL RESEARCH COUNCIL
- ✓ 10% off print titles
- ✓ Custom notification of new releases in your field of interest
- ✓ Special offers and discounts

Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences. Request reprint permission for this book

ASSESSING 21ST CENTURY SKILLS

Summary of a Workshop

Judith Anderson Koenig, *Rapporteur*

Committee on the Assessment of 21st Century Skills

Board on Testing and Assessment

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Award No. N01-OD-4-2139, TO #199 between the National Academy of Sciences and the National Institutes of Health; and Award No. DRL-0956233 between the National Academy of Sciences and the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-21790-3

International Standard Book Number-10: 0-309-21790-3

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, <http://www.nap.edu>.

Copyright 2011 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

Suggested citation: National Research Council. (2011). *Assessing 21st Century Skills: Summary of a Workshop*. J.A. Koenig, Rapporteur. Committee on the Assessment of 21st Century Skills. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON THE ASSESSMENT OF 21ST CENTURY SKILLS

Joan L. Herman (*Chair*), CRESST, University of California, Los Angeles
Greg J. Duncan, University of California, Irvine
Deirdre J. Knapp, HumRRO
Patrick C. Kyllonen, Center for Academic and Workplace Readiness
and Success, Educational Testing Service
Paul R. Sackett, University of Minnesota
Juan I. Sanchez, Florida International University
Steven L. Wise, Northwest Evaluation Association

Judith A. Koenig, *Study Director*
Stuart Elliott, *Director*, Board on Testing and Assessment
Margaret Hilton, *Senior Program Officer*
Kelly Iverson, *Senior Program Assistant*

Preface

Change is omnipresent in today's world. Knowledge is growing exponentially as technology continually transforms the way we live and work. From local to state and national perspectives, global markets and forces are transcendent. Stunning scientific and engineering advances have brought with them vexing social, political, and economic dilemmas. Individually and collectively, citizens in a democracy need to be able to respond to these changing conditions, make informed decisions, and take action to solve current and future challenges.

It would seem to go without saying that students of today must be prepared to take hold of life's demands and thrive in tomorrow's world. The changing nature of the workplace is a prime case in point. The routine jobs of yesterday are being replaced by technology and/or shipped offshore. In their place, job categories that require knowledge management, abstract reasoning, and personal services seem to be growing. These jobs involve skills that cannot easily be automated, such as adaptive problem solving, critical thinking, complex decision making, ethical reasoning, and innovation. Technology cannot be programmed to serve as supervisors or to perform tasks that rely on effective human interactions. It cannot easily be trained to negotiate, persuade, or perceptively handle person-to-person interactions. It cannot teach a classroom of students, treat the sick, care for the elderly, wait on tables, or provide other such services. These are all tasks for humans.

Effectiveness in the workforce also requires the ability to work autonomously, be self-motivating and self-monitoring, and engage in lifelong

learning. Individuals must be able to adapt to new work environments, communicate using a variety of mediums, and interact effectively with others from diverse cultures. Increasingly, workers must be able to work remotely in virtual teams.

This broad set of cognitive and affective capabilities that undergirds success today often is referred to as “21st century skills.” Numerous reports from higher education, the business community, and labor market researchers alike argue that such skills are valued by employers, critical for success in higher education, and underrepresented in today’s high school graduates.

The National Research Council (NRC) has conducted a series of activities to address the issue of 21st century skills in education today. In October 2005, the NRC convened a planning meeting intended to explore the role of K-12 education in developing these skills. Participants identified three critical unanswered questions and encouraged that they be further explored:

1. Is there a body of evidence supporting a taxonomy of 21st century skills coupled to individual and societal well-being?
2. Do we have evidence of effective models to teach 21st century skills through science, technology, engineering, and mathematics (STEM) education?
3. How can we assess 21st century skills?

The first question was addressed as part of a 2-day workshop held in 2007 (see National Research Council, 2008), and the second question was explored during a 2-day workshop held in 2009 (see National Research Council, 2010). These two workshops identified and defined a set of five broad skills that included adaptability, complex communication and social skills, nonroutine problem solving, self-management and self-development, and systems thinking.

The third question was the focus of the present workshop. Jointly funded by the National Institutes of Health (NIH) and the National Science Foundation (NSF), the workshop was designed to address the following questions:

- How can 21st century skills be assessed?
- What assessments of these skills are currently available and how well do they work?
- What needs to be done in order to develop additional assessments of these skills?
- How should the assessment results be used?

The goal for this workshop was to capitalize on the prior efforts and explore strategies for assessing the five skills identified at the earlier workshops. The Committee on the Assessment of 21st Century Skills was asked to organize a workshop that reviewed assessments and related research for each of the five skills, with special attention to recent developments in technology-enabled assessment of critical thinking and problem-solving skills. The workshop was conducted in two parts. The first, held in January in Irvine, California, was a 2-day activity that focused on research and measurement issues associated with assessing these skills. The second, held in May in Washington, DC, was a half-day discussion of policy and practice issues.

Many people contributed to the success of these activities. We first thank the sponsors for their support of this work, NIH and NSF. We particularly thank Bruce Fuchs with NIH and Gerhard Salinger with NSF for their commitment to and assistance with the committee's organization of the workshop. This workshop would not have become a reality without their generous support.

The committee also thanks the four scholars who wrote papers and discussed them at the workshop: Eric Anderman, Ohio State University; Stephen Fiore, University of Central Florida; Rick Hoyle, Duke University; and Nathan Kuncel, University of Minnesota.

We also greatly appreciate the work of the presenters who discussed examples of assessments of 21st century skills: John Behrens, Cisco Systems; Deborah Boisvert, Boston Area Advanced Technical Education; Heather Butler, Claremont McKenna College; Susan Case, National Conference of Bar Examiners; Tim Cleary, University of Wisconsin–Milwaukee; Lynn Gracin Collins, SH&A/Fenestra; Joachim Funke, University of Heidelberg; Art Graesser, University of Memphis; Bob Lenz, Envision Schools; Filip Lievens, Ghent University, Belgium; Gerald Matthews, University of Cincinnati; Richard Murnane, Harvard University; Candice Odgers, University of California, Irvine; and Louise Yarnall, SRI.

We are also grateful to senior staff members of the NRC's Division of Behavioral and Social Sciences and Education (DBASSE) who helped to move this project forward. Robert Hauser, executive director, and Patricia Morison, associate executive director for reports and communication, provided support and guidance at key stages in this project. The committee also thanks the NRC staff members that worked directly on this project. Kelly Iverson, senior project assistant, provided deft organizational skills and careful attention to detail that helped to ensure the success of the workshop. We sincerely appreciate Kelly's help in handling all of the logistical and contractual issues with the workshop and her assistance with manuscript preparation. We thank Judy Koenig, study director, who organized the workshop. We are also grateful to Stuart Elliott, Board on

Testing and Assessment (BOTA) director, and Margaret Hilton, BOTA senior program officer, for their contributions in formulating the design of the workshop and making them both a reality. We particularly wish to recognize Alix Beatty for her assistance in writing Chapter 3 of the workshop report.

Finally, as chair of the committee, I thank the committee members for their dedication and outstanding contributions to this project. They gave generously of their time in planning the workshops and actively participated in workshop presentations and discussions. Their varied experiences and perspectives contributed immeasurably to the success of the project and made them a delightful set of colleagues for this work. I learned a lot from each of them, and for that, I am especially grateful.

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the NRC's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We thank the following individuals for their review of this report: Edward Haertel, School of Education, Stanford University; Milt Hakel, President, Alliance for Organizational Psychology and Professor and Ohio Eminent Scholar Emeritus, Department of Psychology, Bowling Green State University; Michael E. McManus, Dean of Academic Programs, University of Queensland; Keith Millis, Department of Psychology, Northern Illinois University; Paul Nichols, Senior Associate, Center for Assessment, National Center for the Improvement of Educational Assessment; Cornelia S. Orr, Executive Director, National Assessment Governing Board; and Mark R. Wilson, Professor of Policy, Organization, Measurement, and Evaluation Cognition and Development, Graduate School of Education, University of California, Berkeley.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report nor did they see the final draft of the report before its release. The review of this report was overseen by Mark Wilson, University of California, Berkeley. Appointed by NRC, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author(s) and the institution.

Joan L. Herman, *Chair*
Committee on the Assessment of 21st Century Skills

Contents

1	Introduction	1
2	Assessing Cognitive Skills	15
3	Assessing Interpersonal Skills	39
4	Assessing Intrapersonal Skills	63
5	Measurement Considerations	93
6	Synthesis and Policy Implications	107
	References	119
Appendixes		
A	Agenda and Participants for the January Workshop	127
B	Agenda and Participants for the May Workshop	139

1

Introduction

There is growing recognition that individuals need a wide array of skills in order to meet the needs of the modern workplace. Gone are the days when a multitude of jobs were available that required workers to perform simple manual tasks. The introduction of technology, particularly the use of computers, has changed the way that workers perform their tasks and the types of training and skills that workers need in order to complete these tasks. Research has shown that the use of computers has eliminated the need for humans to perform tasks that involve solving routine problems or communicating straightforward information (Autor, Levy, and Murnane, 2003; Levy and Murnane, 2004). Non-routine problem-solving and complex communication and social skills are becoming increasingly valuable in the labor market. The modern workplace requires workers to have broad cognitive and affective skills. Often referred to as “21st century skills,” these skills include being able to solve complex problems, to think critically about tasks, to effectively communicate with people from a variety of different cultures and using a variety of different techniques, to work in collaboration with others, to adapt to rapidly changing environments and conditions for performing tasks, to effectively manage one’s work, and to acquire new skills and information on one’s own.

The National Research Council (NRC) has convened two prior workshops on the topic of 21st century skills. The first, held in 2007, was designed to examine research on the skills required for the 21st century workplace and the extent to which they are meaningfully different from

earlier eras and require corresponding changes in educational experiences. One theme from that workshop was that across the entire labor market—from high-wage biotechnology scientists and computer sales engineers to low-wage restaurant servers and elder caregivers—five skills appear to be increasingly valuable: adaptability, complex communication skills, nonroutine problem-solving skills, self-management/self-development; and systems thinking (National Research Council, 2008).

The second workshop, held in 2009, was designed to explore demand for these types of skills, consider intersections between science education reform goals and 21st century skills, examine models of high-quality science instruction that may develop the skills, and consider science teacher readiness for 21st century skills. A message that emerged from this workshop was that although some new assessments incorporate items that appear promising as potential measures of students' 21st century skills, additional research may be needed in order to more clearly define the constructs and to develop frameworks for assessment of these skills (National Research Council, 2010).

The present workshop was intended to delve more deeply into the topic of assessment. The goal for this workshop was to capitalize on the prior efforts and explore strategies for assessing the five skills identified earlier. The Committee on the Assessment of 21st Century Skills was asked to organize a workshop that reviewed the assessments and related research for each of the five skills identified at the previous workshops, with special attention to recent developments in technology-enabled assessment of critical thinking and problem-solving skills.

In designing the workshop, the committee collapsed the five skills into three broad clusters as shown below:

Cognitive skills: nonroutine problem solving, critical thinking, systems thinking

Interpersonal skills: complex communication, social skills, teamwork, cultural sensitivity, dealing with diversity

Intrapersonal skills: self-management, time management, self-development, self-regulation, adaptability, executive functioning

The committee commissioned a set of papers to examine the research on assessing skills within each of these broad clusters and identified examples of assessments of the skills to feature at the workshop. The workshop was held in two parts. The first, convened in Irvine, California, in January 2011, was more technical in focus. The second, held in Washington, DC, in May 2011, was more policy focused. This report provides an integrated summary of the presentations and discussions from both parts of the workshop.

The remainder of this chapter is intended to provide context for the report, describing the changes in both the labor force and the workplace over the past few decades and discussing the skills that workers need to adequately perform in the currently available jobs. Chapter 2 discusses the skills included within the cognitive cluster. The chapter first explores issues related to defining these constructs, then presents four examples of assessments of these constructs, and concludes with a discussion of the strengths and weaknesses of these assessments. Chapters 3 and 4 follow the same format for skills within the interpersonal and intrapersonal clusters, respectively. Chapter 5 summarizes two workshop presentations that focused on key measurement issues to consider when developing assessments of 21st century skills. Chapter 6 concludes with workshop participants' synthesis of important points raised over the course of the two workshops and a discussion of the policy implications.

It is important to be specific about the nature of this report, which is intended to document the information presented in the workshop presentations and discussions and lay out the key ideas that emerged from the workshop. As such, the report is confined to the material presented by the workshop speakers and participants. Neither the workshop nor this summary is intended as a comprehensive review of what is known about assessing 21st century skills, although it is a general reflection of the literature. The presentations and discussions were limited by the time available for the workshop.

This summary was prepared by an independent rapporteur, and it does not represent findings or recommendations that can be attributed to the steering committee. The steering committee was responsible only for the quality of the agenda and the selection of participants. The workshop was not designed to generate consensus conclusions or recommendations but focused instead on the identification of ideas, themes, and considerations that contribute to an understanding of assessing 21st century skills.

THE CHANGING NATURE OF THE WORKPLACE

Richard Murnane, an economist with the Harvard School of Education, opened the workshop with a presentation about the changes that are occurring in the workplace and the types of skills workers will need to perform these tasks. He began by presenting two graphs—one for men and one for women—that displayed average hourly wages from 1979 through 2007 for individuals grouped by their education level.¹ These graphs, reproduced as Figures 1-1 and 1-2, show wage informa-

¹Murnane's presentation is available at http://www7.nationalacademies.org/bota/21st_Century_Workshop_Murnane.pdf [August 2011].

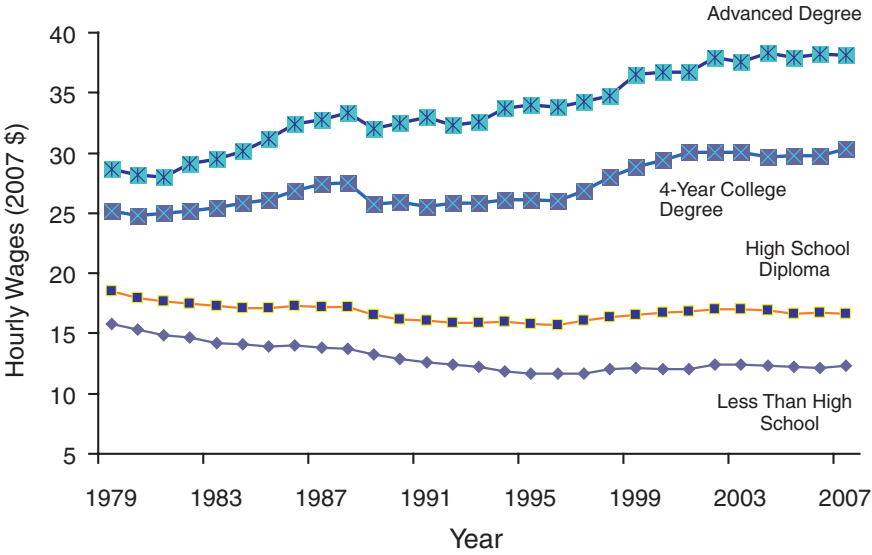


FIGURE 1-1 Men's real hourly wage by education, 1979-2007 (2007 dollars).
SOURCE: Richard Murnane's presentation. Used with permission.

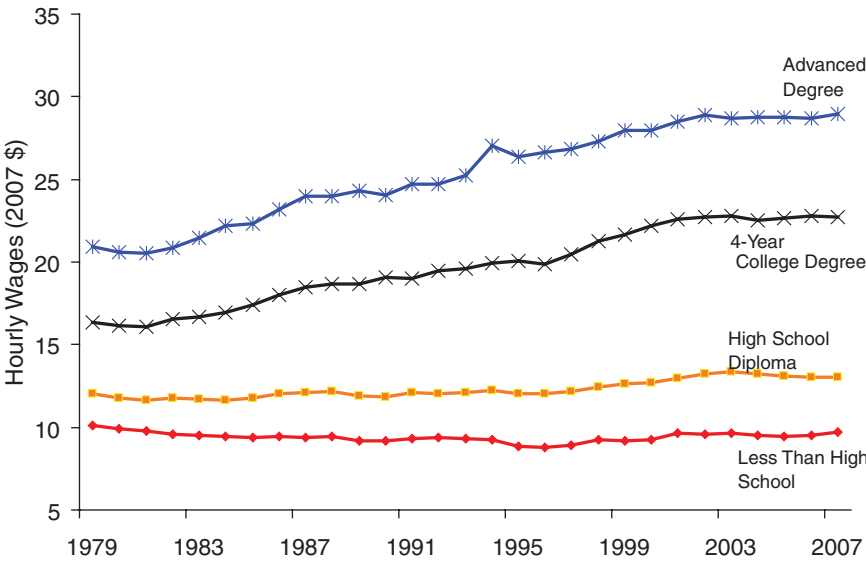


FIGURE 1-2 Women's real hourly wage by education, 1979-2007 (2007 dollars).
SOURCE: Richard Murnane's presentation. Used with permission.

tion for individuals with less than a high school diploma (bottom line, marked with diamonds), a high school diploma (next line up, marked with squares), an undergraduate college degree (4 years of college, line marked with x's), and advanced degrees (top line). The graphs reveal a steady increase in the differences in wages by education level for both men and women.

Over the years, the average hourly wages for men with high school diplomas or less changed very little, and by 2007, were slightly lower than in 1979. However, average hourly wages for men with at least a college degree steadily increased over the years to nearly \$30 for those with college degrees and nearly \$40 for those with advanced degrees. In 2007, men with advanced degrees made more than 2½ times as much per hour as men with less than a high school diploma. The same pattern holds for women, although women averaged lower hourly pay at each education level than their male counterparts. Murnane interpreted this information as indicating that educational attainment appears to play a larger role today in explaining average earnings than it did in 1979, noting “the gap between the premium [that] employers pay college graduates relative to high school graduates” has grown.

Economists tend to think in terms of supply and demand. In this context, “supply” refers to the characteristics and qualifications of individuals available to work, in other words, the characteristics of the labor force. Likewise, “demand” refers to the characteristics and qualifications that employers are looking for in their employees. In the labor force the two work together to influence wages. When demand for certain types of skills is high but the supply of workers with these skills is low, employers will pay more to get the workers they need. When there is a large supply of workers with certain skills but little demand for these skills, employers will pay less. Murnane suggested one explanation for the trends seen in the graphs is that the demand-side of what employers wanted did not change, while the supply-side of the available labor force did. That is, it could be that the labor force includes fewer college graduates relative to high school graduates than in the past, creating a situation where employers needed to pay higher wages to the relatively small proportion of available individuals with the needed qualifications. The data do not support this explanation, however.

To explain, Murnane displayed a graph comparing the educational attainment of the U.S. labor force in 1980 and 2000. Figure 1-3 shows the percentage of the labor force that dropped out of high school, graduated from high school, completed some college, and completed 4 years or more of college. For each education level, the left-most bar shows the percent for 1980 and the right-most bar shows the percent for 2000. As Figure 1-3 shows, the percentage of the labor force with at least some college

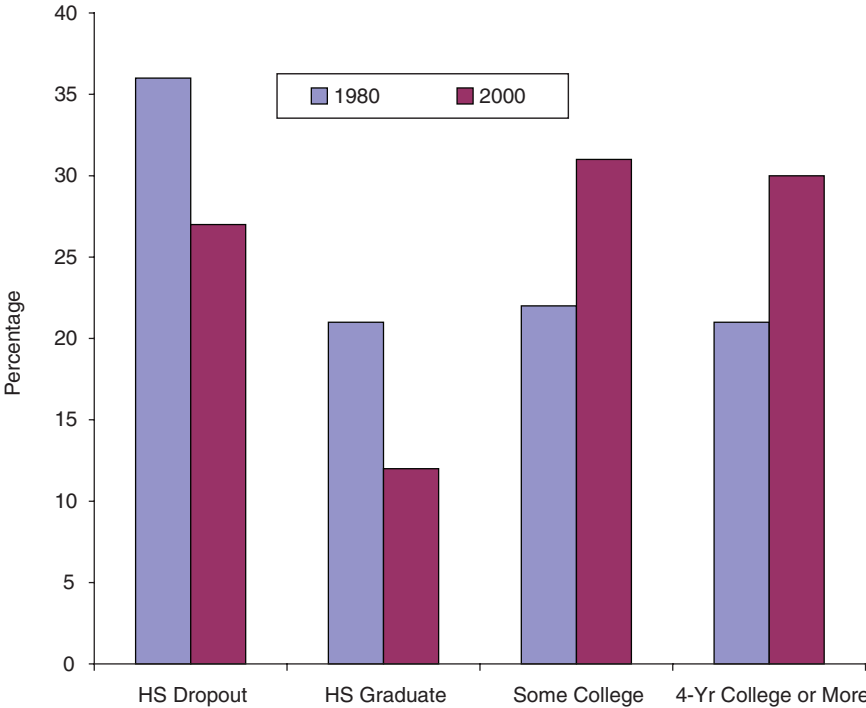


FIGURE 1-3 Educational attainment of the U.S. labor force, 1980 (left bar) and 2000 (right bar).
SOURCE: Richard Murnane’s presentation. Used with permission.

has increased since 1980; thus, it does not appear to be that the supply of college-educated people has decreased. Instead, Murnane believes the pay differences are more likely related to changes on the demand-side of the equation: employers are increasingly interested in individuals not just with a college education but who have certain types of skills.

If this is indeed the explanation, what changes have occurred on the demand-side that would help to account for it? Murnane suggested two related factors. First, there is increased use of computers and other forms of technology, and workers need training in order to make use of these technologies. Those who have this training are more suited to the available jobs, more attractive to employers, and more likely to obtain the higher-paying jobs.

Second, the use of computers interacts with the kinds of jobs available. Computers are particularly good at performing some types of jobs,

such as those that require routine tasks, rely on rule-based logic, and can be programmed. Increasingly, computers are replacing humans in performing these types of jobs. For instance, Murnane explained, airline passengers rarely get boarding passes from humans any longer, the use of automated self-checkout lines at the grocery store is growing, and most people do their banking with automated teller machines. Computers are not appropriate for other types of jobs, however, such as those that do not follow rule-based logic, those that require on-the-spot judgments, and those in which human interaction is essential. Some of these kinds of jobs—such as personnel managers and classroom teachers—require advanced training. Others—such as waiting on tables, caring for the elderly, and serving as a short order cook—require little advanced training.

Murnane said the growing income difference is due to an increased need for individuals to work in jobs that require technological skills, while, at the same time, there is a decreased need for individuals to perform routine tasks that can be computerized. Individuals without advanced training are employed in service jobs for which pay has been steady over time. Individuals with advanced training are working in the other jobs, in which pay has steadily increased.

Murnane argued that data on the types of jobs available supports this hypothesis. Figure 1-4 shows the percentage of people working in seven major job categories in 1979 and 2009. The job categories are arranged in order (left to right) from lowest paying to highest paying. In 1979, nearly 50 percent of the labor force was employed in blue collar and administrative support jobs. By 2009, the occupational distribution had shifted con-

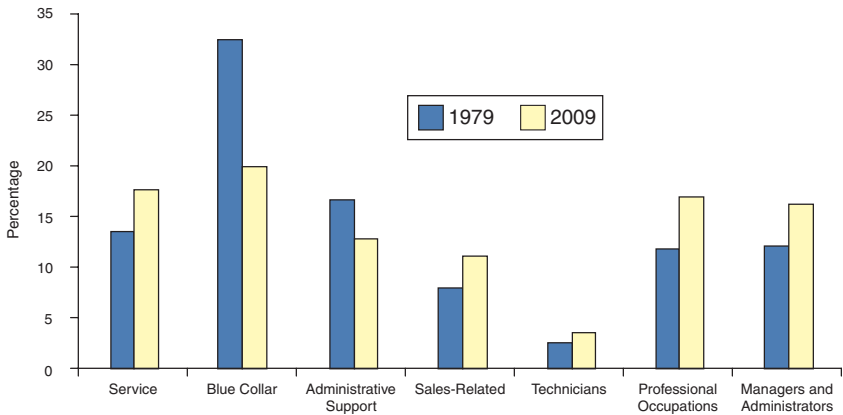


FIGURE 1-4 Distribution of occupations in the United States, 1979 and 2009.
SOURCE: Richard Murnane’s presentation. Used with permission.

siderably, with large declines in the percentage of individuals employed in blue color or administrative work and increases in the percentages of individuals employed in service occupations, professional occupations, and as managers or administrators.

So, which skills do people need in order to be attractive to employers and to perform well in these jobs? With his colleagues Autor and Levy, Murnane has studied the tasks required for various jobs. The researchers group the tasks into four categories:

- 1. Routine cognitive tasks, such as bookkeeping and filing
- 2. Routine manual tasks, such as assembly line work
- 3. Tasks that require expert thinking, such as identifying and solving new problems
- 4. Tasks that require complex communication, such as eliciting critical information and conveying a convincing interpretation of it

The researchers compiled data on the percentage of available jobs that require these four types of tasks and tracked the trends over a 30-year period (from 1969 to 1998). This information is displayed in Figure 1-5. In this figure the x-axis notes the years studied. The y-axis notes the change in the percentage of jobs that require the tasks, using 1969 as the

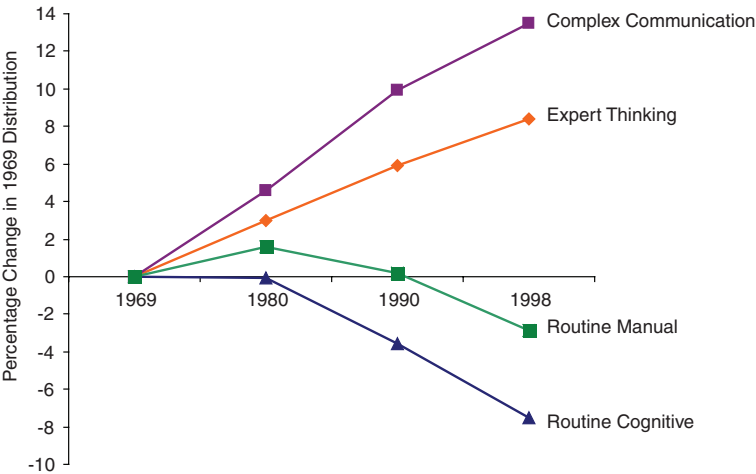


FIGURE 1-5 Economy-wide measures of routine and nonroutine task input: 1969-1998 (1969 = 0).

SOURCE: Levi and Murnane (2004). Reprinted with permission of Princeton University.

base year. Thus, the figure shows that the percentage of jobs that require routine cognitive tasks (line marked with black triangles) was steady from 1969 to 1980 and then began a steady decline. Likewise, the percentage of jobs that require routine manual tasks (line marked with gray squares) was relatively steady until 1990 and then began to decline. The top two lines show that the percentages of jobs that require expert thinking (line marked with gray diamonds) and complex communication (line marked with black squares) have steadily increased since 1969. Murnane interpreted this information as demonstrating that expert thinking and complex communication are clearly tasks that are increasingly in demand by employers.

Murnane has done additional work to explore the components of expert thinking and complex communication in order to better understand the attributes that are most important for the available jobs. His studies reveal that the components of expert thinking include the following:

- Within a domain, workers need a deep understanding of the domain and relationships within it
- Pattern recognition
- A sense of initiative (i.e., when you see a new task, is this a challenge you are anxious to take on or one you shy away from?)
- Metacognition (i.e., monitoring your own problem solving)

Likewise, the components of complex communication include the following:

- Observing and listening
- Eliciting critical information
- Interpreting the information
- Conveying the interpretation to others

At the workshop, Deborah Boisvert, a researcher with the Boston Advanced Technological Education Connection (BATEC), presented survey results that provide additional insight on the skills workers need in the current job market.² In 2007, BATEC conducted a survey designed to learn more about the skills employers sought in their employees. The skills rated most highly by the survey respondents included the following:

²Boisvert's presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Boisvert.pdf [August 2011]. Additional information about BATEC is available at <http://www.BATEC.org> [August 2011].

- Communication skills (oral and written)
- Ability to work productively in teams and groups (teamwork skills)
- Customer and business focus (understanding the big picture)
- Ability to listen for meaning and comprehension
- Ability to prioritize work and self-evaluate (self-reflection and time management)
- Development of original solutions to novel problems (problem solving)
- Ability to lead and act responsibility (leadership and ethics)

Boisvert said that in follow-up interviews, survey respondents told her and her colleagues “while technical skills may help someone get an interview, it is the soft skills [such as those listed above] that get the person the job.”

Further evidence of the importance of these skills is documented in a recent study that Murnane discussed. Lindqvist and Westman (2011) conducted a study on the labor market outcomes for men in Sweden using a large sample of 18-year-old males enlisted in the country’s military. The study examined the relationships between cognitive and noncognitive skills and labor market outcomes. The noncognitive skills assessed were

- Willingness to assume responsibility
- Independence
- Outgoing character
- Persistence
- Emotional stability
- Initiative
- Social skills

Their research findings indicated that compared to measures of cognitive skills, measures of noncognitive skills were stronger predictors of wages,³ stronger predictors of employment status,⁴ and stronger predictors of annual earnings.⁵

³A one standard deviation increase in the measure of noncognitive skills predicted an increase in wages by 9 percent, or one third of a standard deviation, compared to 5 percent for cognitive ability.

⁴A one standard deviation increase in the measure of noncognitive skills predicted a decrease in the probability of receiving employment support by 3.3 percentage points, compared to 1.1 percentage points for cognitive skills. Men with higher scores on the measure of noncognitive skills had shorter periods of unemployment, while cognitive ability had no statistically significant effect on the duration of unemployment.

⁵A one standard deviation increase in the measure of noncognitive skills predicted a

Murnane concluded his remarks by noting that he had focused his presentation on the relationships between 21st century skills and labor market outcomes, in part because labor market research provides a rich source of evidence about the importance of these skills. Nevertheless, he said he would argue that 21st century skills are needed in many aspects of life besides the workplace. As he put it, these skills are essential for “leading a contributing life in a pluralistic democracy.” He enumerated the complex set of problems that the country faces, including such issues as immigration, global warming, and proliferation of nuclear weapons. In his view, understanding these problems and participating in their solutions requires a well-educated citizenry adept at expert thinking and complex communication.

PREPARING STUDENTS FOR THE MODERN WORKPLACE

Are students graduating from high school with adequate preparation in these 21st century skills? At the workshop, representatives from the cosponsors of the project—the National Science Foundation (NSF) and the National Institutes of Health (NIH)—provided some insight on this issue. Gerhard Salinger, program director with the Directorate for Education and Human Resources at NSF, discussed his work with the advanced technological education program, an effort focused on technician education at the 2-year college level. This program is intended to educate students for middle skill jobs, occupations that require more than a secondary education but not necessarily 4 years of higher education. Middle skill jobs account for 50 percent of all jobs in the United States, Salinger said. He noted that the feedback he and his colleagues routinely receive from employers is that employees are lacking in 21st century skills. Furthermore, industry representatives have advised Salinger that these skills are not easily learned on the job. Based on his research and discussions with faculty members, Salinger judges that these skills are best learned in an academic setting. Salinger advocates for ensuring that students learn these skills before they leave high school. In part, this is because he believes that these are skills that everyone needs—not just for the workplace, higher education, or vocational/technical training—but for dealing with all aspects of life.

Bruce Fuchs, director of the Office of Science Education at NIH, presented data on the problem-solving skills of students in the United States. The Programme for International Student Assessment (PISA) has rou-

decrease in the probability that annual earnings fall short of the tenth percentile of the earnings distribution by 4.7 percentage points. The corresponding figure for cognitive ability fell from 1.5 to 0.2 percentage points.

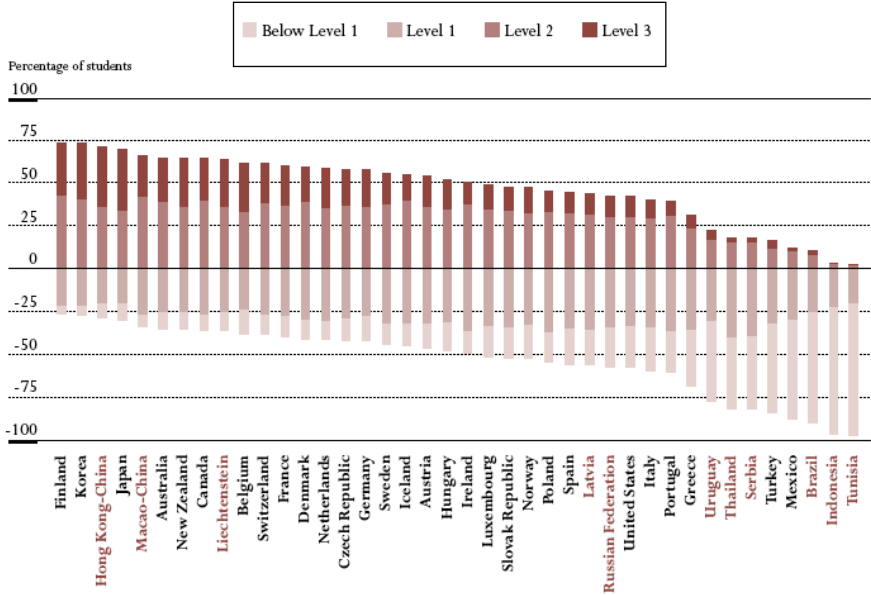


FIGURE 1-6 Percentage of students at each level of proficiency on the problem-solving scale of PISA 2003.
SOURCE: Organisation for Economic Co-operation and Development (2005). Problem Solving for Tomorrow's World: First Measures of Cross-Curricular Competencies from PISA 2003, <http://dx.doi.org/10.1787/9789264006430-en>.

tinely conducted assessments in mathematics, reading, and science. In 2003 an assessment of problem-solving skills was included. Fuchs said that he was “stunned” at the results for U.S. students, which he characterized as much lower than he had expected.

PISA results are reported using four performance levels: “Level 3” (highest), “Level 2,” “Level 1,” and “below Level 1” (lowest). Figure 1-6 shows the percentage of students from each participating country that scored at each performance level. The solid horizontal line at zero on the y-axis indicates the percentage of students at or below Level 1 (below the line) and at or above Level 2 (above the line). On the x-axis, the countries are ranked in descending order by the percentage of 15-year-olds in Levels 2 and 3. Fuchs highlighted three pieces of information on the graph. First, he noted that U.S. students rank ordered 29th compared to students in other countries. Second, he pointed out that 57 percent of the U.S. students taking the test scored below Level 2 (below the solid black line). Third, he called attention to the small percentage of students scoring

at Level 3, which he described as only one-third to one-half of that for the top scoring countries on the assessment.⁶

To exemplify the types of skills that are assessed, he described one of the items that was administered to the 15-year-olds taking the assessment. The item presented students with a map in which six fictional towns were noted (Kado, Lapat, Angaz, Megal, Piras, and Nuben), and a mileage legend that indicated the road distance of the towns from each other. The item presented students with two tasks:

- 1. Calculate the shortest distance by road between Nuben and Kado.
- 2. Zoe lives in Angaz. She wants to visit Kado and Lapat. She can only travel up to 300 kilometers in any one day but can break her journey by camping overnight anywhere between towns. Zoe will stay for two nights in each town so that she can spend one whole day sightseeing in each town. Show Zoe’s itinerary by completing the following table to indicate where she stays each night.

Day	Overnight Stay
1	Camp site between Angaz and Kado
2	
3	
4	
5	
6	
7	Angaz

Fuchs said that the sample item was one of the more complicated items on the assessment and was classified as a Level 3 item. Given that few of the U.S. students scored at a Level 3, most U.S. students would not have been able to answer this question correctly.

During discussion sessions, participants commented that the workshop was being held at an opportune time. Several commented about two reform movements currently underway. First, the National Gover-

⁶Results from more recent administrations of PISA are similar. For results from the 2006 assessment, see the Organisation for Economic Co-operation and Development (2007) *PISA 2006: Science Competencies for Tomorrow’s World Executive Summary*, available at <http://www.oecd.org/dataoecd/15/13/39725224.pdf> [July 2011]. For results from the 2009 assessment, see the Organisation for Economic Co-operation and Development (2010), *PISA 2009 Results: Executive Summary*, available at <http://www.oecd.org/dataoecd/34/60/46619703.pdf> [July 2011]. Also see, ACT (2011) *Affirming the Goal*, available at <http://www.act.org/research/policymakers/pdf/AffirmingtheGoal.pdf> [July 2011].

nors Association (NGA) and the Council of Chief State School Officers (CCSSO) have led an effort by the states to change the standards for educating K-12 students in reading and math. Known as the “Common Core Standards Initiative,” this effort is working first to identify the skills that students need and have all states in the country adopt these standards and second to develop assessments of these skills.⁷ Second, the Race to the Top initiative sponsored by the U.S. Department of Education is capitalizing on this effort in supporting consortia of states in their work to design assessments to measure these standards.⁸ The focus of both efforts is to ensure that students graduate from high school with skills that make them college and career ready. Participants also pointed out that the National Assessment of Educational Progress (NAEP) has been working to define and develop an assessment of college and career readiness, and assessing college readiness has been a prime focus of organizations such as ACT, the College Board, and the Educational Testing Service (ETS). Thus, there is considerable work underway on this issue.

Developing assessments of these skills was an issue that several participants highlighted as critical. As one workshop participant put it, “what is tested is taught and what is not tested is not taught.” Assessments often serve the purpose of defining the standards and laying out priorities for instruction. If assessments focus solely on students’ achievements in factual knowledge, this type of information will be the focus of teaching. To ensure that students acquire and show progress in 21st century skills, assessments need to be available to evaluate their performance in these areas. Participants noted that this should include assessments designed for both summative and formative uses.⁹ The remaining chapters of this report focus on developing assessments of these skills. Specifically: How can these skills be assessed? What assessments are currently available and how well do they work? What needs to be done in order to develop these types of assessments? And how should the results be used?

⁷Further information can be found at <http://www.corestandards.org/> [June 2011].

⁸Authorized under the American Recovery and Reinvestment Act of 2009 (ARRA), the Race to the Top Assessment Program provides funding to consortia of states to develop assessments that are valid, support and inform instruction, provide accurate information about what students know and can do, and measure student achievement against standards designed to ensure that all students gain the knowledge and skills needed to succeed in college and the workplace. (See <http://www2.ed.gov/programs/racetothetop-assessment/index.html> [May 2011].)

⁹See Chapter 5 for an explanation of formative and summative assessment.

2

Assessing Cognitive Skills

As described in Chapter 1, the steering committee grouped the five skills identified by previous efforts (National Research Council, 2008, 2010) into the broad clusters of cognitive skills, interpersonal skills, and intrapersonal skills. Based on this grouping, two of the identified skills fell within the cognitive cluster: nonroutine problem solving and systems thinking. The definition of each, as provided in the previous report (National Research Council, 2010, p. 3), appears below:

Nonroutine problem solving: A skilled problem solver uses expert thinking to examine a broad span of information, recognize patterns, and narrow the information to reach a diagnosis of the problem. Moving beyond diagnosis to a solution requires knowledge of how the information is linked conceptually and involves metacognition—the ability to reflect on whether a problem-solving strategy is working and to switch to another strategy if it is not working (Levy and Murnane, 2004). It includes creativity to generate new and innovative solutions, integrating seemingly unrelated information, and entertaining possibilities that others may miss (Houston, 2007).

Systems thinking: The ability to understand how an entire system works; how an action, change, or malfunction in one part of the system affects the rest of the system; adopting a “big picture” perspective on work (Houston, 2007). It includes judgment and decision making, systems analysis, and systems evaluation as well as abstract reasoning about how the different elements of a work process interact (Peterson et al., 1999).

After considering these definitions, the committee decided a third cognitive skill, critical thinking, was not fully represented. The committee added critical thinking to the list of cognitive skills, since competence in critical thinking is usually judged to be an important component of both skills (Mayer, 1990). Thus, this chapter focuses on assessments of three cognitive skills: problem solving, critical thinking, and systems thinking.

DEFINING THE CONSTRUCT

One of the first steps in developing an assessment is to define the construct and operationalize it in a way that supports the development of assessment tasks. Defining some of the constructs included within the scope of 21st century skills is significantly more challenging than defining more traditional constructs, such as reading comprehension or mathematics computational skills. One of the challenges is that the definitions tend to be both broad and general. To be useful for test development, the definition needs to be specific so that there can be a shared conception of the construct for use by those writing the assessment questions or preparing the assessment tasks.

This set of skills also generates debate about whether they are domain general or domain specific. A predominant view in the past has been that critical thinking and problem-solving skills are domain general: that is, that they can be learned without reference to any specific domain and, further, once they are learned, can be applied in any domain. More recently, psychologists and learning theorists have argued for a domain-specific conception of these skills, maintaining that when students think critically or solve problems, they do not do it in the absence of subject matter: instead, they think about or solve a problem in relation to some topic. Under a domain-specific conception, the learner may acquire these skills in one domain as he or she acquires expertise in that domain, but acquiring them in one domain does not necessarily mean the learner can apply them in another.

At the workshop, Nathan Kuncel, professor of psychology with University of Minnesota, and Eric Anderman, professor of educational psychology with Ohio State University, discussed these issues. The sections below summarize their presentations and include excerpts from their papers,¹ dealing first with the domain-general and domain-specific con-

¹For Kuncel's presentation, see http://www7.national-academies.org/bota/21st_Century_Workshop_Kuncel.pdf. For Kuncel's paper, see http://www7.national-academies.org/bota/21st_Century_Workshop_Kuncel_Paper.pdf. For Anderman's presentation, see http://www7.national-academies.org/bota/21st_Century_Workshop_Anderman.pdf. For Anderman's paper, see <http://nrc51/xpedio/groups/dbasse/documents/webpage/060387~1.pdf> [August 2011].

ceptions of critical thinking and problem solving and then with the issue of transferring skills from one domain to another.

Critical Thinking: Domain-Specific or Domain-General

It is well established, Kuncel stated, that foundational cognitive skills in math, reading, and writing are of central importance and that students need to be as proficient as possible in these areas. Foundational cognitive abilities, such as verbal comprehension and reasoning, mathematical knowledge and skill, and writing skills, are clearly important for success in learning in college as well as in many aspects of life. A recent study documents this. Kuncel and Hezlett (2007) examined the body of research on the relationships between traditional measures of verbal and quantitative skills and a variety of outcomes. The measures of verbal and quantitative skills included scores on six standardized tests—the GRE, MCAT, LSAT, GMAT, MAT, and PCAT.² The outcomes included performance in graduate school settings ranging from Ph.D. programs to law school, medical school, business school, and pharmacy programs. Figure 2-1 shows the correlations between scores on the standardized tests and the various outcome measures, including (from bottom to top) first-year graduate GPA (1st GGPA), cumulative graduate GPA (GGPA), qualifying or comprehensive examination scores, completion of the degree, estimate of research productivity, research citation counts, faculty ratings, and performance on the licensing exam for the profession. For instance, the top bar shows a correlation between performance on the MCAT and performance on the licensing exam for physicians of roughly .65, the highest of the correlations reported in this figure. The next bar indicates the correlation between performance on the LSAT and performance on the licensing exam for lawyers is roughly .35. Of the 34 correlations shown in the figure, all but 11 are over .30. Kuncel characterized this information as demonstrating that verbal and quantitative skills are important predictors of success based on a variety of outcome measures, including performance on standardized tests, whether or not people finish their degree program, how their performance is evaluated by faculty, and their contribution to the field.

Kuncel has also studied the role that broader abilities have in predicting future outcomes. A more recent review (Kuncel and Hezlett, 2010) examined the body of research on the relationships between measures of general cognitive ability (historically referred to as IQ) and job outcomes,

²Respectively, the Graduate Record Exam, Medical College Admission Test, Law School Admission Test, Graduate Management Admission Test, Miller Analogies Test, and Pharmacy College Admission Test.

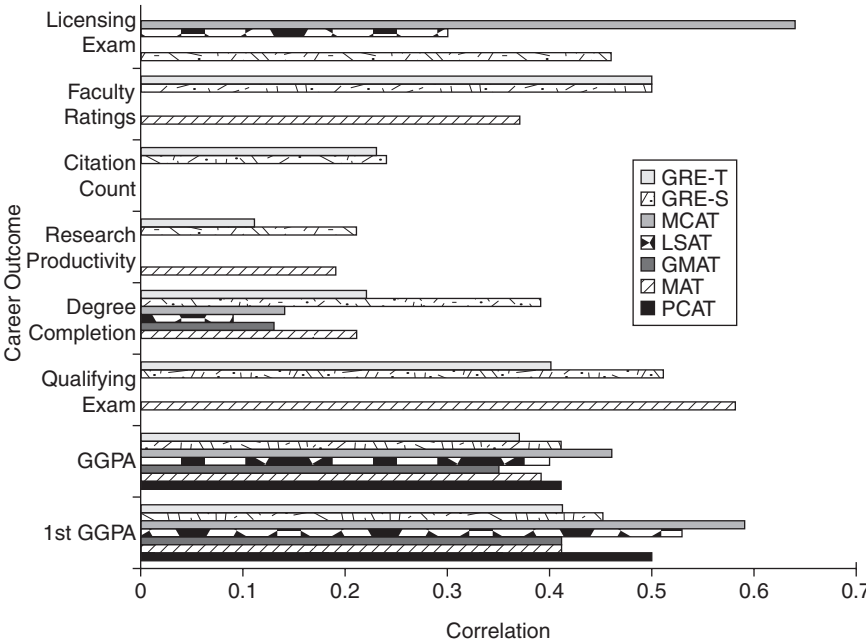


FIGURE 2-1 Correlations between scores on standardized tests and academic and job outcome measures.
 SOURCE: Kuncel and Hezlett (2007). Reprinted with permission of American Association for the Advancement of Science.

including performance in high, medium, and low complexity jobs; training success in civilian and military settings; how well leaders perform on objective measures; and evaluations of the creativity of people’s work. Figure 2-2 shows the correlations between performance on a measure of general cognitive ability and these outcomes. All of the correlations are above .30, which Kuncel characterized as demonstrating a strong relationship between general cognitive ability and job performance across a variety of performance measures. Together, Kuncel said, these two reviews present a body of evidence documenting that verbal and quantitative skills along with general cognitive ability are predictive of college and career performance.

Kuncel noted that other broader skills, such as critical thinking or analytical reasoning, may also be important predictors of performance, but he characterizes this evidence as inconclusive. In his view, the problems lie both with the conceptualization of the constructs as domain-general (as opposed to domain-specific) as well as with the specific definition of the construct. He finds the constructs are not well defined and have not

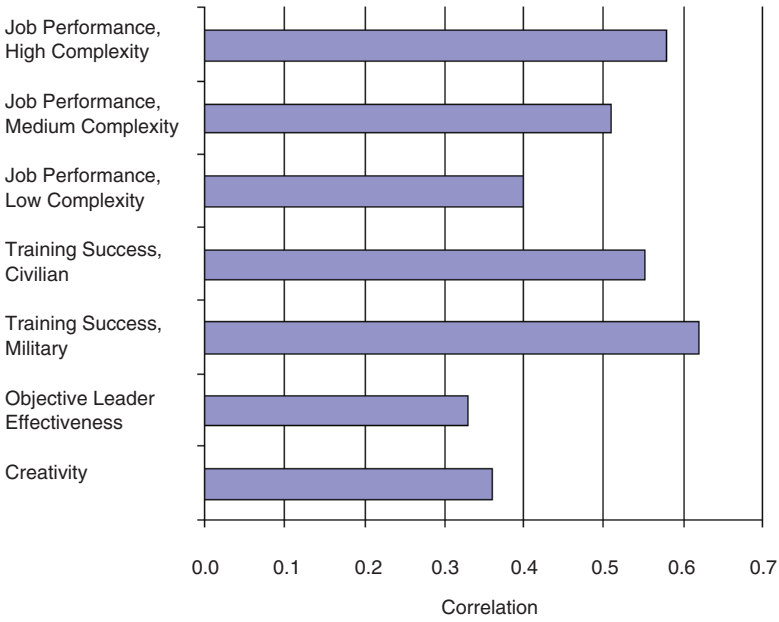


FIGURE 2-2 Correlations between measures of cognitive ability and job performance.
SOURCE: Kuncel and Hezlett (2011). Copyright 2010 by Sage Publications. Reprinted with permission of Sage Publications.

been properly validated. For instance, a domain-general concept of the construct of “critical thinking” is often indistinguishable from general cognitive ability or general reasoning and learning skills. To demonstrate, Kuncel presented three definitions of critical thinking that commonly appear in the literature:

1. “[Critical thinking involves] cognitive skills or strategies that increase the probability of a desirable outcome—in the long run, critical thinkers will have more desirable outcomes than ‘noncritical’ thinkers. . . . Critical thinking is purposeful, reasoned, and goal-directed. It is the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions” (Halpern, 1998, pp. 450-451).
2. “Critical thinking is reflective and reasonable thinking that is focused on deciding what to believe or do” (Ennis, 1985, p. 45).
3. “Critical thinking [is] the ability and willingness to test the validity of propositions” (Bangert-Drowns and Bankert, 1990, p. 3).

He characterized these definitions both very general and very broad. For instance, Halpern's definition essentially encompasses all of problem solving, judgment, and cognition, he said. Others are more specific and focus on a particular class of tasks (e.g., Bangert-Drowns and Bankert, 1990). He questioned the extent to which critical thinking so conceived is distinct from general cognitive ability (or general intelligence).

Kuncel conducted a review of the literature for empirical evidence of the validity of the construct of critical thinking. The studies in the review examined the relationships between various measures of critical thinking and measures of general intelligence and expert performance. He looked for two types of evidence—convergent validity evidence³ and discriminant validity⁴ evidence.

Kuncel found several analyses of the relationships among different measures of critical thinking (see Bondy et al., 2001; Facione, 1990; and Watson and Glaser, 1994). The assessments that were studied included the Watson-Glaser Critical Thinking Appraisal (WGCTA), the Cornell Critical Thinking Test (CCTT), the California Critical Thinking Skills Test (CCTST), and the California Critical Thinking Disposition Inventory (CCTDI). The average correlation among the measures was .41. Considering that all of these tests purport to be measures of the same construct, Kuncel judged this correlation to be low. For comparison, he noted a correlation of .71 between two subtests of the SAT intended to measure critical thinking (the SAT-critical reading test and the SAT-writing test).

With regard to discriminant validity, Kuncel conducted a literature search that yielded 19 correlations between critical-thinking skills and traditional measures of cognitive abilities, such as the Miller Analogies Test and the SAT (Adams et al., 1999; Bauer and Liang, 2003; Bondy et al., 2001; Cano and Martinez, 1991; Edwards, 1950; Facione et al., 1995, 1998; Spector et al., 2000; Watson and Glaser, 1994). He separated the studies into those that measured critical-thinking skills and those that measured critical-thinking dispositions (i.e., interest and willingness to use one's critical-thinking skills). The average correlation between gen-

³Convergent validity indicates the degree to which an operationalized construct is similar to other operationalized constructs that it theoretically should also be similar to. For instance, to show the convergent validity of a test of critical thinking, the scores on the test can be correlated with scores on other tests that are also designed to measure critical thinking. High correlations between the test scores would be evidence of convergent validity.

⁴Discriminant validity evaluates the extent to which a measure of an operationalized construct differs from measures of other operationalized constructs that it should differ from. In the present context, the interest is in verifying that critical thinking is a construct distinct from general intelligence and expert performance. Thus, discriminant validity would be examined by evaluating the patterns of correlations between and among scores on tests of critical thinking and scores on tests of the other two constructs (general intelligence and expert performance).

eral cognitive ability measures and critical-thinking skills was .48, and the average correlation between general cognitive ability measures and critical-thinking dispositions was .21.

Kuncel summarized these results as demonstrating that different measures of critical thinking show lower correlations with each other (i.e., average of .41) than they do with traditional measures of general cognitive ability (i.e., average of .48). Kuncel judges that these findings provide little support for critical thinking as a domain-general construct distinct from general cognitive ability. Given this relatively weak evidence of convergent and discriminant validity, Kuncel argued, it is important to determine if critical thinking is correlated differently than cognitive ability with important outcome variables like grades or job performance. That is, do measures of critical-thinking skills show incremental validity beyond the information provided by measures of general cognitive ability?

Kuncel looked at two outcome measures: grades in higher education and job performance. With regard to higher education, he examined data from 12 independent samples with 2,876 subjects (Behrens, 1996; Gadzella et al., 2002, 2004; Kowalski and Taylor, 2004; Taube, 1997; Williams, 2003). Across these studies, the average correlation between critical-thinking skills and grades was .27 and between critical-thinking dispositions and grades was .24. To put these correlations in context, the SAT has an average correlation with 1st year college GPA between .26 to .33 for the individual scales and .35 when the SAT scales are combined (Kobrin et al., 2008).⁵

There are very limited data that quantify the relationship between critical-thinking measures and subsequent job performance. Kuncel located three studies with the Watson-Glaser Appraisal (Facione and Facione, 1996, 1997; Giancarlo, 1996). They yielded an average correlation of .32 with supervisory ratings of job performance ($N = 293$).

Kuncel described these results as “mixed” but not supporting a conclusion that assessments of critical thinking are better predictors of college and job performance than other available measures. Taken together with the convergent and discriminant validity results, the evidence to support critical thinking as an independent construct distinct from general cognitive ability is weak.

Kuncel believes these correlational results do not tell the whole story, however. First, he noted, a number of artifactual issues may have contributed to the relatively low correlation among different assessments of critical thinking, such as low reliability of the measures themselves, restriction in range, different underlying definitions of critical thinking, overly broad

⁵It is important to note that when corrected for restriction in range, these coefficients increase to .47 to .51 for individual scores and .51 for the combined score.

definitions that are operationalized in different ways, different kinds of assessment tasks, and different levels of motivation in test takers.

Second, he pointed out, even though two tests correlate highly with each other, they may not measure the same thing. That is, although the critical-thinking tests correlate .48, on average, with cognitive ability measures, it does not mean that they measure the same thing. For example, a recent study (Kuncel and Grossbach, 2007) showed that ACT and SAT scores are highly predictive of nursing knowledge. But, obviously, individuals who score highly on a college admissions test do not have all the knowledge needed to be a nurse. The constructs may be related but not overlap entirely.

Kuncel explained that one issue with these studies is they all conceived of critical thinking in its broadest sense and as a domain-general construct. He said this conception is not useful, and he summarized his meta-analysis findings as demonstrating little evidence that critical thinking exists as a domain-general construct distinct from general cognitive ability. He highlighted the fact that some may view critical thinking as a specific skill that, once learned, can be applied in many situations. For instance, many in his field of psychology mention the following as specific critical-thinking skills that students should acquire: understanding the law of large numbers, understanding what it means to affirm the consequent, being able to make judgments about sample bias, understanding control groups, and understanding Type I versus Type II errors. However, Kuncel said many tasks that require critical thinking would not make use of any of these skills.

In his view, the stronger argument is for critical thinking as a domain-specific construct that evolves as the person acquires domain-specific knowledge. For example, imagine teaching general critical-thinking skills that can be applied across all reasoning situations to students. Is it reasonable, he asked, to think a person can think critically about arguments for different national economic policies without understanding macroeconomics or even the current economic state of the country? At one extreme, he argued, it seems clear that people cannot think critically about topics for which they have no knowledge, and their reasoning skills are intimately tied to the knowledge domain. For instance, most people have no basis for making judgments about how to conduct or even prioritize different experiments for CERN's Large Hadron Collider. Few people understand the topic of particle physics sufficiently to make more than trivial arguments or decisions. On the other hand, perhaps most people could try to make a good decision about which among a few medical treatments would best meet their needs.

Kuncel also talked about the kinds of statistical and methodological reasoning skills learned in different disciplines. For instance, chemists,

engineers, and physical scientists learn to use these types of skills in thinking about the laws of thermodynamics that deal with equilibrium, temperature, work, energy, and entropy. On the other hand, psychologists learn to use these skills in thinking about topics such as sample bias and self-selection in evaluating research findings. Psychologists who are adept at thinking critically in their own discipline would have difficulty thinking critically about problems in the hard sciences, unless they have specific subject matter knowledge in the discipline. Likewise, it is difficult to imagine that a scientist highly trained in chemistry could solve a complex problem in psychology without knowing some subject matter in psychology.

Kuncel said it is possible to train specific skills that aid in making good judgments in some situations, but the literature does not demonstrate that it is possible to train universally effective critical thinking skills. He noted, "I think you can give people a nice toolbox with all sorts of tools they can apply to a variety of tasks, problems, issues, decisions, citizenship questions, and learning those things will be very valuable, but I dissent on them being global and trainable as a global skill."

Transfer from One Context to Another

There is a commonplace assumption, Eric Anderman noted in his presentation, that learners readily transfer the skills they have learned in one course or context to situations and problems that arise in another. Anderman argued research on human learning does not support this assumption. Research suggests such transfer seldom occurs naturally, particularly when learners need to transfer complex cognitive strategies from one domain to another (Salomon and Perkins, 1989). Transfer is only likely to occur when care is taken to facilitate that transfer: that is, when students are specifically taught strategies that facilitate the transfer of skills learned in one domain to another domain (Gick and Holyoak, 1983).

For example, Anderman explained, students in a mathematics class might be taught how to solve a problem involving the multiplication of percentages (e.g., $4.79\% \times 0.25\%$). The students then might encounter a problem in their social studies courses that involves calculating compounded interest (such as to solve a problem related to economics or banking). Although the same basic process of multiplying percentages might be necessary to solve both problems, it is unlikely that students will naturally, on their own, transfer the skills learned in the math class to the problem encountered in the social studies class.

In the past, Anderman said, there had been some notion that critical-thinking and problem-solving skills could be taught independent of context. For example, teaching students a complex language such as Latin,

a computer programming language such as LOGO, or other topics that require complex thinking might result in an overall increase in their ability to think critically and problem solve.

Both Kuncel and Anderman maintained that the research does not support this idea. Instead, the literature better supports a narrower definition in which critical thinking is considered a finite set of specific skills. These skills are useful for effective decision making for many, but by no means all, tasks or situations. Their utility is further curtailed by task-specific knowledge demands. That is, a decision maker often has to have specific knowledge to make more than trivial progress with a problem or decision.

Anderman highlighted four important messages emerging from recent research. First, research documents that it is critical that students learn basic skills (such as basic arithmetic skills like times tables) so the skills become automatic. Mastery of these skills is required for the successful learning of more complex cognitive skills. Second, the use of general practices intended to improve students' thinking are not usually successful as a means of improving their overall cognitive abilities. The research suggests students may become more adept in the specific skill taught, but this does not transfer to an overall increase in cognitive ability. Third, when general problem-solving strategies are taught, they should be taught within meaningful contexts and not as simply rote algorithms to be memorized. Finally, educators need to actively teach students to transfer skills from one context to another by helping students to recognize that the solution to one type of problem may be useful in solving a problem with similar structural features (Mayer and Wittrock, 1996).

He noted that instructing students in general problem-solving skills can be useful but more elaborate scaffolding and domain-specific applications of these skills are often necessary. Whereas general problem-solving and critical-thinking strategies can be taught, research indicates these skills will not automatically or naturally transfer to other domains. Anderman stressed that educators and trainers must recognize that 21st century skills should be taught within specific domains; if they are taught as general skills, he cautioned, then extreme care must be taken to facilitate the transfer of these skills from one domain to another.

ASSESSMENT EXAMPLES

The workshop included examples of four different types of assessments of critical-thinking and problem-solving skills—one that will be used to make international comparisons of achievement, one used to license lawyers, and two used for formative purposes (i.e., intended to support instructional decision making). The first example was the com-

puterized problem-solving component of the Programme for International Student Assessment (PISA). This assessment is still under development but is scheduled for operational administration in 2012.⁶ Joachim Funke, professor of cognitive, experimental, and theoretical psychology with the Heidelberg University in Germany, discussed this assessment.

The second example was the Multistate Bar Exam, a paper-and-pencil test that consists of both multiple-choice and extended-response components. This test is used to qualify law students for practice in the legal profession. Susan Case, director of testing with the National Conference of Bar Exams, made this presentation.

The two formative assessments both make use of intelligent tutors, with assessments embedded into instruction modules. The “Auto Tutor” described by Art Graesser, professor of psychology with the University of Memphis, is used in instructing high school and higher education students in critical thinking skills in science. The Auto Tutor is part of a system Graesser has developed called Operation ARIES! (Acquiring Research Investigative and Evaluative Skills). The “Packet Tracer,” described by John Behrens, director of networking academy learning systems development with Cisco, is intended for individuals learning computer networking skills.

Problem Solving on PISA

For the workshop, Joachim Funke supplied the committee with the draft framework for PISA (see Organisation for Economic Co-operation and Development, 2010⁷) and summarized this information in his presentation.⁸ The summary below is based on both documents.

PISA, Funke explained, defines problem solving as an individual’s capacity to engage in cognitive processing to understand and resolve problem situations where a solution is not immediately obvious. The definition includes the willingness to engage with such situations in order to achieve one’s potential as a constructive and reflective citizen (Organisation for Co-operation and Development, 2010, p. 12). Further, the PISA 2012 assessment of problem-solving competency will not test simple reproduction of domain-based knowledge, but will focus on the cognitive skills required to solve unfamiliar problems encountered in life and lying outside traditional curricular domains. While prior knowledge

⁶For a full description of the PISA program, see http://www.oecd.org/pages/0,3417,en_32252351_32235731_1_1_1_1_1,00.html [August 2011].

⁷Available at <http://www.oecd.org/dataoecd/8/42/46962005.pdf> [August 2011].

⁸Available at http://www7.national-academies.org/bota/21st_Century_Workshop_Funke.pdf [August 2011].

is important in solving problems, problem-solving competency involves the ability to acquire and use new knowledge or to use old knowledge in a new way to solve novel problems. The assessment is concerned with nonroutine problems, rather than routine ones (i.e., problems for which a previously learned solution procedure is clearly applicable). The problem solver must actively explore and understand the problem and either devise a new strategy or apply a strategy learned in a different context to work toward a solution. Assessment tasks center on everyday situations, with a wide range of contexts employed as a means of controlling for prior knowledge in general.

The key domain elements for PISA 2012 are as follows:

- The problem context: whether it involves a technological device or not, and whether the focus of the problem is personal or social
- The nature of the problem situation: whether it is interactive or static (defined below)
- The problem-solving processes: the cognitive processes involved in solving the problem

The PISA 2012 framework (pp. 18-19) defines four processes that are components of problem solving. The first involves information retrieval. This process requires the test taker to quickly explore a given system to find out how the relevant variables are related to each other. The test taker must explore the situation, interact with it, consider the limitations or obstacles, and demonstrate an understanding of the given information. The objective is for the test taker to develop a mental representation of each piece of information presented in the problem. In the PISA framework, this process is referred to as exploring and understanding.

The second process is model building, which requires the test taker to make connections between the given variables. To accomplish this, the examinee must sift through the information, select the information that is relevant, mentally organize it, and integrate it with relevant prior knowledge. This requires the test taker to represent the problem in some way and formulate hypotheses about the relevant factors and their interrelationships. In the PISA framework, this dimension is called representing and formulating.

The third process is called forecasting and requires the active control of a given system. The framework defines this process as setting goals, devising a strategy to carry them out, and executing the plan. In the PISA framework, this dimension is called planning and executing.

The fourth process is monitoring and reflecting. The framework defines this process as checking the goal at each stage, detecting unexpected events, taking remedial action if necessary, and reflecting on solu-

tions from different perspectives by critically evaluating assumptions and alternative solutions.

Each of these processes requires the use of reasoning skills, which the framework describes as follows (Organisation for Economic Co-operation and Development, 2010, p. 19):

In understanding a problem situation, the problem solver may need to distinguish between facts and opinion, in formulating a solution, the problem solver may need to identify relationship between variables, in selecting a strategy, the problem solver may need to consider cause and effect, and in communicating the results, the problem solver may need to organize information in a logical manner. The reasoning skills associated with these processes are embedded within problem solving. They are important in the PISA context since they can be taught and modeled in classroom instruction (e.g., Adey et al., 2007; Klauer and Phye, 2008).

For any given test taker, the test lasts for 40 minutes. PISA is a survey-based assessment that uses a balanced rotation design. A total of 80 minutes of material is organized into four 20-minute clusters, with each student taking two clusters.

The items are grouped into units around a common stimulus that describes the problem. Reading and numeracy demands are kept to a minimum. The tasks all consist of authentic stimulus items, such as refueling a moped, playing on a handball team, mixing a perfume, feeding cats, mixing elements in a chemistry lab, taking care of a pet, and so on. Funke noted that the different contexts for the stimuli are important because test takers might be motivated differentially and might be differentially interested depending on the context. The difficulty of the items is manipulated by increasing the number of variables or the number of relations that the test taker has to deal with.

PISA 2012 is a computer-based test in which items are presented by computer and test takers respond on the computer. Approximately three-quarters of the items are in a format that the computer can score (simple or complex multiple-choice items). The remaining items are constructed-response, and test takers enter their responses into text boxes.

Scoring of the items is based on the processes that the test taker uses to solve the problem and involves awarding points for the use of certain processes. For information retrieval, the focus is on identifying the need to collect baseline data (referred to in PISA terminology as identifying the “zero round”) and the method of manipulating one variable at a time (referred to in PISA terminology as “varying one thing at a time” or VOTAT). Full credit is awarded if the subject uses VOTAT strategy and makes use of zero rounds. Partial credit is given if the subject uses VOTAT but does not make use of zero rounds.

For model building, full credit is awarded if the generated model is correct. If one or two errors are present in the model, partial credit is given. If more than two errors are present, then no credit is awarded.

For forecasting, full credit is given if the target goals are reached. Partial credit is given if some progress toward the target goals can be registered, and no credit is given if there is no progress toward target goals at all.

PISA items are classified as static versus interactive. In static problems, all the information the test taker needs to solve the problem is presented at the outset. In contrast, interactive problems require the test taker to explore the problem to uncover important relevant information (Organisation for Economic Co-operation and Development, 2010, p. 15). Two sample PISA items appear in Box 2-1.

Funke and his colleagues have conducted analyses to evaluate the construct validity of the assessment. They have examined the internal structure of the assessment using structural equation modeling, which evaluates

BOX 2-1 **Sample Problem-Solving Items for PISA 2012**

Digital Watch—interactive:

A simulation of a digital watch is presented. The watch is controlled by four buttons, the functions of which are unknown to the student at the outset of the problems. The student is required to (Q1) determine through guided exploration how the buttons work in TIME mode, (Q2) complete a diagram showing how to cycle through the various modes, and (Q3) use this knowledge to control the watch (set the time).

Q1 is intended to measure exploring and understanding, Q2 measures representing and formulating, Q3 measures planning and executing.

Basketball—static

The rules for a basketball tournament relating to the way in which match time should be distributed between players are given. There are two more players than required (5) and each player must be on court for at least 25 of the 40 minutes playing time. Students are required to (Q1) create a schedule for team members that satisfies the tournament rules, and (Q2) reflect on the rules by critiquing an existing schedule.

Q1 is designed to measure planning and executing, Q2 measures monitoring and reflecting.

SOURCE: Organisation for Economic Co-operation and Development (2010, p. 28). Reprinted with permission of Organisation for Economic Co-operation and Development.

the extent to which the items measure the dimensions they are intended to measure. The results indicate the three dimensions are correlated with each other. Model Building and Forecasting correlate at .77; Forecasting and Information Retrieval correlate at .71; and Information Retrieval and Model Building correlate at .75. Funke said that the results also document that the items “load on” the three dimensions in the way the test developers hypothesized. He indicated some misfit related to the items that measure Forecasting, and he attributes this to the fact that the Forecasting items have a skewed distribution. However, the fit of the model does not change when these items are removed.

Funke reported results from studies of the relationship between test performance and other variables, including school achievement and two measures of problem solving on the PISA German National Extension on Complex Problem Solving. The latter assessment, called HEIFI, measures knowledge about a system and the control of the system separately. Scores on the PISA Model Building dimension are statistically significant ($p < .05$) related to school achievement ($r = .64$) and to scores on the HEIFI knowledge component ($r = .48$). Forecasting is statistically significant ($p < .05$) related to both of the HEIFI scores ($r = .48$ for HEIFI knowledge and $r = .36$ for HEIFI control). Information Retrieval is statistically significant ($p < .05$) related to HEIFI control ($r = .38$). The studies also show that HEIFI scores are not related to school achievement.

Funke closed by discussing the costs associated with the assessment. He noted it is not easy to specify the costs because in a German university setting, many costs are absorbed by the department and its equipment. Funke estimates that development costs run about \$13 per unit,⁹ plus \$6.5 for the Cognitive Labs used to pilot test and refine the items.¹⁰ The license for the Computer Based Assessment (CBA) Item-builder and the execution environment is given for free for scientific use from DIPF¹¹ Frankfurt.

The Bar Examination for Lawyers¹²

The Bar examination is administered by each jurisdiction in the United States as one step in the process to license lawyers. The National Council of Bar Examiners (NCBE) develops a series of three exams for use by the jurisdictions. Jurisdictions may use any or all of these three

⁹A unit consists of stimulus materials, instructions, and the associated questions.

¹⁰Costs are in American dollars.

¹¹DIPF stands for the Deutsches Institut für Internationale Pädagogische Forschung, which translates to the German Institute for Educational Research and Educational Information.

¹²The summary is based on a presentation by Susan Case, see http://www7.national-academies.org/bota/21st_Century_Workshop_Case.pdf [August 2011].

exams or may administer locally developed exam components if they wish. The three major components developed by the NCBE include the Multi-state Bar Exam (MBE), the Multi-state Essay Exam (MEE), and the Multi-state Performance Test (MPT). All are paper-and-pencil tests. Examinees pay to take the test, and the costs are \$54 for the MBE, \$20 for the MEE, and \$20 for the MPT.

Susan Case, who has spent her career working on licensing exams—first the medical licensing exam for physicians and then the bar exam for lawyers—noted the Bar examination is like other tests used to award professional licensure. The focus of the test is on the extent to which the test taker has the knowledge and skills necessary to be licensed in the profession on the day of the test. The test is intended to ensure the newly licensed professional knows what he/she needs to know to practice law. The test is not designed to measure the curriculum taught in law schools, but what licensed professionals need to know. When they receive the credential, lawyers are licensed to practice in all fields of law. This is analogous to medical licensing in which the licensed professional is eligible to practice any kind of medicine.

The Bar exam includes both multiple-choice and constructed-response components. Both require examinees to be able to gather and synthesize information and apply their knowledge to the given situation. The questions generally follow a vignette that describes a case or problem and asks the examinee to determine the issues to resolve before advising the client or to determine other information needed in order to proceed. For instance, what questions should be asked next? What is the best strategy to implement? What is the best defense? What is the biggest obstacle to relief? The questions may require the examinee to synthesize the law and the facts to predict outcomes. For instance, is the ordinance constitutional? Should a conviction be overturned?

The MBE

The purpose of the MBE is to assess the extent to which an examinee can apply fundamental legal principles and legal reasoning to analyze a given pattern of facts. The questions focus on the understanding of legal principles rather than memorization of local case or statutory law. The MBE consists of 60 multiple-choice questions and lasts a full day.

A sample question follows:

A woman was told by her neighbor that he planned to build a new fence on his land near the property line between their properties. The woman said that, although she had little money, she would contribute something toward the cost. The neighbor spent \$2,000 in materials and a day of his time to construct the fence. The neighbor now wants her to pay half the cost of the materials. Is she liable for this amount?

The MEE

The purpose of the MEE is to assess the examinee's ability to (1) identify legal issues raised by a hypothetical factual situation; (2) separate material that is relevant from that which is not; (3) present a reasoned analysis of the relevant issues in a clear, concise, and well-organized composition; and (4) demonstrate an understanding of the fundamental legal principles relevant to the probable resolution of the issues raised by the factual situation.

The MEE lasts for 6 hours and consists of nine 30-minute questions. An excerpt from a sample question follows:

The CEO/chairman of the 12-member board of directors (the Board) of a company plus three other members of the Board are senior officers of the company. The remaining eight members of the Board are wholly independent directors.

Recently, the Board decided to hire a consulting firm to market a new product . . .

The CEO disclosed to the Board that he had a 25% partnership interest in the consulting firm. The CEO stated that he would not be involved in any work to be performed by the consulting firm. He knew but did not disclose to the Board that the consulting firm's proposed fee for this consulting assignment was substantially higher than it normally charged for comparable work . . .

The Board discussed the relative merits of the two proposals for 10 minutes. The Board then voted unanimously (CEO abstaining) to hire the consulting firm . . .

1. Did the CEO violate his duty of loyalty to his company? Explain.
2. Assuming the CEO breached his duty of loyalty to his company, does he have any defense to liability? Explain.
3. Did the other directors violate their duty of care? Explain.

The MPT

The purpose of the MPT is to assess fundamental lawyering skills in realistic situations by asking the candidate to complete a task that a beginning lawyer should be able to accomplish. The MPT requires applicants to sort detailed factual materials; separate relevant from irrelevant facts; analyze statutory, case, and administrative materials for relevant principles of law; apply relevant law to the facts in a manner likely to resolve a client's problem; identify and resolve ethical dilemmas; communicate effectively in writing; and complete a lawyering task within time constraints.

Each task is completely self-contained and includes a file, a library, and a task to complete. The task might deal with a car accident, for

example, and therefore might include a file with pictures of the accident scene and depositions from the various witnesses, as well as a library with relevant case law. Examinees are given 90 minutes to complete each task.

For example, in a case involving a slip and fall in a store, the task might be to prepare an initial draft of an early dispute resolution for a judge. The draft should candidly discuss the strengths and weaknesses of the client's case. The file would contain the instructional memo from the supervising attorney, the local rule, the complaint, an investigator's report, and excerpts of the depositions of the plaintiff and a store employee. The library would include a jury instruction concerning the premises liability with commentary on contributory negligence.

Scoring

The MBE is a multiple-choice test and thus scored by machine. However, the other two components require human scoring. The NCBE produces the questions and the grading guidelines for the MEE and MPT, but the essays and performance tests are scored by the jurisdictions themselves. The scorers are typically lawyers who are trained during grading seminars held at the NCBE offices, after the exam is administered. At this time, they review sample papers and receive training on how to apply the scoring guidelines in a consistent fashion.

Each component of the Bar examination (MBE, MEE, MPT) is intended to assess different skills. The MBE focuses on breadth of knowledge, the MEE focuses on depth of knowledge, and the MPT focuses on the ability to demonstrate practical skills. Together, the three formats cover the different types of tasks that a new lawyer needs to do.

Determinations about weighting the three components are left to the jurisdictions; however, the NCBE urges them to weight the MBE score by 50 percent and the MEE and MPT by 25 percent each. The recommendation is an attempt to balance a number of concerns, including authenticity, psychometric considerations, logistical issues, and economic concerns. The recommendation is to award the highest weight to the MBE because it is the most psychometrically sound. The reliability of scores on the MBE is generally over .90, much higher than scores on the other portions, and the MBE is scaled and equated across time. The recommended weighting helps to ensure high decision consistency and comparability of pass/fail decisions across administrations.

Currently the MBE is used by all but three jurisdictions (Louisiana, Washington, and Puerto Rico). The essay exam is used by 27 jurisdictions, and the performance test is used by 34 jurisdictions.

Test Development

Standing test development committees that include practicing lawyers, judges, and lawyers on staff with law schools write the test questions. The questions are reviewed by outside experts, pretested on appropriate populations, analyzed and revised, and professionally edited before operational use. Case said the test development procedures for the Bar exam are analogous to those used for the medical licensure exams.

Operation ARIES! (Acquiring Research Investigative and Evaluative Skills)

The summary below is based on materials provided by Art Graesser, including his presentation¹³ and two background papers he supplied to the committee (Graesser et al., 2010; Millis et al., in press).

Operation ARIES! is a tutorial system with a formative assessment component intended for high school and higher education students, Graesser explained. It is designed to teach and assess critical thinking about science. The program operates in a game environment intended to be engaging to students. The system includes an “Auto Tutor,” which makes use of animated characters that converse with students. The Auto Tutor is able to hold conversations with students in natural language, interpret the student’s response, and respond in a way that is adaptive to the student’s response. The designers have created a science fiction setting in which the game and exercises operate. In the game, alien creatures called “Fuaths” are disguised as humans. The Fuaths disseminate bad science through various media outlets in an attempt to confuse humans about the appropriate use of the scientific method. The goal for the student is to become a “special agent of the Federal Bureau of Science (FBS), an agency with a mission to identify the Fuaths and save the planet” (Graesser et al., 2010, p. 328).

The system addresses scientific inquiry skills, developing research ideas, independent and dependent variables, experimental control, the sample, experimenter bias, and relation of data to theory. The focus is on use of these skills in the domains of biology, chemistry, and psychology. The system helps students to learn to evaluate evidence intended to support claims. Some examples of the kinds of research questions/claims that are evaluated include the following:

¹³For Graesser’s presentation, see <http://nrc51/xpedio/groups/dbasse/documents/webpage/060267~1.pdf> [August 2011].

From Biology:

- Do chemical and organic pesticides have different effects on food quality?
- Does milk consumption increase bone density?

From Chemistry:

- Does a new product for winter roads prevent water from freezing?
- Does eating fish increase blood mercury levels?

From Psychology:

- Does using cell phones hurt driving?
- Is a new cure for autism effective?

The system includes items in real-life formats, such as articles, advertisements, blogs, and letters to the editor, and makes use of different types of media where it is common to see faulty claims.

Through the system, the student encounters a story told by video, combined with communications received by e-mail, text message, and updates. The student is engaged through the Auto Tutor, which involves a “tutor agent” that serves as a narrator, and a “student agent” that serves in different roles, depending on the skill level of the student.

The system makes use of three kinds of modules—interactive training, case studies, and interrogations. The interactive training exchanges begin with the student reading an e-book, which provides the requisite information used in later modules. After each chapter, the student responds to a set of multiple-choice questions intended to assess the targeted skills. The text is interactive in that it involves “trialogs” (three-way conversations) between the primary agent, the student agent, and the actual (human) student. It is adaptive in that the strategy used is geared to the student’s performance. If the student is doing poorly, the two auto-tutor agents carry on a conversation that promotes vicarious learning: that is, the tutor agent and the student agent interact with each other, and the human student observes. If the student is performing at an intermediate level, normal tutoring occurs in which the student carries on a conversational exchange with the tutor agent. If the student is doing very well, he or she may be asked to teach the student agent, under the notion that the act of teaching can help to perfect one’s skills.

In the case study modules, the student is expected to apply what he or she has learned. The case study modules involve some type of flawed science, and the student is to identify the flaws by applying information learned from the interactive text in the first module. The student responds by verbally articulating the flaws, and the system makes use of advances in computational linguistics to analyze the meaning of the response. The researchers adopted the case study approach because it “allows learners to encode and discover the rich source of constraints and interdependen-

cies underlying the target elements (flaws) within the cases. [Prior] cases provide a knowledge base for assessing new cases and help guide reasoning, problem solving, interpretation and other cognitive processes" (Millis et al., in press, p. 17).

In the interrogation modules, insufficient information is provided, so students must ask questions. Research is presented in an abbreviated fashion, such as through headlines, advertisements, or abstracts. The student is expected to identify the relevant questions to ask and to learn to discriminate good research from flawed research. The storyline is advanced by e-mails, dialogues, and videos that are interspersed among the learning activities.

Through the three kinds of modules, the system interweaves a variety of key principles of learning that Graesser said have been shown to increase learning. These include

- Self-explanation (where the learner explains the material to another student, such as the automated student)
- Immediate feedback (through the tutoring system)
- Multimedia effects (which tend to engage the student)
- Active learning (in which students actually participate in solving a problem)
- Dialog interactivity (in which students learn by engaging in conversations and tutorial dialogs)
- Multiple, real-life examples (intended to help students transfer what they learn in one context to another context and to real world situations)

Graesser closed by saying that he and his colleagues are beginning to collect data from evaluation studies to examine the effects of the Auto Tutor. Research has focused on estimating changes in achievement before and after use of the system, and, to date, the results are promising.

Packet Tracer

The summary below is based on materials provided by John Behrens, including his presentation¹⁴ and a background paper he forwarded in preparation for the workshop (Behrens et al., in press).

To help countries around the world train their populations in networking skills, Cisco created the Networking Academy. The academy is a public/private partnership through which Cisco provides free online

¹⁴For Behrens' presentation, see http://www7.national-academies.org/bota/21st_Century_Workshop_Behrens.pdf [August 2011].

curricula and assessments. Behrens pointed out that in order to become adept with networking, students need both a conceptual understanding of networking and the skills to apply this knowledge to real situations. Thus, hands-on practice and assessment on real equipment are important components of the academy's instructional program. Cisco also wants to provide students with time for out-of-class practice and opportunities to explore on their own using online equipment that is not typically available in the average classroom setting. In the Networking Academy, students work with an online instructor, and they proceed through an established curriculum that incorporates numerous interactive activities.

Behrens talked specifically about a new program Cisco has developed called "Packet Tracer," a computer package that uses simulations to provide instruction and includes an interactive and adaptable assessment component. Cisco has incorporated Packet Tracer activities into the curricula for training networking professionals. Through this program, instructors and students can construct their own activities, and students can explore problems on their own. In Cisco's Networking Academy, assessments can be student-initiated or instructor-initiated. Student-initiated assessments are primarily embedded in the curriculum and include quizzes, interactive activities, and "challenge labs," which are a feature of Packet Tracer. The student-initiated assessments are designed to provide feedback to the student to help his or her learning. They use a variety of technologies ranging from multiple-choice questions (in the quizzes) to complex simulations (in the challenge labs). Before the development of Packet Tracer, the instructor-initiated assessments consisted either of hands-on exams with real networking equipment or multiple-choice exams in the online assessment system. Packet Tracer provides more simulation-based options, and also includes detailed reporting and grade-book integration features.

Each assessment consists of one extensive network configuration or troubleshooting activity that may require up to 90 minutes to complete. Access to the assessment is associated with a particular curricular unit, and it may be re-accessed repeatedly based on instructor authorization. The system provides simulations of a broad range of networking devices and networking protocols, including features set around the Cisco IOS (Internet Operating System). Instructions for tasks can be presented through HTML-formatted text boxes that can be preauthored, stored, and made accessible by the instructor at the appropriate time.

Behrens presented an example of a simulated networking problem in which the student needs to obtain the appropriate cable. To complete this task, the student must determine what kind of cable is needed, where on the computer to plug it in, and how to connect it. The student's performance is scored, and his or her interactions with the problem are

tracked in a log. The goal is not to simply assign a score to the student's performance but to provide detailed feedback to enhance learning and to correct any misinterpretations. The instructors can receive and view the log in order to evaluate how well the student understands the tasks and what needs to be done.

Packet Tracer can simulate a broad range of devices and networking protocols, including a wide range of PC facilities covering communication cards, power functionality, web browsers, and operating system configurations. The particular devices, configurations, and problem states are determined by the author of the task (e.g., the instructor) in order to address whatever proficiencies the chapter, course, or instruction targets. When icons of the devices are touched in the simulator, more detailed pictures are presented with which the student can interact. The task author can program scoring rules into the system. Students can be observed trying and discarding potential solutions based on feedback from the game resulting in new understandings. The game encourages students to engage in problem-solving steps (such as problem identification, solution generation, and solution testing). Common incorrect strategies can be seen across recordings.

3

Assessing Interpersonal Skills

The second cluster of skills—broadly termed interpersonal skills—are those required for relating to other people. These sorts of skills have long been recognized as important for success in school and the workplace, said Stephen Fiore, professor at the University of Central Florida, who presented findings from a paper about these skills and how they might be assessed (Salas, Bedwell, and Fiore, 2011).¹ Advice offered by Dale Carnegie in the 1930s to those who wanted to “win friends and influence people,” for example, included the following: be a good listener; don’t criticize, condemn, or complain; and try to see things from the other person’s point of view. These are the same sorts of skills found on lists of 21st century skills today. For example, the Partnership for 21st Century Skills includes numerous interpersonal capacities, such as working creatively with others, communicating clearly, and collaborating with others, among the skills students should learn as they progress from preschool through postsecondary study (see Box 3-1 for the definitions of the relevant skills in the organization’s P-21 Framework).

It seems clear that these are important skills, yet definitive labels and definitions for the interpersonal skills important for success in schooling and work remain elusive: They have been called social or people skills, social competencies, soft skills, social self-efficacy, and social intelligence, Fiore said (see, e.g., Ferris, Witt, and Hochwarter, 2001; Hochwarter et al.,

¹See http://www7.national-academies.org/bota/21st_Century_Workshop_Salas_Fiore_Paper.pdf [August 2011].

BOX 3-1 Interpersonal Capacities in the Partnership for 21st Century Skills Framework

Work Creatively with Others

- Develop, implement, and communicate new ideas to others effectively
- Be open and responsive to new and diverse perspectives; incorporate group input and feedback into the work
- Demonstrate originality and inventiveness in work and understand the real-world limits to adopting new ideas
- View failure as an opportunity to learn; understand that creativity and innovation is a long-term, cyclical process of small successes and frequent mistakes

Communicate Clearly

- Articulate thoughts and ideas effectively using oral, written, and nonverbal communication skills in a variety of forms and contexts
- Listen effectively to decipher meaning, including knowledge, values, attitudes, and intentions
- Use communication for a range of purposes (e.g., to inform, instruct, motivate, and persuade)
- Utilize multiple media and technologies, and know how to judge their effectiveness a priori as well as to assess their impact
- Communicate effectively in diverse environments (including multilingual)

Collaborate with Others

- Demonstrate ability to work effectively and respectfully with diverse teams
- Exercise flexibility and willingness to be helpful in making necessary compromises to accomplish a common goal
- Assume shared responsibility for collaborative work, and value the individual contributions made by each team member

2006; Klein et al., 2006; Riggio, 1986; Schneider, Ackerman, and Kanfer, 1996; Sherer et al., 1982; Sternberg, 1985; Thorndike, 1920). The previous National Research Council (NRC) workshop report that offered a preliminary definition of 21st century skills described one broad category of interpersonal skills (National Research Council, 2010, p. 3):

Complex communication/social skills: Skills in processing and interpreting both verbal and nonverbal information from others in order to respond appropriately. A skilled communicator is able to select key pieces of a complex idea to express in words, sounds, and images, in order to build shared understanding (Levy and Murnane, 2004). Skilled communicators negotiate positive outcomes with customers, subordinates, and superiors through social perceptiveness, persuasion, negotiation, instructing, and service orientation (Peterson et al., 1999).

Adapt to Change

- Adapt to varied roles, jobs responsibilities, schedules, and contexts
- Work effectively in a climate of ambiguity and changing priorities

Be Flexible

- Incorporate feedback effectively
- Deal positively with praise, setbacks, and criticism
- Understand, negotiate, and balance diverse views and beliefs to reach workable solutions, particularly in multicultural environments

Interact Effectively with Others

- Know when it is appropriate to listen and when to speak
- Conduct themselves in a respectable, professional manner

Work Effectively in Diverse Teams

- Respect cultural differences and work effectively with people from a range of social and cultural backgrounds
- Respond open-mindedly to different ideas and values
- Leverage social and cultural differences to create new ideas and increase both innovation and quality of work

Guide and Lead Others

- Use interpersonal and problem-solving skills to influence and guide others toward a goal
- Leverage strengths of others to accomplish a common goal
- Inspire others to reach their very best via example and selflessness
- Demonstrate integrity and ethical behavior in using influence and power

Be Responsible to Others

- Act responsibly with the interests of the larger community in mind

SOURCE: Excerpted from P21 Framework Definitions, Partnership for 21st Century Skills December 2009 [copyrighted—available at http://www.p21.org/index.php?option=com_content&task=view&id=254&Itemid=120 [August 2011].

These and other available definitions are not necessarily at odds, but in Fiore's view, the lack of a single, clear definition reflects a lack of theoretical clarity about what they are, which in turn has hampered progress toward developing assessments of them. Nevertheless, appreciation for the importance of these skills—not just in business settings, but in scientific and technical collaboration, and in both K-12 and postsecondary education settings—has been growing. Researchers have documented benefits these skills confer, Fiore noted. For example, Goleman (1998) found they were twice as important to job performance as general cognitive ability. Sonnentag and Lange (2002) found understanding of cooperation strategies related to higher performance among engineering and software development teams, and Nash and colleagues (2003) showed that collaboration skills were key to successful interdisciplinary research among scientists.

WHAT ARE INTERPERSONAL SKILLS?

The multiplicity of names for interpersonal skills and ways of conceiving of them reflects the fact that these skills have attitudinal, behavioral, and cognitive components, Fiore explained. It is useful to consider 21st century skills in basic categories (e.g., cognitive, interpersonal, and intrapersonal), but it is still true that interpersonal skills draw on many capacities, such as knowledge of social customs and the capacity to solve problems associated with social expectations and interactions. Successful interpersonal behavior involves a continuous correction of social performance based on the reactions of others, and, as Richard Murnane had noted earlier, these are cognitively complex tasks. They also require self-regulation and other capacities that fall into the intrapersonal category (discussed in Chapter 4). Interpersonal skills could also be described as a form of “social intelligence,” specifically social perception and social cognition that involve processes such as attention and decoding. Accurate assessment, Fiore explained, may need to address these various facets separately.

The research on interpersonal skills has covered these facets, as researchers who attempted to synthesize it have shown. Fiore described the findings of a study (Klein, DeRouin, and Salas, 2006) that presented a taxonomy of interpersonal skills based on a comprehensive review of the literature. The authors found a variety of ways of measuring and categorizing such skills, as well as ways to link them both to outcomes and to personality traits and other factors that affect them. They concluded that interpersonal effectiveness requires various sorts of competence that derive from experience, instinct, and learning about specific social contexts. They put forward their own definition of interpersonal skills as “goal-directed behaviors, including communication and relationship-building competencies, employed in interpersonal interaction episodes characterized by complex perceptual and cognitive processes, dynamic verbal and non verbal interaction exchanges, diverse roles, motivations, and expectancies” (p. 81).

They also developed a model of interpersonal performance, shown in Figure 3-1, that illustrates the interactions among the influences, such as personality traits, previous life experiences, and the characteristics of the situation; the basic communication and relationship-building skills the individual uses in the situation; and outcomes for the individual, the group, and the organization. To flesh out this model, the researchers distilled sets of skills for each area, as shown in Table 3-1.

Fiore explained that because these frameworks focus on behaviors intended to attain particular social goals and draw on both attitudes and cognitive processes, they provide an avenue for exploring what goes into the development of effective interpersonal skills in an individual. They

TABLE 3-1 Taxonomy of Interpersonal Skills

Interpersonal Skill	Description	Related Skills
Communication Skills		
Active Listening	Paying close attention to what is being said, asking the other party to explain exactly what he or she means, and requesting that ambiguous ideas or statements are repeated	Listening with empathy and sympathy; listening for understanding
Oral Communication	Sending verbal messages constructively	Enunciating; expressing yourself clearly; communicating emotion; interpersonal communication
Written Communication	Writing clearly and appropriately	Clarity; communicating intended meaning
Assertive Communication	Directly expressing one’s feelings, preferences, needs, and opinions in a way that is neither threatening nor punishing to another person	Proposing ideas; social assertiveness; defense of rights; directive; asserting your needs
Nonverbal Communication	Reinforcing or replacing spoken communication through the use of body language, gestures, voice, or artifacts	Expression of feelings; perception/recognition of feelings; facial regard
Relationship-Building Skills		
Cooperation and Coordination	Understanding and working with others in groups or teams; includes offering help to those who need it and pacing activities to fit the needs of the team	Adaptability; shared situational awareness; performance monitoring and feedback; interpersonal relations; communication; decision making; cohesion; group problem solving; being a team player
Trust	An individual’s faith or belief in the integrity or reliability of another person or thing; willingness of a party to be vulnerable to the actions of another party based on the expectation that certain actions important to the trustor will be performed	Self-awareness; self-disclosure; swift trust

continued

TABLE 3-1 Continued

Interpersonal Skill	Description	Related Skills
<i>Intercultural Sensitivity</i>	Appreciating individual differences among people	Acceptance; openness to new ideas; sensitivity to others; cross-cultural relations
<i>Service Orientation</i>	A set of basic individual predispositions and an inclination to provide service, to be courteous and helpful in dealing with customers, clients, and associates	Exceeding customer's expectations; customer satisfaction skills; ability to maintain positive client relationship; selling; building rapport; representing the organization to customers and the public
<i>Self-Presentation</i>	Process by which individuals attempt to influence the reactions and images people have of them and their ideas; managing these impressions encompasses a wide range of behaviors designed to create a positive influence on work associates	Self-expression; face-saving and impression management; managing perceptions; self-promotion
<i>Social Influence</i>	Guiding people toward the adoption of specific behaviors, beliefs, or attitudes; influencing the distribution of advantages and disadvantages within an organization through one's actions	Business etiquette; reasoning; friendliness; coalition building; bargaining; appeals to higher authority; imposing sanctions; networking; persuasion, positive political skills
<i>Conflict Resolution and Negotiation</i>	Advocating one's position with an open mind, not taking personally other members' disagreements, putting oneself in the other's shoes, following rational argument and avoiding premature evaluation, and trying to synthesize the best ideas from all viewpoints and perspectives	Conflict-handling style; conflict management; conflict prevention; compromising; problem solving; integrative bargaining; principled negotiation; cultural negotiation; mediation

SOURCE: Klein, DeRouin, and Salas (2006). Reprinted with permission of John Wiley & Sons, Ltd.

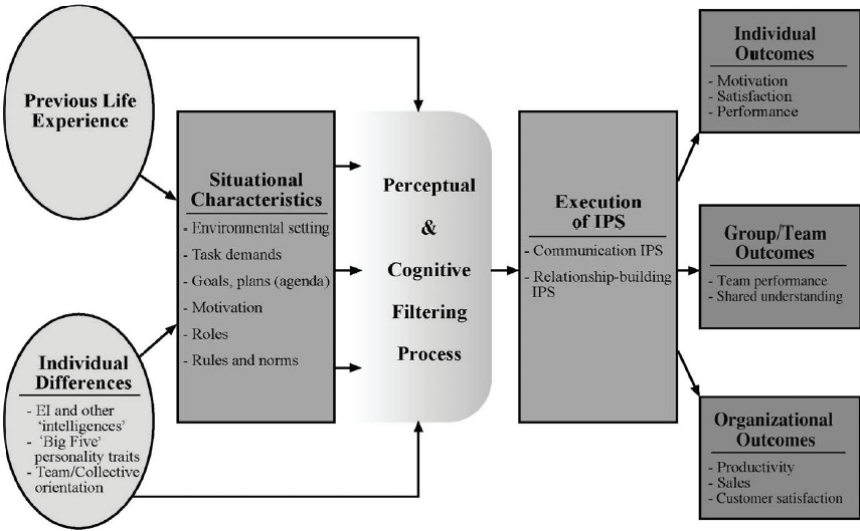


FIGURE 3-1 Model of interpersonal performance.

NOTE: Big Five personality traits = openness, conscientiousness, extraversion, agreeableness, and neuroticism; EI = emotional intelligence; IPS = interpersonal skills.

SOURCE: Stephen Fiore’s presentation. Klein, DeRouin, and Salas (2006). Copyright 2006, Wiley & Sons, Ltd. Reprinted with permission of John Wiley & Sons, Ltd.

also allow for measurement of specific actions in a way that could be used in selection decisions, performance appraisals, or training. More specifically, Figure 3-1 sets up a way of thinking about these skills in the contexts in which they are used. The implication for assessment is that one would need to conduct the measurement in a suitable, realistic context in order to be able to examine the attitudes, cognitive processes, and behaviors that constitute social skills.

ASSESSMENT APPROACHES AND ISSUES

One way to assess these skills, Fiore explained, is to look separately at the different components (attitudinal, behavioral, and cognitive). For example, as the model in Figure 3-1 indicates, previous life experiences, such as the opportunities an individual has had to engage in successful and unsuccessful social interactions, can be assessed through reports (e.g., personal statements from applicants or letters of recommendation from prior employers). If such narratives are written in response to specific

questions about types of interactions, they may provide indications of the degree to which an applicant has particular skills. However, it is likely to be difficult to distinguish clearly between specific social skills and personality traits, knowledge, and cognitive processes. Moreover, Fiore added, such narratives report on past experience and may not accurately portray how one would behave or respond in future experiences.

The research on teamwork (or collaboration)—a much narrower concept than interpersonal skills—has used questionnaires that ask people to rate themselves and also ask for peer ratings of others on dimensions such as communication, leadership, and self-management. For example, Kantrowitz (2005) collected self-report data on two scales: performance standards for various behaviors, and comparison to others in the subjects' working groups. Loughry, Ohland, and Moore (2007) asked members of work teams in science and technical contexts to rate one another on five general categories: contribution to the team's work; interaction with teammates; contribution to keeping the team on track; expectations for quality; and possession of relevant knowledge, skills, and abilities.

Another approach, Fiore noted, is to use situational judgment tests (SJTs), which are multiple-choice assessments of possible reactions to hypothetical teamwork situations to assess capacities for conflict resolution, communication, and coordination, as Stevens and Campion (1999) have done. The researchers were able to demonstrate relationships between these results and both peers' and supervisors' ratings and to ratings of job performance. They were also highly correlated to employee aptitude test results.

Yet another approach is direct observation of team interactions. By observing directly, researchers can avoid the potential lack of reliability inherent in self- and peer reports, and can also observe the circumstances in which behaviors occur. For example, Taggar and Brown (2001) developed a set of scales related to conflict resolution, collaborative problem solving, and communication on which people could be rated.

Though each of these approaches involve ways of distinguishing specific aspects of behavior, it is still true, Fiore observed, that there is overlap among the constructs—skills or characteristics—to be measured. In his view, it is worth asking whether it is useful to be "reductionist" in parsing these skills. Perhaps more useful, he suggested, might be to look holistically at the interactions among the facets that contribute to these skills, though means of assessing in that way have yet to be determined. He enumerated some of the key challenges in assessing interpersonal skills.

The first concerns the precision, or degree of granularity, with which interpersonal expertise can be measured. Cognitive scientists have provided models of the progression from novice to expert in more concrete skill areas, he noted. In K-12 education contexts, assessment developers

have looked for ways to delineate expectations for particular stages that students typically go through as their knowledge and understanding grow more sophisticated. Hoffman (1998) has suggested the value of a similar continuum for interpersonal skills. Inspired by the craft guilds common in Europe during the Middle Ages, Hoffman proposed that assessment developers use the guidelines for novices, journeymen, and master craftsmen, for example, as the basis for operational definitions of developing social expertise. If such a continuum were developed, Fiore noted, it should make it possible to empirically examine questions about whether adults can develop and improve in response to training or other interventions.

Another issue is the importance of the context in which assessments of interpersonal skills are administered. By definition, these skills entail some sort of interaction with other people, but much current testing is done in an individualized way that makes it difficult to standardize. Sophisticated technology, such as computer simulations, or even simpler technology can allow for assessment of people's interactions in a standardized scenario. For example, Smith-Jentsch and colleagues (1996) developed a simulation of an emergency room waiting room, in which test takers interacted with a video of actors following a script, while others have developed computer avatars that can interact in the context of scripted events. When well executed, Fiore explained, such simulations may be able to elicit emotional responses, allowing for assessment of people's self-regulatory capacities and other so-called soft skills.

Workshop participants noted the complexity of trying to take the context into account in assessment. For example, one noted both that behaviors may make sense only in light of previous experiences in a particular environment, and that individuals may display very different social skills in one setting (perhaps one in which they are very comfortable) than another (in which they are not comfortable). Another noted that the clinical psychology literature would likely offer productive insights on such issues.

The potential for technologically sophisticated assessments also highlights the evolving nature of social interaction and custom. Generations who have grown up interacting via cell phone, social networking, and tweeting may have different views of social norms than their parents had. For example, Fiore noted, a telephone call demands a response, and many younger people therefore view a call as more intrusive and potentially rude than a text message, which one can respond to at his or her convenience. The challenge for researchers is both to collect data on new kinds of interactions and to consider new ways to link the content of interactions to the mode of communication, in order to follow changes in what constitutes skill at interpersonal interaction. The existing definitions

and taxonomies of interpersonal skills, he explained, were developed in the context of interactions that primarily occur face to face, but new technologies foster interactions that do not occur face to face or in a single time window.

In closing, Fiore returned to the conceptual slippage in the terms used to describe interpersonal skills. Noting that the etymological origins of both “cooperation” and “collaboration” point to a shared sense of working together, he explained that the word “coordination” has a different meaning, even though these three terms are often used as if they were synonymous. The word “coordination” captures instead the concepts of ordering and arranging—a key aspect of teamwork. These distinctions, he observed, are a useful reminder that examining the interactions among different facets of interpersonal skills requires clarity about each facet.

ASSESSMENT EXAMPLES

The workshop included examples of four different types of assessments of interpersonal skills intended for different educational and selection purposes—an online portfolio assessment designed for high school students; an online assessment for community college students; a situational judgment test used to select students for medical school in Belgium; and a collection of assessment center approaches used for employee selection, promotion, and training purposes.

The first example was the portfolio assessment used by the Envision High School in Oakland, California, to assess critical thinking, collaboration, communication, and creativity. At Envision Schools, a project-based learning approach is used that emphasizes the development of deeper learning skills, integration of arts and technology into core subjects, and real-world experience in workplaces.² The focus of the curriculum is to prepare students for college, especially those who would be the first in their family to attend college. All students are required to assemble a portfolio in order to graduate. Bob Lenz, cofounder of Envision High School, discussed this online portfolio assessment.

The second example was an online, scenario-based assessment used for community college students in science, technology, engineering, and mathematics (STEM) programs. The focus of the program is on developing students’ social/communication skills as well as their technical skills. Louise Yarnall, senior research scientist with SRI, made this presentation.

Filip Lievens, professor of psychology at Ghent University in Belgium, described the third example, a situational judgment test designed

²See <http://www.envisionschools.org/site/> [August 2011] for additional information about Envision Schools.

to assess candidates' skill in responding to health-related situations that require interpersonal skills. The test is used for high-stakes purposes.

The final presentation was made by Lynn Gracin Collins, chief scientist for SH&A/Fenestra, who discussed a variety of strategies for assessing interpersonal skills in employment settings. She focused on performance-based assessments, most of which involve role-playing activities.

Online Portfolio Assessment of High School Students³

Bob Lenz described the experience of incorporating in the curriculum and assessing several key interpersonal skills in an urban high school environment. Envision Schools is a program created with corporate and foundation funding to serve disadvantaged high school students. The program consists of four high schools in the San Francisco Bay area that together serve 1,350 primarily low-income students. Sixty-five percent qualify for free or reduced-price lunch, and 70 percent are expected to be the first in their families to graduate from college. Most of the students, Lenz explained, enter the Envision schools at approximately a sixth-grade level in most areas. When they begin the Envision program, most have exceedingly negative feelings about school; as Lenz put it they "hate school and distrust adults." The program's mission is not only to address this sentiment about schools, but also to accelerate the students' academic skills so that they can get into college and to develop the other skills they will need to succeed in life.

Lenz explained that tracking students' progress after they graduate is an important tool for shaping the school's approach to instruction. The first classes graduated from the Envision schools 2 years ago. Lenz reported that all of their students meet the requirements to attend a 4-year college in California (as opposed to 37 percent of public high school students statewide), and 94 percent of their graduates enrolled in 2- or 4-year colleges after graduation. At the time of the presentation, most of these students (95 percent) had re-enrolled for the second year of college. Lenz believes the program's focus on assessment, particularly of 21st century skills, has been key to this success.

The program emphasizes what they call the "three Rs": rigor, relevance, and relationships. Project-based assignments, group activities, and workplace projects are all activities that incorporate learning of interpersonal skills such as leadership, Lenz explained. Students are also asked to assess themselves regularly. Researchers from the Stanford Center for Assessment, Learning, and Equity (SCALE) assisted the Envision staff in

³Lenz's presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Lenz.pdf [August 2011].

developing a College Success Assessment System that is embedded in the curriculum. Students develop portfolios with which they can demonstrate their learning in academic content as well as 21st century skill areas. The students are engaged in three goals: mastery knowledge, application of knowledge, and metacognition.

The components of the portfolio, which is presented at the end of 12th grade, include

- A student-written introduction to the contents
- Examples of “mastery-level” student work (assessed and certified by teachers prior to the presentation)
- Reflective summaries of work completed in five core content areas
- An artifact of and a written reflection on the workplace learning project
- A 21st century skills assessment

Students are also expected to defend their portfolios, and faculty are given professional development to guide the students in this process. Eventually, Lenz explained, the entire portfolio will be archived online.

Lenz showed examples of several student portfolios to demonstrate the ways in which 21st century skills, including interpersonal ones, are woven into both the curriculum and the assessments. In his view, teaching skills such as leadership and collaboration, together with the academic content, and holding the students to high expectations that incorporate these sorts of skills, is the best way to prepare the students to succeed in college, where there may be fewer faculty supports.

STEM Workforce Training Assessments⁴

Louise Yarnall turned the conversation to assessment in a community college setting, where the technicians critical to many STEM fields are trained. She noted the most common approach to training for these workers is to engage them in hands-on practice with the technologies they are likely to encounter. This approach builds knowledge of basic technical procedures, but she finds that it does little to develop higher-order cognitive skills or the social skills graduates need to thrive in the workplace.

Yarnall and a colleague have outlined three categories of primary skills that technology employers seek in new hires (Yarnall and Ostrander, in press):

⁴Yarnall’s presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Yarnall.pdf [August 2011].

Social-Technical

- Translating client needs into technical specifications
- Researching technical information to meet client needs
- Justifying or defending technical approach to client

Social

- Reaching consensus on work team
- Polling work team to determine ideas

Technical

- Using tools, languages, and principles of domain
- Generating a product that meets specific technical criteria
- Interpreting problems using principles of domain

In her view, new strategies are needed to incorporate these skills into the community college curriculum. To build students' technical skills and knowledge, she argued, faculty need to focus more on higher-order thinking and application of knowledge, to press students to demonstrate their competence, and to practice. Cooperative learning opportunities are key to developing social skills and knowledge. For the skills that are both social and technical, students need practice with reflection and feedback opportunities, modeling and scaffolding of desirable approaches, opportunities to see both correct and incorrect examples, and inquiry-based instructional practices.

She described a project she and colleagues, in collaboration with community college faculty, developed that was designed to incorporate this thinking, called the Scenario-Based Learning Project (see Box 3-2). This team developed eight workplace scenarios—workplace challenges that were complex enough to require a team response. The students are given a considerable amount of material with which to work. In order to succeed, they would need to figure out how to approach the problem, what they needed, and how to divide up the effort. Students are also asked to reflect on the results of the effort and make presentations about the solutions they have devised. The project begins with a letter from the workplace manager (the instructor plays this role and also provides feedback throughout the process) describing the problem and deliverables that need to be produced. For example, one task asked a team to produce a website for a bicycle club that would need multiple pages and links.

Yarnall noted they encountered a lot of resistance to this approach. Community college students are free to drop a class if they do not like the instructor's approach, and because many instructors are adjunct faculty,

BOX 3-2**Sample Constructs, Evidence of Learning, and Assessment Task Features for Scenario-Based Learning Projects****Technical Skills***Sample knowledge/skills/abilities (KSAs):*

Ability to document system requirements using a simplified use case format; ability to address user needs in specifying system requirements.

Sample evidence:

Presented with a list of user's needs/uses, the student will correctly specify web functionalities that address each need.

Sample task features:

The task must engage students in the use of tools, procedures, and knowledge representations employed in Ajax programming; the assessment task requires students to summarize the intended solution.

Social Skills*Sample social skill KSAs:*

Ability to listen to team members with different viewpoints and to propose a consensus.

Sample evidence:

Presented with a group of individuals charged with solving a problem, the student will demonstrate correctly indicators of active listening and collaboration skills, including listening attentively, waiting an adequate amount of time for problem solutions, summarizing ideas, and questioning to reach a decision.

Sample social skill characteristic task features:

The assessment task will be scenario-based and involve a group of individuals charged with solving a work-related problem. The assessment will involve a conflict among team members and require the social processes of listening, negotiation, and decision making.

Social-Technical Skills*Sample social-technical skill KSAs:*

Ability to ask questions to specify user requirements, and ability to engage in software design brainstorming by generating examples of possible user interactions with the website.

Sample social-technical skill evidence:

Presented with a client interested in developing a website, the student will correctly define the user's primary needs. Presented with a client interested in developing a website, the student will correctly define the range of possible uses for the website.

Sample social-technical skill characteristic task features:

The assessment task will be scenario-based and involve the design of a website with at least two constraints. The assessment task will require the use of "querying" to determine client needs. The assessment task will require a summation of client needs.

SOURCE: Adapted from Louise Yarnall's presentation. Used with permission.

their positions are at risk if their classes are unpopular. Scenario-based learning can be risky, she explained, because it can be demanding, but at the same time students sometimes feel unsure that they are learning enough. Instructors also sometimes feel unprepared to manage the teams, give appropriate feedback, and track their students' progress.

Furthermore, Yarnall continued, while many of the instructors did enjoy developing the projects, the need to incorporate assessment tools into the projects was the least popular aspect of the program. Traditional assessments in these settings tended to measure recall of isolated facts and technical procedures, and often failed to track the development or application of more complex cognitive skills and professional behaviors, Yarnall explained. She and her colleagues proposed some new approaches, based on the theoretical framework known as evidence-centered design.⁵ Their goal was to guide the faculty in designing tasks that would elicit the full range of knowledge and skills they wanted to measure, and they turned to what are called reflection frameworks that had been used in other contexts to elicit complex sets of skills (Herman, Aschbacher, and Winters, 1992).

They settled on an interview format, which they called Evidence-Centered Assessment Reflection, to begin to identify the specific skills required in each field, to identify the assessment features that could produce evidence of specific kinds of learning, and then to begin developing specific prompts, stimuli, performance descriptions, and scoring rubrics for the learning outcomes they wanted to measure. The next step was to determine how the assessments would be delivered and how they would be validated. Assessment developers call this process a domain analysis—its purpose was to draw from the instructors a conceptual map of what they were teaching and particularly how social and social-technical skills fit into those domains.

Based on these frameworks, the team developed assessments that asked students, for example, to write justifications for the tools and procedures they intended to use for a particular purpose; rate their teammates' ability to listen, appreciate different points of view, or reach a consensus; or generate a list of questions they would ask a client to better understand his or her needs. They used what Yarnall described as "coarse, three-level rubrics" to make the scoring easy to implement with sometimes-reluctant faculty, and have generally averaged 79 percent or above in inter-rater agreement.

Yarnall closed with some suggestions for how their experience might be useful for a K-12 context. She noted the process encouraged thinking about how students might apply particular knowledge and skills, and

⁵See Mislevy and Risconscente (2006) for an explanation of evidence-centered design.

how one might distinguish between high- and low-quality applications. Specifically, the developers were guided to consider what it would look like for a student to use the knowledge or skills successfully—what qualities would stand out and what sorts of products or knowledge would demonstrate a particular level of understanding or awareness.

Assessing Medical Students' Interpersonal Skills⁶

Filip Lievens described a project conducted at Ghent University in Belgium, in which he and colleagues developed a measure of interpersonal skills in a high-stakes context: medical school admissions. The project began with a request from the Belgian government, in 1997, for a measure of these skills that could be used not only to measure the current capacities of physicians, but also to predict the capacities of candidates and thus be useful for selection. Lievens noted the challenge was compounded by the fact the government was motivated by some negative publicity about the selection process for medical school.

One logical approach would have been to use personality testing, often conducted through in-person interviews, but that would have been very difficult to implement with the large numbers of candidates involved, Lievens explained. A paper on another selection procedure, called “low-fidelity simulation” (Motowidlo et al., 1990), suggested an alternative. This approach is also known as a situational judgment test, mentioned above, in which candidates select from a set of possible responses to a situation that is described in writing or presented using video. It is based on the proposition that procedural knowledge of the advantages and disadvantages of possible courses of action can be measured, and that the results would be predictive of later behaviors, even if the instrument does not measure the complex facets that go into such choices. A sample item from the Belgian assessment, including a transcription of the scenario and the possible responses, is shown in Box 3-3. In the early stages of the project, the team used videotaped scenarios, but more recently they have experimented with presenting them through other means, including in written format.

Lievens noted a few differences between medical education in Belgium and the United States that influenced decisions about the assessment. In Belgium, prospective doctors must pass an admissions exam at age 18 to be accepted for medical school, which begins at the level that for Americans is the less structured 4-year undergraduate program. The government-run exam is given twice a year to approximately 4,000 stu-

⁶Lievens' presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Lievens.pdf [August 2011].

BOX 3-3
Sample Item from the Situational Judgment Test Used
for Admissions to Medical School in Belgium

Situation:

Patient: So, this physiotherapy is really going to help me?

Physician: Absolutely, even though the first days it might still be painful.

Patient: Yes, I suppose it will take a while before it starts working.

Physician: That is why I am going to prescribe a painkiller. You should take three painkillers per day.

Patient: Do I really have to take them? I have already tried a few things. First, they didn't help me. And second, I'm actually opposed to taking any medication. I'd rather not take them. They are not good for my health.

Question:

What is the best way for you (as a physician) to react to this patient's refusal to take the prescribed medication?

- a. Ask her if she knows something else to relieve the pain.
- b. Give her the scientific evidence as to why painkillers will help.
- c. Agree not to take them now but also stress the importance of the physiotherapy.
- d. Tell her that, in her own interest, she will have to start changing her attitude.

SOURCE: Louise Yarnall's presentation. Used with permission.

dents in total, and it has a 30 percent pass rate. Once accepted for medical school, students may choose the university at which they will study—the school must accept all of the students who select it.

The assessment's other components include 40 items covering knowledge of chemistry, physics, mathematics, and biology and 50 items covering general cognitive ability (verbal, numerical, and figural reasoning). The two interpersonal skills addressed—in 30 items—are building and maintaining relationships and exchanging information.

Lievens described several challenges in the development of the interpersonal component. First, it was not possible to pilot test any items because of a policy that students could not be asked to complete items that did not count toward their scores. In response to both fast-growing numbers of candidates as well as technical glitches with video presentations, the developers decided to present all of the prompts in a paper-and-pencil format. A more serious problem was feedback they received ques-

tioning whether each of the test questions had only one correct answer. To address this, the developers introduced a system for determining correct answers through consensus among a group of experts.

Because of the high stakes for this test, they have also encountered problems with maintaining the security of the test items. For instance, Lievens reported that items have appeared for sale on eBay, and they have had problems with students who took the test multiple times simply to learn the content. Developing alternate test forms was one strategy for addressing this problem.

Lievens and his colleagues have conducted a study of the predictive validity of the test in which they collected data on four cohorts of students (a total of 4,538) who took the test and entered medical school (Lievens and Sackett, 2011). They examined GPA and internship performance data for 519 students in the initial group who completed the 7 years required for the full medical curriculum as well as job performance data for 104 students who later became physicians. As might be expected, Lievens observed, the cognitive component of the test was a strong predictor, particularly for the first years of the 7-year course, whereas the interpersonal portion was not useful for predicting GPA (see Figure 3-2). However, Figure 3-3 shows this component of the test was much better at predicting the students' later performance in internships and in their first 9 years as practicing physicians.

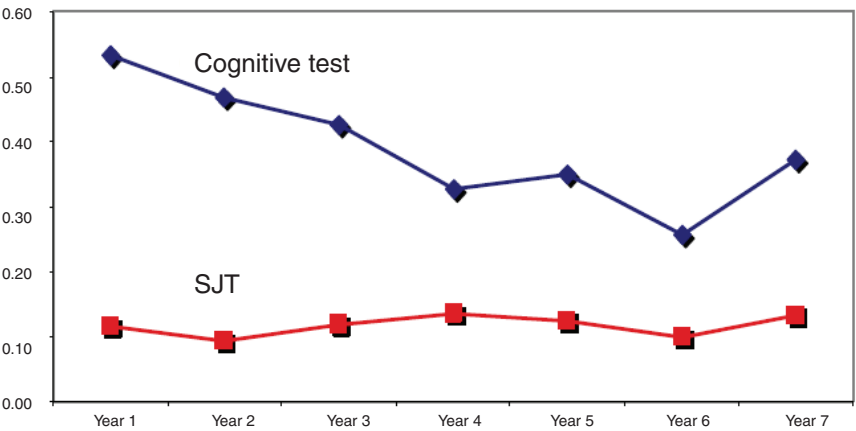


FIGURE 3-2 Correlations between cognitive and interpersonal components (situational judgment test, or SJT) of the medical school admission test and medical school GPA.

SOURCE: Filip Lievens' presentation. Used with permission.

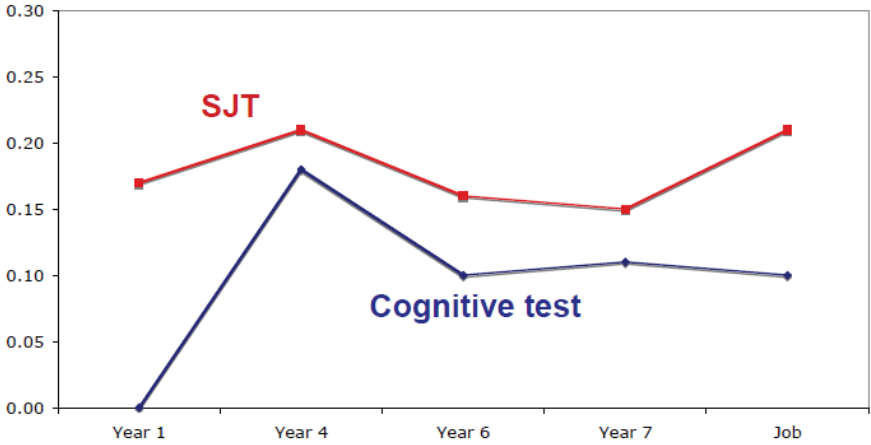


FIGURE 3-3 Correlations between the cognitive and interpersonal components (situational judgment test, or SJT) of the medical school admission test and internship/job performance.

SOURCE: Filip Lievens' presentation. Used with permission.

Lievens also reported the results of a study of the comparability of alternate forms of the test. The researchers compared results for three approaches to developing alternate forms. The approaches differed in the extent to which the characteristics of the situation presented in the items were held constant across the forms. The correlations between scores on the alternate forms ranged from .34 to .68, with the higher correlation occurring for the approach that maintained the most similarities in the characteristics of the items across the forms. The exact details of this study are too complex to present here, and the reader is referred to the full report (Lievens and Sackett, 2007) for a more complete description.

Lievens summarized a few points he has observed about the addition of the interpersonal skills component to the admissions test:

- While cognitive assessments are better at predicting GPA, the assessments of interpersonal skills were superior at predicting performance in internships and on the job.⁷
- Applicants respond favorably to the interpersonal component of the test—Lievens did not claim this component is the reason but noted a sharp increase in the test-taking population.

⁷Lievens mentioned but did not show data indicating (1) that the predictive validity of the interpersonal items for later performance was actually greater than the predictive validity of the cognitive items for GPA, and (2) that women perform slightly better than men on the interpersonal items.

- Success rates for admitted students have also improved. The percentage of students who successfully passed the requirements for the first academic year increased from 30 percent, prior to having the exam in place, to 80 percent after the exam was installed. While not making a causal claim, Lievens noted that the increased pass rate may be due to the fact that universities have also changed their curricula to place more emphasis on interpersonal skills, especially in the first year.

Assessment Centers⁸

Lynn Gracin Collins began by explaining what an assessment center is. She noted the International Congress on Assessment Center Methods describes an assessment center as follows⁹:

a standardized evaluation of behavior based on multiple inputs. Several trained observers and techniques are used. Judgments about behavior are made, in major part, from specifically developed assessment simulations. These judgments are pooled in a meeting among the assessors or by a statistical integration process. In an integration discussion, comprehensive accounts of behavior—and often ratings of it—are pooled. The discussion results in evaluations of the assessee's performance on the dimensions or other variables that the assessment center is designed to measure.

She emphasized that key aspects of an assessment center are that they are standardized, based on multiple types of input, involve trained observers, and use simulations. Assessment centers had their first industrial application in the United States about 50 years ago at AT&T. Collins said they are widely favored within the business community because, while they have guidelines to ensure they are carried out appropriately, they are also flexible enough to accommodate a variety of purposes. Assessment centers have the potential to provide a wealth of information about how someone performs a task. An important difference with other approaches is that the focus is not on “what *would* you do” or “what *did* you do”; instead, the approach involves watching someone actually perform the tasks. They are commonly used for the purpose of (1) selection and promotion, (2) identification of training and development needs, and (3) skill enhancement through simulations.

Collins said participants and management see them as a realistic job

⁸Collins' presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Collins.pdf [August 2011].

⁹See <http://www.assessmentcenters.org/articles/whatisassess1.asp> [July 2011].

preview, and when used in a selection context, prospective employees actually experience what the job would entail. In that regard, Collins commented it is not uncommon for candidates—during the assessment—to “fold up their materials and say if this is what the job is, I don’t want it.” Thus, the tasks themselves can be instructive, useful for experiential learning, and an important selection device.

Some examples of the skills assessed include the following:

- *Interpersonal*: communication, influencing others, learning from interactions, leadership, teamwork, fostering relationships, conflict management
- *Cognitive*: problem solving, decision making, innovation, creativity, planning and organizing
- *Intrapersonal*: adaptability, drive, tolerance for stress, motivation, conscientiousness

To provide a sense of the steps involved in developing assessment center tasks, Collins laid out the general plan for a recent assessment they developed called the Technology Enhanced Assessment Center (TEAC). The steps are shown in Box 3-4.

BOX 3-4
Steps involved in Developing the Technology Enhanced Assessment Center

Week 1:	Scoping out the task and planning
Weeks 3-12:	Job analysis/define the dimensions Create assessment plan/build exercises Conduct assessment reviews Revise assessment materials Develop benchmarks and interpretation guide for scoring protocol Conduct content validation Load simulation materials into technology platform Establish center schedule Pilot test the assessment Train assessors
Week 13-ongoing:	Implementation Conduct the assessments
Ongoing:	Trend analysis and support for improvement planning

SOURCE: Adapted from presentation by Lynn Gracin Collins. Used with permission.

Assessment centers make use of a variety of types of tasks to simulate the actual work environment. One that Collins described is called an “inbox exercise,” which consists of a virtual desktop showing received e-mail messages (some of which are marked “high priority”), voice messages, and a calendar that includes some appointments for that day. The candidate is observed and tracked as he or she proceeds to deal with the tasks presented through the inbox. The scheduled appointments on the calendar are used for conducting role-playing tasks in which the candidate has to participate in a simulated work interaction. This may involve a phone call, and the assessor/observer plays the role of the person being called. With the scheduled role-plays, the candidate may receive some information about the nature of the appointment in advance so that he or she can prepare for the appointment. There are typically some unscheduled role-playing tasks as well, in order to observe the candidate’s on-the-spot performance. In some instances, the candidate may also be expected to make a presentation. Assessors observe every activity the candidate performs.

Everything the candidate does at the virtual desktop is visible to the assessor(s) in real time, although in a “behind the scenes” manner that is blind to the candidate. The assessor can follow everything the candidate does, including what they do with every message in the inbox, any responses they make, and any entries they make on the calendar.

Following the inbox exercise, all of the observers/assessors complete evaluation forms. The forms are shared, and the ratings are discussed during a debriefing session at which the assessors come to consensus about the candidate. Time is also reserved to provide feedback to the candidate and to identify areas of strengths and weaknesses.

Collins reported that a good deal of information has been collected about the psychometric qualities of assessment centers. She characterized their reliabilities as adequate, with test-retest reliability coefficients in the .70 range. She said a wide range of inter-rater reliabilities have been reported, generally ranging from .50 to .94. The higher inter-rater reliabilities are associated with assessments in which the assessors/raters are well trained and have access to training materials that clearly explain the exercises, the constructs, and the scoring guidelines. Providing behavioral summary scales, which describe the actual behaviors associated with each score level, also help the assessors more accurately interpret the scoring guide.

She also noted considerable information is available about the validity of assessment centers. The most popular validation strategy is to examine evidence of content validity, which means the exercises actually measure the skills and competencies that they are intended to measure. A few studies have examined evidence of criterion-related validity, looking at the relationship between performance on the assessment center exer-

cises and job performance. She reported validities of .41 to .48 for a recent study conducted by her firm (SH&A/Fenestra, 2007) and .43 for a study by Byham (2010). Her review of the research indicates that assessment center results show incremental validity over personality tests, cognitive tests, and interviews.

One advantage of assessment center methods is they appear not to have adverse impact on minority groups. Collins said research documents that they tend to be unbiased in predictions of job performance. Further, they are viewed by participants as being fairer than other forms of assessment, and they have received positive support from the courts and the Equal Employment Opportunity Commission (EEOC).

Assessment centers can be expensive and time intensive, which is one of the challenges associated with using them. An assessment center in a traditional paradigm (as opposed to a high-tech paradigm) can cost between \$2,500 and \$10,000 per person. The features that affect cost are the number of assessors, the number of exercises, the length of the assessment, the type of report, and the type of feedback process. They can be logistically difficult to coordinate, depending on whether they use a traditional paradigm in which people need to be brought to a single location as opposed to a technology paradigm where much can be handled remotely and virtually. The typical assessment at a center lasts a full day, which means they are resource intensive and can be difficult to scale up to accommodate a large number of test takers.

4

Assessing Intrapersonal Skills

The third cluster of skills—intrapersonal skills—are talents or abilities that reside within the individual and aid him or her in problem solving. The previous workshop report that defined a set of 21st century skills (National Research Council, 2010) identified two broad skills that fall within this cluster:

Adaptability: The ability and willingness to cope with uncertain, new, and rapidly changing conditions on the job, including responding effectively to emergencies or crisis situations and learning new tasks, technologies, and procedures. Adaptability also includes handling work stress; adapting to different personalities, communication styles, and cultures; and physical adaptability to various indoor or outdoor work environments (Houston, 2007; Pulakos et al., 2000).

Self-management/self-development: The ability to work remotely, in virtual teams; to work autonomously; and to be self-motivating and self-monitoring. One aspect of self-management is the willingness and ability to acquire new information and skills related to work (Houston, 2007).

These kinds of skills operate across contexts, as Rick Hoyle, professor of psychology and neuroscience at Duke University, who presented findings from a paper about them and how they might be assessed, pointed out (Hoyle and Davisson, 2011).¹ They are “transportable,” he explained,

¹See http://www7.national-academies.org/bota/21st_Century_Workshop_Hoyle_Paper.pdf [August 2011].

automatically transferred from one context to the next so that the very same skills that serve a person well in the social arena, for example, serve the person well in health decisions and in schooling and academics. Furthermore, he added, these skills ultimately contribute to adaptive behavior and productivity in that they counteract undesired influences that may arise from within the person or from the environment. Intrapersonal skills support volitional behavior, which Hoyle defined as discretionary behavior aimed at accomplishing the goals an individual sets for himself or herself. Examples of intrapersonal skills include attributes such as planfulness, self-discipline, delay of gratification, the ability to deal with and overcome distractions, and the ability to adjust one's strategy or approach as needed. In Hoyle's view, the common thread among these attributes is a skill called self-regulation.

Even though the field of psychology has studied self-regulation since the late 1960s, Hoyle said, disagreement about how to define it remains. To provide the audience with the broad spectrum of definitions, he presented varying points of view that four prominent researchers have put forth:

- "The capacity of individuals to guide themselves, in any way possible, toward important goal states" (Fitzsimons and Bargh, 2004)
- "The capacity to plan, guide, and monitor one's behavior flexibly in the face of changing circumstances" (Brown, 1998)
- "Self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals" (Zimmerman, 2000)
- "The process by which one monitors, directs attention, maintains, and modifies behaviors to approach a desirable goal" (Ilkowska and Engle, 2010)

Hoyle identified some common threads among the definitions. They all recognize that people need to monitor their behavior and that they are doing this in the service of goal pursuit. In addition, they all acknowledge that flexibility is needed. Most importantly, they all involve affect. Hoyle emphasized that self-regulation does not just involve cognition but also involves feelings and emotions.

Hoyle prefers the following definition: the processes by which people remain on course in their pursuit of the goals they have adopted. In some cases, such as a school setting, these goals may not be the student's own, but they are put before students. The question is if they capable and ready to do the things that need to be done to pursue those goals and to move forward on them.

WHY IS SELF-REGULATION IMPORTANT?

Invariably the goals we adopt or are given to us are challenged in a number of ways. We may have counterproductive impulses, such as eating dessert even though we have a goal to lose weight. We may encounter situational hurdles, obstacles that interfere with the ongoing pursuit of some goal. We may have competing goals, so that satisfying one goal detracts from accomplishing another. Thus, people must manage the conflict between goals. And, in some cases, progress may be so slow that it is difficult to sustain motivation. Remaining on course toward goal pursuit requires a set of strategies that, collectively, constitute self-regulation.

Not every good behavior involves self-regulation, Hoyle clarified; self-regulation is behavior over which the individual exercises some level of discretion. Self-regulation requires considerable cognitive energy and effort. If the individual is constantly self-regulating, it is impossible to sustain momentum toward accomplishing a goal. It is most effective for the individual to move many behaviors outside the realm of the processes that require self-regulation.

For example, some behaviors are contingent on cues in the environment and are simply habits. The individual performs a habit when he or she links a behavior with some cue in the environment, Hoyle explained, and thus can accomplish the behavior without having to draw on self-regulation. Also, many behaviors are attributable to impulse. Impulses may be productive or they may detract from goal pursuit in some way, but they occur without self-regulation. Other behaviors are strongly influenced by normative pressure. This is frequently seen among adolescents, who experience a critical push/pull between the normative environment and their own individual goals. Finally, there are behaviors that are determined by social, political, or religious systems within which people live. In some cases those systems serve the role of regulating behavior, thereby circumventing the need for self-regulation.

Trends in society demonstrate some of the consequences that result from a lack of self-regulation. As Hoyle put it, “I don’t think we need much convincing that a lot of what we see around us seems to involve a failure of self-control at a fairly large level.” For example, Hoyle noted, U.S. consumers have not exerted much self-regulation when it comes to debt levels. In the late 1970s, consumer revolving credit debt was \$54 billion. By the end of the 1990s, it rose to more than \$600 billion, and now approaches \$1 trillion.² Likewise, Hoyle observed, obesity rates are at crisis levels. In 1990, no state had an obesity prevalence rate above 15 percent. By 2007, only one state had an obesity prevalence rate less than 20 percent, and 30 states had a preva-

²See http://www.federalreserve.gov/releases/g19/hist/cc_hist_sa.html [July 2011].

lence rate of 25 percent or more.³ In addition, a notable number of deaths can be attributed to failure of self-regulation. According to a report by the Centers for Disease Control and Prevention (Anderson, 2002), 33 percent of deaths were attributable to obesity, physical inactivity, and tobacco use. In addition, 8 percent of deaths were attributable to a cluster of behavioral causes including alcohol consumption, motor vehicle crashes, incidents involving firearms, sexual behaviors, and use of illicit drugs.

Furthermore, in Hoyle's view, the current economic crisis can be considered as a failure of self-regulation on a grand scale. Systemic, circumstantial, and societal issues all contributed to the crisis, but the excessive borrowing and lending and high-risk investments made with little or no concern for potential long-term consequences are all hallmarks of a lack of self-regulation. Hoyle emphasized that these examples all provide evidence of the importance of equipping children to be better self-regulated citizens as they approach adulthood. Hoyle also noted that considerable evidence in the literature underscores the value and importance of self-regulation. He focused on three studies.

One longitudinal study, conducted by Walter Mischel, began in the late 1950s and focused on delay of gratification (Mischel, 1958; Mischel et al., in press). Mischel used a variety of paradigms to study delay of gratification, and Hoyle described one that involved a set of preschoolers. The children were presented with an object they desired (e.g., a piece of candy or a marshmallow) but were told that they must wait until the experimenter returned to the room before they could have it. The experimenter left the room, closed the door, and intentionally did not return. Mischel collected data on how long each child waited before reaching for the object. After 10 to 12 years, Mischel contacted the parents of the participants and gathered information about their academic and social competence. He found that adolescent behavior was significantly predicted by the duration of the self-imposed delay in gratification. That is, the longer the preschooler was able to delay gratification, the better he or she fared as an adolescent in terms of a variety of self-regulation characteristics, such as attentiveness, planfulness, and reasoning ability.

A second longitudinal study by Caspi, Moffitt, and others (Caspi et al., 1997) with youngsters in Dunedin, New Zealand, is currently underway. The researchers are studying an entire birth cohort, collecting data every 2 to 3 years. At age 3, the children's temperament was evaluated, with some classified as "under-controlled." At age 18, the children who fell into the under-controlled classification rated high on a number of qualities that indicate poor self-regulation, including impulsivity and danger-seeking behavior, aggression, and interpersonal alienation. At

³See <http://www.cdc.gov/obesity/data/trends.html> [July 2011].

age 21, these individuals were more likely to be engaged in activities that show evidence of a failure to control their behavior, such as alcohol dependence, dangerous driving, violent behavior, and having unsafe sex. The likelihood of engaging in these behaviors was double that of the other children in the birth cohort.

Third, Hoyle cited work by economist Jim Heckman, who argues that noncognitive abilities are equally, if not more, important than traditional cognitive abilities when it comes to predicting educational and socioeconomic outcomes. The noncognitive attributes that Heckman refers to—attentiveness, persistence, impulse control, and social competence—are all evidence of self-regulation from a psychological perspective. Heckman's work shows that a gap between the disadvantaged and the advantaged begins to emerge very early. The point he makes is that there is a window of opportunity during which we can invest in those children. They can be taught to self-regulate, which Heckman finds will eventually result in significant dividends in terms of economic productivity, life success, and the like. Heckman (2006) reported findings from a study of children who participated in the Perry Preschool Program, which included a significant component of training of self-regulatory skills. He found students who participated in this program were less likely to drop out of school, spend time in jail, smoke, and participate in other self-destructive behaviors. In terms of economic productivity, the Perry Preschool Program participants were 15 to 17 percent higher than children who did not participate. Heckman argues that from an economic perspective there was a nine-fold payoff in what it costs to operate the Perry Preschool Program versus the payoff in economic productivity down the line.

DEFINING SELF-REGULATION

Although there has been considerable work on the topic of self-regulation in the field—with 114 chapters in edited volumes between 1998 and 2010⁴ and about 120 published articles each year—Hoyle said the field has no current consensus regarding a single definition of self-regulation. His review of the body of work revealed a definition is sometimes, but not always, provided. He finds no evidence of even minimal acceptance of a common definition, and even the same authors sometimes use different definitions. Furthermore, he thinks self-regulation has been applied

⁴Such as *Motivation and Self-Regulation Across the Life Span* published in 1998; the *Handbook of Self-Regulation* published in 2000; *Self-Regulation of Health and Illness Behaviour* published in 2003; the *Handbook of Self-Regulation: Research, Theory, and Applications* published in 2004; *Self-Regulation in Health Behavior* published in 2006; and the *Handbook of Personality and Self-Regulation* published in 2010.

far too broadly and, in many cases, inappropriately. Hoyle believes the current state of the conceptualization of self-regulation is the primary obstacle to producing assessments of it.

Hoyle laid out a conceptualization of self-regulation, which he emphasized was not really a model or a theory, but a framework that might help move forward in developing assessments. This conceptualization is presented in Figure 4-1. Understanding these components of self-regulation helps to provide a basis for defining constructs that might be assessed. Hoyle explained each of the components.

In the leftmost column (“Foundations”) are a series of variables or traits the individual “brings to the table.” These include (1) executive function, (2) temperament, and (3) personality characteristics. Hoyle added it is not clear whether these foundations are susceptible to change, but they are the “raw materials” that self-regulation draws upon.

Executive function is a set of cognitive processes and propensities that originate early in life (Goldman-Rakic, 1987; for a review, see Best and Miller, 2010). Three core functions underlie the processes involved in most acts of self-regulation (Miyake et al., 2000). *Inhibition* involves stopping ongoing thoughts and actions either when prompted by an external signal or upon determining that continuation would lead to an error (Logan and Cowan, 1984). *Working memory* involves keeping information active in primary memory while searching and retrieving information stored in secondary memory (Unsworth and Engle, 2007). Because keeping relevant information active while ignoring or suppressing competing information that is not relevant involves inhibition, inhibition and working memory

Foundations	Processes	Consequences
<ul style="list-style-type: none">➤ executive function<ul style="list-style-type: none">• inhibition• working memory• shifting➤ temperament<ul style="list-style-type: none">• effortful control• reactive control➤ personality<ul style="list-style-type: none">• higher-order• lower-order	<ul style="list-style-type: none">➤ receiving information➤ evaluating information➤ triggering change➤ searching for options➤ formulating a plan➤ implementing the plan➤ assessing effectiveness	<ul style="list-style-type: none">➤ normative➤ domain specific➤ idiosyncratic

FIGURE 4-1 A conceptualization of self-regulation.
SOURCE: Adapted from Rick Hoyle’s presentation. Used with permission.

are related. Complex tasks require the coordination of information relevant to multiple task components, requiring working memory to be flexible and controlled. Finally, *shifting* involves moving back and forth between mental states, rules, or tasks (Miyake et al., 2000). The importance of these basic capacities is evident in a cornerstone of self-regulation, the delay of gratification, which requires the inhibition of an impulse to act in response to a temptation in the immediate environment in favor of one or more longer-term goals or priorities (Mischel et al., in press).

Variability in executive function is expressed as individual differences in temperament, which Hoyle said is defined as individual differences in emotional and motor reactivity and in the attentional capacities that support self-regulation (Rothbart and Hwang, 2002, p. 113). One of the most important capacities is referred to as effortful control, which Hoyle explained is apparent when the child “is able to say no to that thing in front of them in service of some other thing that needs attention at that moment.” A related dimension of temperament is reactive control, which Hoyle described as the “relatively involuntary influence of approach and avoidance motives.” Extreme forms of reactive control can result in overcontrolled reactivity, such as shyness, or undercontrolled reactivity, such as impulsivity.

Hoyle defines personality as tendencies of thought, feeling, and action that are moderately stable across the lifespan (Roberts and DelVecchio, 2000), and he noted they can be separated into higher-order dimensions and lower-order dimensions. Research has shown that there are between three and seven higher-order dimensions (depending on the model and classification strategy) into which all personality traits fall. The dimension most relevant for self-regulation is conscientiousness, which generally concerns the ways people manage their behavior. Individuals who are high on conscientiousness tend to be confident, disciplined, orderly, and planful (Costa and McCrae, 1992).

A large number of narrower (lower-order) personality constructs also tend to facilitate or impede self-regulation. One of the most important is impulsivity, which Hoyle said might be viewed as the absence of self-regulation. Other lower-order personality constructs are relevant to self-regulation—those that concern self-regulatory style—and how (rather than whether) self-regulation is accomplished. They are foundational in their provision of the basic capacities and tendencies on which the processes involved directly in self-regulation draw.

In the middle column of Figure 4-1, Hoyle provided a list of the processes individuals go through as they try to accomplish a goal, although he cautioned that there is no agreement in his field on the exact nature of these processes. He noted the list helps to understand what

would be involved in an assessment of how effective an individual is at self-regulation.

The process generally begins with forethought, when the individual receives information, evaluates it, considers options, sets goals, and formulates a plan to achieve these goals. This is followed by performance, in which the individual implements the plan. From a self-regulation perspective, performance involves exercising self-control for the purpose of engaging in goal-relevant behaviors while avoiding behaviors irrelevant to or in conflict with the goal. Hoyle said a critical aspect of performance is self-observation or self-reflection, when the individual assesses the effectiveness of his or her performance and re-engages the process for subsequent attempts at goal pursuit. This model assumes a cyclical process whereby the individual repeatedly moves from forethought to performance to self-reflection, realizing progress toward the goal with each successive cycle.

The rightmost column of Figure 4-1 is labeled “Consequences,” which Hoyle maintains is probably the quickest approach to getting at a person’s skill level at self-regulation. What observable evidence is there that an individual is skilled or unskilled at self-regulation? He classified consequences into three categories.

One type of consequence is normative: that is, certain behaviors are evidence of a well-regulated individual regardless of the context or the particular population. Examples include academic success as evidenced by regularly completing assignments as instructed on schedule; social success in the form of routine relationship maintenance behaviors; and good health as evidenced by proper diet and exercise and general avoidance of health-risk behaviors.

Another type of consequence is domain-specific, such as self-regulation in the context of health behavior. For instance, hypertension patients often are prescribed a regimen that includes control of diet and regular intake of medications. Certain forms of psychotherapy might prescribe goals and behavioral evidence of their pursuit. In such instances, self-regulation is necessary and evidence of successful self-regulation is concrete and specific.

The final category is the idiosyncratic goals that each person decides on his or her own to pursue.

APPROACHES FOR ASSESSING SELF-REGULATION

Hoyle described a number of approaches for assessing self-regulation. One frequently used approach is self-report. In the typical use of this strategy, the respondent is given a set of statements and asked to select one of the provided response options to indicate extent of agreement or

disagreement with the statement or the degree to which the statement accurately describes him or her. There are advantages and disadvantages to this strategy, and Hoyle described several. It is often the least expensive approach in terms of materials as well as time and space requirements. There is also an implicit assumption that an individual is uniquely positioned to report on his or her standing on statements about the constructs and may well be the best source for the information. On the other hand, Hoyle noted, individuals are biased in both how they think about their own behavior and what they think is the task before them when they are responding to questionnaire items. There is evidence that people often do not have access to higher-order processes and therefore are unable to report about them accurately (Nisbett and Wilson, 1977). Hoyle said that there is also an age issue in that young children may lack the cognitive skills and reading ability to understand the statements they are asked to rate and the use of rating scales to do so.

Another approach is informant reports, which, Hoyle said, share many of the qualities of self-reports and address some of the limitations of the self-report strategy. One advantage of informant reports is that they eliminate the self-referential biases that may undermine the validity of self-reports. That is, Hoyle explained, well-trained informants who observe the target across time and situations may be able to infer and accurately report on characteristics of the target that the target is unable to accurately report about himself or herself. Another advantage Hoyle cited is that the informant report strategy allows for assessment of preverbal children, as well as of individuals who for other reasons may be unable to read and understand the statements on which they are to be rated. A clear drawback of the strategy, Hoyle noted, is the limited access most informants have to the individuals they are rating. For example, teachers only observe children in academic settings, parents see them primarily in the home, and peers are privy to behavior only in selected settings. Further, Hoyle stated, it may be difficult to extract information about specific skills and abilities from complex behavior sequences. That is, sometimes it is difficult to know, even after extensive observation, what is actually going on in the head of the person one is observing.

A third approach is behavioral task performances, which, Hoyle said, are designed so that they require only the capacity or skill of interest. Hoyle noted that these tasks are most often used to assess constructs in the foundations (see Figure 4-1), generally those capacities that constitute executive function. Speed and efficiency in completing these tasks is assumed to measure strength of the capacity being assessed. According to Hoyle, the tasks are tailored to the age group being assessed, and they generally do not require verbal skills or awareness by the individual of

his/her use of the capacity. The tasks are typically scored in terms of objective characteristics of performance (e.g., time to completion, number of mistakes). The positive features of assessments based on behavioral task performance are offset somewhat by two shortcomings, Hoyle cautioned. First, behavioral tasks tend to be tailored to the age group being assessed, which interferes with the ability to track performance over time. A second shortcoming concerns the purity of capacities assessed by the tasks. Complex tasks likely require multiple, interdependent capacities, thereby producing scores that cannot be used to pinpoint standing on specific capacities (Garon, Bryson, and Smith, 2008). They have the advantages of not requiring verbal skills, they do not require the person to report on higher order mental activity, and the scores tend to be objective (e.g., time to completion, number of mistakes). The measures tend to be things like Mischel's delay of gratification, which was the amount of time before the individual reached for the tempting object on the table. The disadvantages of this approach, Hoyle said, are that the tasks must be tailored to the age of the respondent and they often tap more than one skill or ability.

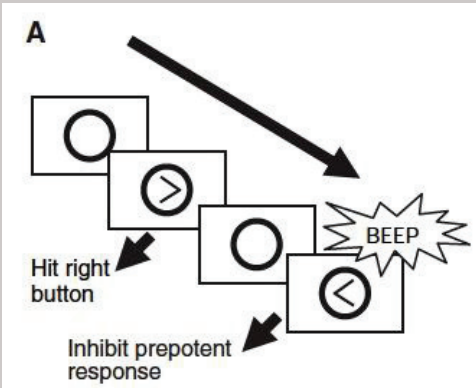
Hoyle described some examples of behavioral tasks performances intended to measure each of the foundational skills (see Figure 4-1). One task, referred to as the "stop signal measure" (see Box 4-1), is designed to measure executive function. In fairly rapid succession, the subject is presented with a series of cards. When the greater sign appears on the card, the subject is to press the right key, and when the lesser sign appears, the subject is to press the left key. At variable intervals, an audible sound occurs at which point the subject is not to press any key. The assessment measures how well they are able to inhibit and not press the key.

Another example, the star counting task, measures working memory. As shown in Box 4-2, the task begins with the number 15. In this case, when the subject reaches a plus sign, he/she is to count in the forward direction (16, 17, 18, etc.); when the subject reaches a minus sign, he/she is to change and count downward (18, 17, 16, etc.). The task is to get the right answer within a minute. A series of these is presented, and then the rules change so that a plus sign indicates to count in the backward direction and a minus sign indicates to count in the forward direction. This task measures the ability to change rules and hold the new rule in memory while overriding the old one.

Hoyle also showed examples of assessments intended to measure self-regulation through process and consequences (see Figure 4-1). The first is a self-report instrument on which the candidate rates him/herself on statements about processes, such as the ones that appear below:

- "I usually keep track of my progress toward my goals."
- "I have personal standards, and try to live up to them."

BOX 4-1
Example of a Stop Signal Task Designed
to Measure Executive Function



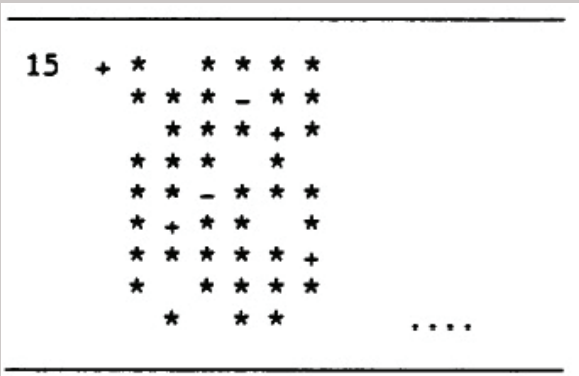
SOURCE: Adapted from Rick Hoyle's presentation. From Chamberlain, S.R. (2006). Neurochemical modulation of response inhibition and probabilistic learning in humans. *Science*, 311, 861. Reprinted with permission of American Association for the Advancement of Science.

- “I am willing to consider other ways of doing things.”
- “I have sought out advice or information about changing.”
- “Once I have a goal, I can usually plan how to reach it.”
- “I get easily distracted from my plans.” (reverse-scored)
- “I don’t seem to learn from my mistakes.” (reverse-scored)

The second is a self-report instrument that includes measures of behavior indicators of conscientiousness. The assumption is that the routine production of these behaviors is a sign of an individual who is either capable or not capable at self-regulation. The test taker rates him/herself on statements such as those shown below:

- “Play sick to avoid doing something” (avoid work)
- “Make a grocery list before going to the store” (organization)
- “Buy something on the spur of the moment” (impulsivity)
- “Clean the inside of the microwave oven” (cleanliness)
- “Work or study on a Friday or Saturday evening” (industriousness)
- “Clean up right after company leaves” (appearance)
- “Allow extra time for getting lost when going to new places” (punctuality)

BOX 4-2
Example of a Star Counting Task Designed
to Measure Working Memory



SOURCE: Adapted from Rick Hoyle's presentation. Reprinted from De Jong, P.F., and Das-Smaal, E.A. (1990). The star counting test: An attention test for children. *Personality and Individual Differences*, 11(6), 597-604. Copyright 1990, with permission from Elsevier.

In concluding, Hoyle noted that measures of foundational constructs are well established and, in many cases, have been adapted for use with infants and children. Measures of the self-regulation process are few and generally have not been adapted for use outside the research context. Behavioral consequences of the skill at self-regulating have not been considered in efforts at conceptualization and assessment.

ASSESSMENT EXAMPLES

At the workshop, four speakers discussed other examples of assessments of intrapersonal skills. Paul Sackett, professor of psychology with the University of Minnesota, made the first presentation and covered a variety of strategies for assessing integrity in employee selection settings. The second presentation, made by Candice Odgers, assistant professor of psychology, social behavior, and education with the University of California at Irvine, focused on strategies for assessing antisocial behaviors and conduct disorders in K-12 and counseling settings. Both of these types of assessments have been used operationally for some time. The remaining

two presenters discussed assessment strategies that are currently under research. Tim Cleary, associate professor of psychology with the University of Wisconsin–Milwaukee, discussed research on assessments of self-regulated learning. Gerald Matthews, professor of psychology with the University of Cincinnati, discussed research on assessing emotional intelligence.

Assessing Integrity in Job Applicants

Sackett began by talking about the origin of assessments like tests of integrity.⁵ He noted that for employers, the goal has always been to hire people likely to be good job performers. More recently, however, there has been a move to consider employees' contribution to an organization beyond simple task completion—not only what they do, but what they *do not* do. For instance, in a retail setting, the employer wants to hire sales clerks who perform their job well but who also do not pilfer money from the cash drawer. A trucking firm wants to hire drivers who deliver the products on time but who also obey traffic laws and drive safely. Sackett said that there are a host of behaviors, which he referred to as “counter-productive work behaviors,” that employers want to avoid in the people they hire, such as drinking or using drugs on the job, stealing, sexually abusing coworkers, lying, and cheating.

In many work settings, people work untended and have access to cash, money, and merchandise. Employers can take a number of preemptive steps to reduce the prevalence of these behaviors, such as installing cameras or other kinds of monitoring and control systems. But, there are limits to how many cameras can be installed and where they can be installed, and many settings are not easily monitored. Thus, Sackett explained, employers have moved in the direction of trying to screen potential employees and eliminate those likely to participate in counter-productive work behaviors. In response to this demand from employers, a set of commercial products emerged generically referred to as integrity tests.

Integrity tests are designed to predict theft and other forms of counterproductive work behavior. The measures are used internally by organizations, so the test taker never receives a score or feedback of any kind. The basis for using the test is predictive validity at the aggregate level. From an organization's point of view, it is not necessary to be able to precisely pinpoint which individuals will lie, cheat, or steal. The focus is on reducing these behaviors in the aggregate. As Sackett put it, “if I use this

⁵Sackett's presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Sackett.pdf [August 2011].

instrument to hire a workforce of 200 people, will I have fewer incidents of wrongdoing than if I did not use it?"

To help the audience understand the construct of integrity, Sackett presented them with a set of scenarios, all of which involved \$20 in cash that clearly belongs to someone else. He asked them to consider what they might do in each situation.

Scenario 1: You go to the gym with a friend for your weekly game of racquetball. After you finish your game, your friend heads for the shower before you, and you are left alone in the locker room. Your friend has not locked his locker, and you see his wallet peaking out of his jacket pocket. Would you reach in and take \$20 out of his wallet?

Scenario 2: You go to a "big box" store, such as Walmart or Target. You are checking out, and you ask the cashier a question. The cashier does not know the answer, leaves the cash drawer open, and leaves the counter to see if a coworker knows the answer. You see a \$20 bill within easy reach. Would you reach in and take the \$20 out of the cash drawer?

Scenario 3: You go to the ATM and follow the procedures to withdraw \$100. Your withdrawal arrives as six \$20 bills. Would you contact the bank to notify them that you received \$20 more than you had requested?

Scenario 4: You are making a purchase at your local grocery store. The cashier totals your purchase, and you pay in cash. When the cashier counts out your change, you are given \$20 more than you are due. You realize this on the spot. Would you let the cashier know of the mistake?

Scenario 5: You are walking down the street and find a wallet. You pick it up and discover that it is full of credit cards, along with an I.D. and \$20 in cash. You use the I.D. to locate the owner. When you return the wallet to the owner, do you keep the \$20 in cash or do you return it?

Sackett said that one of the readily apparent points is that integrity can depend on the situation. Many people would never take \$20 from a friend's wallet but might take the extra \$20 from a bank or a cashier, justifying it as "the ATM screwed up" or "if the sales clerk cannot count, it's not my problem." There are some situations that virtually no one reports that he or she would do, and others that many people report they would do. Situational features affect the percentage of individuals who say they would engage in the counterproductive behavior. But within any one situation, individual differences influence who does and does not engage in these behaviors.

Sackett described several different types of integrity measures. He noted that integrity tests were originally developed within the polygraph industry during the 1960s and 1970s. The practice of administering pre-employment polygraphs to prospective employees was banned by many states during this period (and nationally in 1988), and thus the polygraph industry began searching for alternative methods for prescreening job applicants. The measures were initially developed to predict theft, out of a demand by employers for pre-employment screenings of job candidates who would have access to cash. Eventually, they were expanded for use with a full range of behaviors. Over time, Sackett said three types of measures have emerged, all self-report.

He described the first as an “overt integrity test,” noting that it is overt in the sense that the intent of the assessment is clear to the job applicant. The assessment consists of a series of questions probing the candidate’s beliefs about the frequency and extent of theft and other forms of wrongdoing and their attitudes about punishment. The test consists of questions that assess seven categories of beliefs and attitudes, as shown below:

- Beliefs about the frequency and extent of theft (e.g., what percentage of people do you think cheat on taxes?)
- Punitiveness toward theft (e.g., an employee is caught taking \$100 from his organization. What punishment should the employee receive?)
- Ruminations about theft
- Perceived ease of theft
- Rationalizations about theft
- Assessments of one’s own honesty
- Admissions

Sackett explained the final category, “Admissions,” is actually a constructed-response item that asks the job candidate about his or her own theft behaviors, such as “what is the dollar value of cash and merchandise you have taken from your previous employer in the last six months?” Sackett said many are surprised to see this type of question and, more so, that anyone responds to it with an actual dollar amount. He hypothesized that people who steal from their employers believe everyone does it; therefore, they believe the employer would think they were lying if they said they took nothing (since everyone does it).

A second type of measure includes “personality-oriented tests,” which have their roots in the psychology discipline rather than the polygraph industry. He described three different commercial products:

- The Personnel Reaction Blank: designed to measure wayward impulses. The items focus on dependability, conscientiousness, and social conformity.
- The Employment Inventory: intended to measure employee deviance. The items deal with trouble with authority, thrill seeking, hostility, unhappy home life, and lack of work motivation.
- The Hogan Personality Inventory Reliability Scale: a measure of organizational delinquency. The items assess levels of hostility, impulse control, and attachment.

A third type of measure, called “conditional reasoning tests,” has emerged only recently. Sackett said the theory underlying these tests is that a person’s standing on a trait affects the justification mechanisms the person uses to explain his or her own behavior. Developed by Larry James at Georgia Institute of Technology, the theory is that people who are prone to engage in counterproductive work behavior will tend to be also high on a construct called “hostile attribution bias.” A sample item appears below:

American cars are now more reliable than they used to be 15 to 20 years ago. Why?

Option A: American car makers knew less about building reliable cars 15 to 20 years ago.

Option B: Prior to the introduction of high-quality foreign cars, American car makers purposely built cars badly in order to sell more repair parts.

According to James’ theory, endorsing option B is a manifestation of hostile attribution bias. The purpose of the assessment is disguised to the candidate. The candidate is told that it is a reasoning test, but, in fact, the focus is on the frequency with which he or she chooses the option with the aggressive or hostile undertone.

Sackett then turned to empirical evidence documenting the validity of integrity tests. He has conducted several literature reviews on this topic. In the first review (Sackett and Decker, 1979), he found six studies of tests of honesty. In subsequent reviews, he found 40 (Sackett and Harris, 1984), 70 (Sackett, Burris, and Callahan, 1989), and most recently 665 (Sackett and Wanek, 1996). At this point, the field of employment testing considers the validity of integrity tests to be well established.

Generally, the findings show validity coefficients in the .20 to .30 range. He said the three types of tests appear to have similar levels of validity. While most of the tests originally focused on identifying job candidates likely to steal, the tests predict a wide array of counterproductive

work behaviors. Some behaviors (e.g., absence from work) are more predictable than others (e.g., theft). He commented that theft, in particular, is difficult to predict because it tends to occur rarely, and detection of it is rare. The detection rate is much lower than the rate of engaging in the behavior, which complicates attempts to study the behavior.

The studies also show the tests predict overall job performance. Sackett believes this relationship is attributable to underlying constructs of conscientiousness and other forms of self-regulation that cause people to perform well at work as well as to avoid wrong-doing.

The studies show minimal subgroup differences on the tests, suggesting that employers do not need to worry about fairness or adverse impact in using them. Generally, women perform higher than men, but the performance differences follow the gender patterns seen in other forms of deviant and criminal behavior.

The tests tend to have a low relation with measures of cognitive ability, indicating they provide information that is not redundant with the other kinds of measures employers often use. There is evidence they are valid for both high- and low-complexity jobs.

One concern with these tests is their reliance on self-reports of attitudes and behaviors, which raises concerns about fakeability. Studies have been conducted to investigate this using a strategy called “instructed faking.” In the classic instructed faking paradigm, subjects are randomly assigned to two conditions. One group is told to try to score as high as possible and not to worry about responding honestly; the other group is told to respond honestly. The results show those in the first group score higher. Sackett said this result demonstrates the tests are conceptually fakeable, but he thinks it is important to evaluate this finding in light of results from the validity studies discussed above. In his view, if faking were prevalent in live applicant contexts, the validity coefficients would be diminished. Sackett added that faking is currently a concern for the first two types of integrity tests (overt and personality oriented), but not for conditional reasoning tests because their purpose is disguised. These tests would become fakeable if test takers were to discern their true purpose.

Sackett closed by noting that integrity testing developed from an applied standpoint, but the field has now shifted to a theoretical orientation. Initially, employers simply sought a device to identify job candidates likely to participate in wrong-doing on the job; they simply wanted a method that worked. The objective of more recent research has been to investigate why integrity tests work and to understand the underlying mechanisms by which they predict counterproductive work behaviors. Much of this work has centered on the self-regulation literature.

Assessing Antisocial Behavior and Conduct Disorders

Candice Odgers' work focuses on antisocial behavior, sometimes referred to as conduct disorders in children.⁶ A list of typical antisocial behaviors appears below:

- Aggression (e.g., fights, is physically cruel to people or animals, bullies, uses weapon, makes threats)
- Theft (e.g., steals, shoplifts, takes things)
- Deceitful (e.g., lies, cheats, blames others)
- Personality problems (e.g., is irritable, loud, jealous, hostile, annoying or demanding, brags and boasts)
- Rule-breaking (e.g., is disobedient, is truant, runs away from home)
- Oppositional (e.g., argues, swears, is stubborn, has tantrums)
- Destructive (e.g., commits vandalism, sets fires)

Odgers explained these are the children who have no ability to delay gratification. As she put it, "in the study by Mischel that Hoyle described, these are the children who would eat the candy or marshmallow before the interviewer left the room and then give the interviewer a defiant look." They are the children who become adults who steal from their employers. Antisocial behavior is apparently quite prevalent, Odgers reported, with an estimated lifetime prevalence of nearly 10 percent (Nock et al., 2006). Research shows antisocial behavior in children is a robust predictor of a number of problematic behaviors in adults, including poor physical health, school failure, and economic problems (Moffitt et al., 2002; Odgers et al., 2008). It is closely linked to difficulties with self-regulation and deficits in executive functioning (Dishion and Connell, 2006; Ellis et al., 2004; Moffitt, 1993). As Odgers put it, "antisocial behavior is clearly a marker of bad things to come."

She pointed out that these behavior problems can be costly. One study showed that they result in an additional expense of about \$70,000 per child in terms of the services used over the course of the 7 years of adolescence (Foster and Jones, 2005). The behaviors translate into unique challenges for families and schools, mental health and justice-related settings, and employers and social welfare systems.

Odgers addressed one question that arose repeatedly during the workshop—whether these kinds of intrapersonal skills are malleable, or specifically, whether these self-regulation skills can be changed. Odgers

⁶Odgers' presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Odgers.pdf [August 2011].

BOX 4-3
Example Items from the Childhood
Behavior Checklist in the ASEBA

Below is a list of items that describe children and youths. For each item that describes your child **now or within the past 6 months**, please circle the **2** if the item is **very true or often true** of your child. Circle the **1** if the item is **somewhat or sometimes true** of your child. If the item is **not true** of your child, circle the **0**. Please answer all items as well as you can, even if some do not seem to apply to your child.

- 0 = Not true (as far as you know)
1 = Somewhat or sometimes true
2 = Very true or often true

- | | |
|--------------|---|
| 0 1 2 | 1. Argues a lot |
| 0 1 2 | 2. Has temper tantrums or a hot temper |
| 0 1 2 | 3. Is stubborn, sullen or irritable |
| 0 1 2 | 4. Doesn't get along well with other kids |
| 0 1 2 | 5. Destroys his/her own things |
| 0 1 2 | 6. Lying or cheating |
| 0 1 2 | 7. Runs away from home |
| 0 1 2 | 8. Physically attacks people |
| 0 1 2 | 9. Cruelty, bullying, or meanness to others |

SOURCE: Copyright T.M. Achenbach. Reproduced with permission.

said the answer depends in part on when interventions are attempted. Her field of developmental psychology has established optimal timing for the development of some of these skills, and a payoff associated with early identification and intervention. Early intervention ultimately reduces the persistence of antisocial behaviors and subsequent involvement with the juvenile justice system.

Fortunately, Odgers noted, antisocial behavior is relatively easy to diagnose. It is assessed in a semi-structured interview setting, and the most widely used instrument is the Achenbach System of Empirically Based Assessments (ASEBA).⁷ The ASEBA has been translated into 85 languages and reported in more than 7,000 articles. It is designed to assess children's academic performance, adaptive functioning, and behavioral/emotional problems. The assessment system uses behavior checklists designed for different age groups. A sample of the checklist items for school-aged children appears in Box 4-3.

⁷See <http://www.aseba.org> [August2011] for further information about the assessment.

The syndrome scales were derived via factor analysis and were normed on large population-based and clinical samples. Reliabilities, based on test-retest estimates and coefficient alpha, are in the .90 range. The assessment takes about 15 minutes to complete. It is also compatible with the Diagnostic and Statistical Manual's definition of conduct disorder and antisocial behavior disorder, which allows people to talk across disciplines. Other psychometric information about the assessment is available on the ASEBA website (see footnote 7).

For young children, parents and teachers complete the checklist. For adolescents and adults, the checklist is self-reported, although there is usually an attempt to gather information from another informant (parents and teachers for adolescents; spouse or significant other for adults).

Odgers has been involved in the longitudinal study in Dunedin, New Zealand, that Hoyle described. The study has followed 1,000 individuals born in 1972 and 1973, and the researchers have just finished the age-38 assessment. Assessments were done at birth and every couple of years thereafter, thus providing a longitudinal perspective of when these skills emerge (or when problems emerge) and how they relate to other skills and deficits. The study has yielded considerable information about the relationships between these skills and life outcomes. Odgers said that they are finding that conduct disorder, particularly persistent conduct disorder across childhood, is one of the most accurate signals of future problems across a wide array of domains, including mental health, physical health, economic functioning, and job prospects.

Odgers presented the graph shown in Figure 4-2 that displays the incidence of conduct problems for the males in the sample, following them from ages 7 to 26. The researchers identified four patterns of behavior: (1) individuals who were consistently low in conduct problems (solid line); (2) individuals who exhibited conduct problems in childhood, but the problems diminished over time (line with triangles); (3) individuals who began exhibiting conduct problems during adolescent years (line with circles); and (4) individuals who persistently exhibited conduct disorders from childhood on into adulthood (line with squares). The researchers have compared outcomes for these four groups.

Odgers said the first finding from this analysis is that antisocial behavior in childhood does not necessarily signal poor outcomes in adulthood. Some children may exhibit conduct problems early on, but these problems are dealt with or as Odgers put it "socialized out." Through the influences of family, school, peers, and other factors, these children develop effective self-regulation skills, and the conduct problems diminish over time. However, this does not happen for all children with early-onset conduct problems, and individuals whose problems persist into adulthood experience difficulties in a number of areas of life.

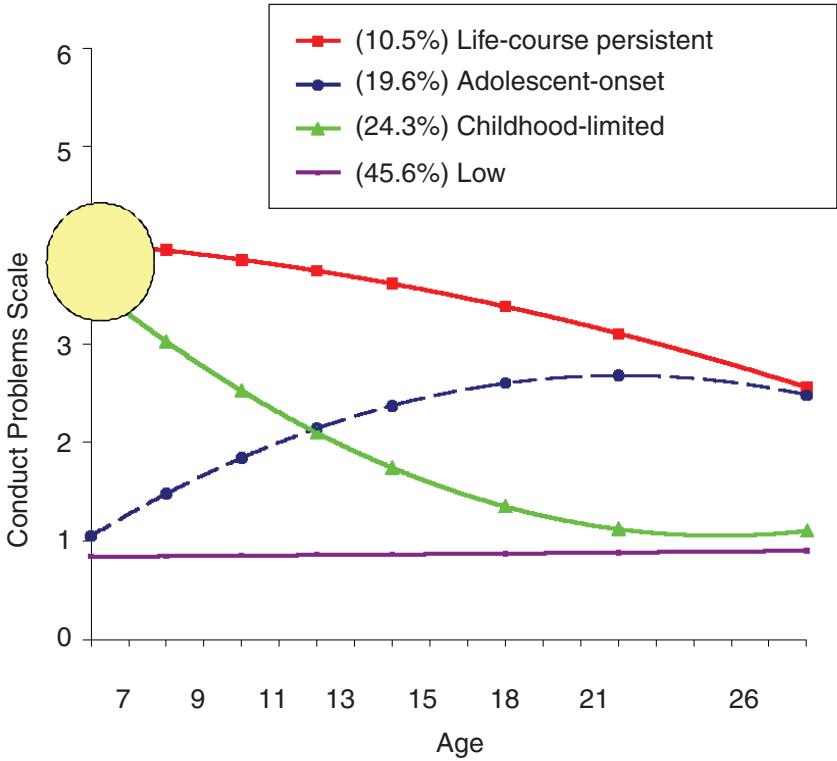


FIGURE 4-2 Incidence of conduct problems between ages 7 and 26 for longitudinal sample of individuals in Dunedin, New Zealand.
SOURCE: Odgers et al. (2007). Reprinted from Odgers, C.L., Milne, B.J., et al. (2007). Predicting prognosis for the conduct-problem boy: Can family history help? *Journal of American Academy of Child and Adolescent Psychiatry*, 46(10), 1240-1249, Copyright 2007, with permission from Elsevier. And from *Archives of General Psychiatry*, 64, 476-484, Copyright 2007, American Medical Association, All rights reserved.

For example, Figures 4-3 and 4-4 use effect sizes to compare health outcomes for males in the different groups. Figure 4-3 compares health outcomes for males with life-course-persistent conduct disorders versus those who scored low on conduct disorders. Figure 4-4 compares health outcomes for males with childhood-limited conduct disorders versus those who scored low on conduct disorders. The figures show the health outcomes for males with childhood-limited conduct disorders are quite similar to the health outcomes for individuals who scored low in conduct problems. On the other hand, the males with life-course persis-

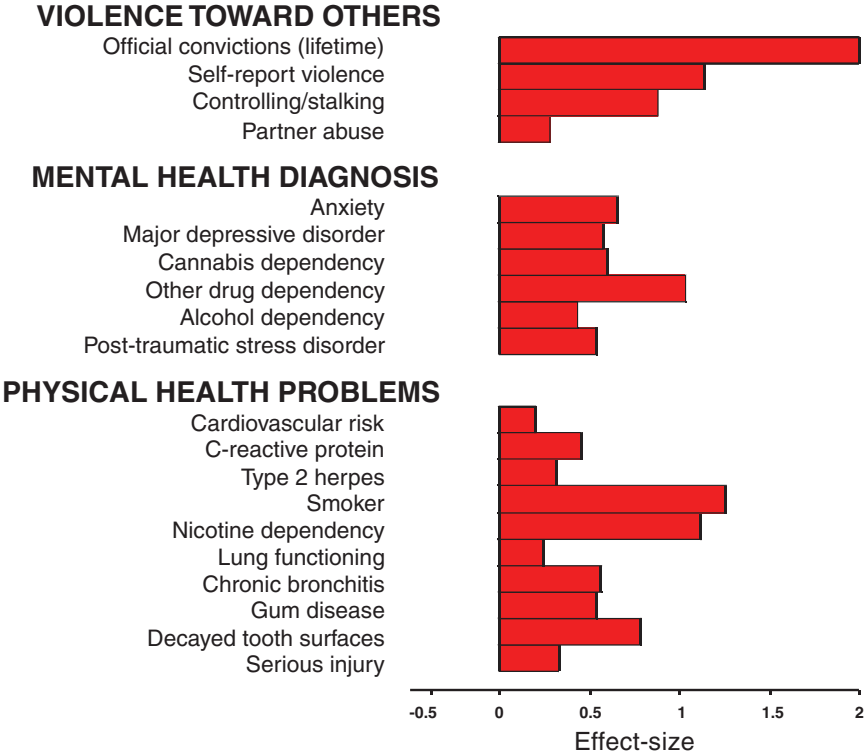


FIGURE 4-3 Health outcomes for males with life-course persistent conduct disorders compared to those who scored low on conduct disorders.
 SOURCE: Data from Odgers et al. (2008). Used with permission.

tent problems tended to be violent toward others and have convictions for this activity. They tended to suffer from anxiety and depression; were more likely to be dependent on alcohol, drugs, and tobacco; and had a greater incidence of health issues associated with these activities. Moreover, by age 32, 59 percent of this group had no educational qualifications⁸ as compared to an average of about 7 percent in the population at large. Only 24 percent of the males with childhood-limited conduct disorders had no educational qualifications, which Odgers noted was higher than average but half that for the males with life-course persistent conduct disorders.

⁸For the study, “No Educational Qualifications” was defined as ending secondary education prior to receiving qualifications (i.e., a diploma) and not having pursued further education.

VIOLENCE TOWARD OTHERS

Official convictions (lifetime)
Self-report violence
Controlling/stalking
Partner abuse

MENTAL HEALTH DIAGNOSIS

Anxiety
Major depressive disorder
Cannabis dependency
Other drug dependency
Alcohol dependency
Post-traumatic stress disorder

PHYSICAL HEALTH PROBLEMS

Cardiovascular risk
C-reactive protein
Type 2 herpes
Smoker
Nicotine dependency
Lung functioning
Chronic bronchitis
Gum disease
Decayed tooth surfaces
Serious Injury

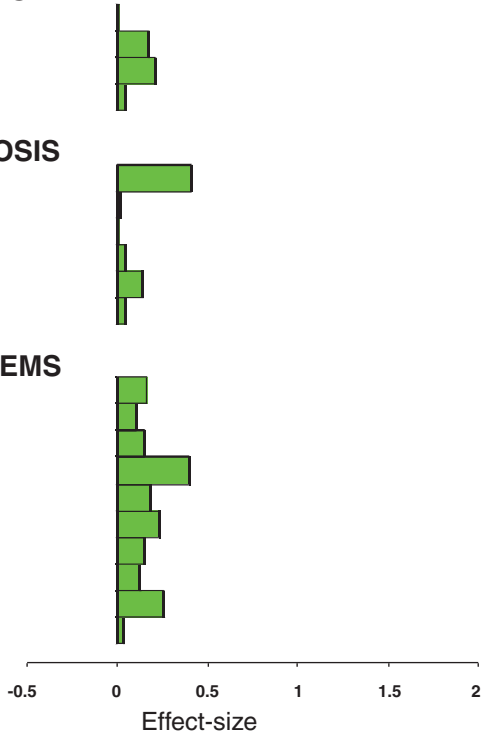


FIGURE 4-4 Health outcomes for males with childhood-limited conduct disorders compared to those who scored low on conduct disorders.

SOURCE: Data from Odgers et al. (2008). Used with permission.

Odgers closed by highlighting some new issues being pursued in her field. Bullying is a topic being intensely explored, including bullying in school and in the workplace, as well as cyber-bullying in all contexts. Assessment strategies are emerging, which are focusing on the traits of being callous and unemotional as a subtype of antisocial behavior in which the person lacks empathy and the ability to read and relate to others. These traits are being considered as a precursor to psychopathy. Odgers noted that children who have both antisocial behavior and this lack of empathy seem to have particularly poor outcomes. There are considerations to adding this characteristic to the conduct disorder diagnosis to help improve prediction of outcomes.

Odgers said that the field is quickly realizing the importance of collecting family history information about children, much in the way that

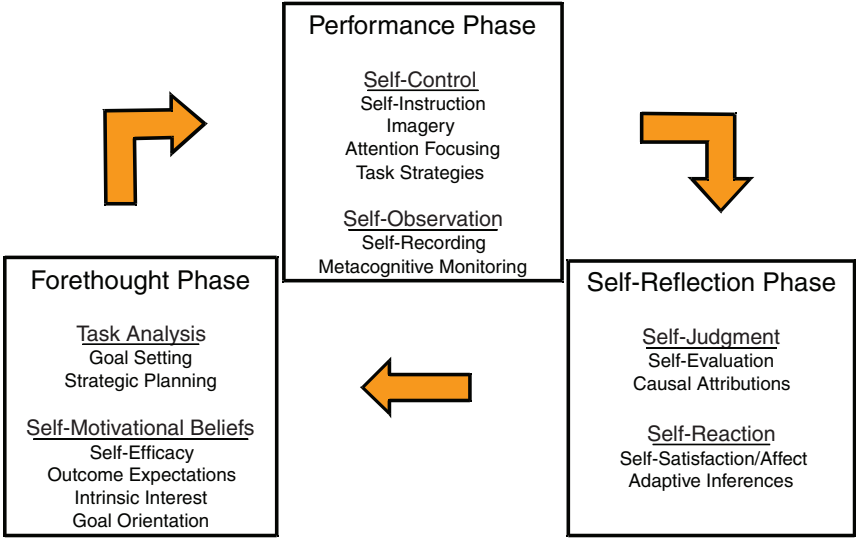


FIGURE 4-5 Three-phase model of self-regulated thought and action.
 SOURCE: Adapted from Zimmerman (2000). Used with permission.

it is done by medicine. Knowing about the parents' levels of antisocial behavior can help considerably in the diagnosis and prediction of long-term outcomes.

Microanalysis of Self-Regulated Learning

A self-regulated learner, Tim Cleary explained,⁹ is an individual who

- sets goals and develops/uses strategic plans;
- is highly self-motivated and proactive;
- engages in forms of self-control;
- monitors strategies, performance, and cognition; and
- frequently participates in self-reflection and analysis.

Cleary presented a three-phase model of self-regulated thought and action, as shown in Figure 4-5, which was developed by Zimmerman (2000) and referred to as a Cyclical Feedback Loop. The three phases of the model are forethought, performance, and self-reflection. The idea is that an individual approaches a task by considering what is involved, what it

⁹Cleary's presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Cleary.pdf [August 2011].

would take to complete the task, and how he or she should approach it. At the same time, the individual's approach to the task is influenced by how motivated he or she is to do the task, how important or valuable it is, and how confident he or she feels about successfully completing it. These ideas and thoughts impact the person's performance. During the performance phase, the individual uses self-regulation in order to complete the task. That is, he or she uses self-control strategies to stay on task and to learn what is being taught; he or she uses self-observation strategies to remain motivated and monitor learning. After a performance—typically after the individual receives some outcome such as a test grade, a quiz grade, or feedback on homework—he or she engages in reflection. At this phase, the individual evaluates the extent to which the goal has been reached, the factors that interfered with or helped with goal attainment, and considers his or her reaction to the performance (good or bad). Reflection is hypothesized to have an impact on subsequent attempts or subsequent strategies and modification of goals before the next learning attempt. Cleary noted this model forms the basis for microanalysis, which essentially focuses on diagnosis or the assessment of self-regulated learning and diagnosis of problems.

Cleary distinguished between two approaches toward measuring self-regulated learning based on work by Winne and Perry (2000): aptitude measures and event measures. The differences are in part related to the conceptualization of the construct. Aptitude measures, Cleary explained, are assessment tools that target self-regulated learning as a relatively global and enduring attribute of a person that predicts future behavior. They typically include self-report scales that rely on retrospective accounts of student behaviors and thoughts in terms of frequency, typicality, and usefulness. They generally capture the characteristics of self-regulated learning but they do so in a decontextualized manner. Some examples of aptitude scales are the (1) Motivated Strategies for Learning Questionnaire, (2) Learning and Study Strategies Inventory, (3) School Motivation and Learning Strategy Inventory, and (4) Self-Regulation Strategy Inventory. He highlighted two potential problems with this approach to measuring self-regulated learning. First, there are validity issues that relate to context-specificity. Research has shown that students' self-reports of self-regulated learning behaviors vary across different content areas as well as across tasks within a course. Second, student self-reports are often not consistent with their actual behaviors (Hadwin et al., 2001; Winne and Jamieson-Noel, 2002).

Event measures are assessment tools that target self-regulated learning as an event, behavior, or cognition that may vary across contexts and tasks. They involve direct assessment of self-regulatory processes as they occur in real time and in authentic contexts (as opposed to self-reports

about past events or behaviors). In Cleary's view, these measures are well equipped to capture the process of self-regulated learning. There are four approaches to obtaining event measures: (1) direct observations of students' actual behaviors in an authentic environment; (2) trace measures, which are overt indicators of student cognition created during task engagement (such as underlining or highlighting text while reading); (3) personal diaries in which students record their study behaviors at home or the types of thoughts they had and actions they took when performing specific tasks; and (4) verbal reports or "think-aloud protocols," which are records of students' thought as they complete authentic activities. Self-regulated learning microanalysis is an event measure that uses a structured interview approach to measure students' beliefs, attitudes, and cognitive regulatory processes before, during, and after some task or activity.

There are essentially four steps to the microanalysis approach that Cleary has studied. The first is to select a task with a clear beginning, middle, and end, such as studying for an exam or writing an essay. The second step is to identify the cyclical phase process that is of interest (see Figure 4-5), and the third step is to develop context-specific assessment questions to target the specific phase and process. Finally, and the most important element to Cleary's approach, is to link the three-phase cycle processes to temporal dimensions of the task: that is, to identify the questions to ask in the forethought phase, the performance phase, and the self-reflection phase. Cleary said it is the matching of the questions and the task in temporal terms that is the most important aspect of this approach.

Cleary and his colleagues have developed a bank of questions that can be adapted to a variety of contexts and tasks. They have administered these questions to school-aged and college samples in order to gather data on their reliability and validity. The reliability estimates, which are coefficient Alpha estimates for metric variables and inter-rater agreement for categorical variables tend to run in the .80 to .90 range. Cleary said they have developed coding manuals and scoring rubrics for training the raters, which helps to produce these high reliability coefficients.

In terms of validity, all of the questions are derived from operational definitions of theoretical constructs from social cognitive theory and expert consensus, which Cleary noted helps to provide evidence of content validity. The researchers have also collected evidence on the differential and predictive validity of self-regulated learning microanalysis. In one recent study with college students, the authors examined the extent to which the microanalytic self-regulation questions accounted for unique variance in student course grades over and above that accounted for by the most commonly used self-report measure of self-regulation, the Motivated Strategies Learning Questionnaire (MSLQ; Cleary et al.,

2010). The analyses indicated that the microanalytic questions Cleary and colleagues developed, which included “attribution” (the reasons why students thought they had received the grade) and “adaptive differences” (the ways that they thought they should do differently), accounted for approximately 30 percent of the variance in final course grades over and above that accounted for by scores on the MSLQ along with several background variables.¹⁰

Cleary and his colleagues have also conducted differential validity studies in the context of motor tasks and physical activities that demonstrate that goal-setting, strategic planning, attributions, and adaptive inferences reliably differentiate low and high achievers (Cleary and Zimmerman, 2001; Cleary, Zimmerman, and Keating, 2006; Kitsantas and Zimmerman, 2002). Different groups of students who had different levels of achievement (novices or experts) showed distinct profiles of regulatory processes.

Cleary closed by stressing that attribution and adaptive differences play an important role in how engaged students are in their studies and the extent to which they have effective strategies to identify their weaknesses and improve their performance.

Assessing Emotional Intelligence

Gerald Matthews began by cautioning the audience that the field of psychology is still in its infancy in terms of defining and assessing emotional intelligence.¹¹ On one hand, no one would want to be referred to as low on emotional intelligence. As he put it, “Saying that somebody has low emotional intelligence is now a pretty standard insult in various public domains.” On the other hand, research on emotional intelligence has not yet yielded a single conception of what it entails or how best to assess it. Thus, he advised, he would provide a “wide-angle” view of the state of the field, but he said there is no basis for coming to clear-cut conclusions about the construct.

In its broadest sense, Matthews explained, emotional intelligence includes abilities, competencies, and skills in perceiving, understanding, and managing emotion; however, there are a multitude of conceptualizations of the construct. One conception considers it as a set of abilities for

¹⁰The R square (variance explained) for the regression equation when background variables and MSLQ scores were included was .082. When information on responses to the attributions and adaptive inferences questions were added to the model, the R square increased to .373. This .291 change in the R square value was statistically significant at $p < .000$.

¹¹Matthews’ presentation is available at http://www7.nationalacademies.org/bota/21st_Century_Workshop_Matthews.pdf [August 2011].

processing emotional stimuli (Mayer et al., 2000) and treats the construct as a standard intelligence, having the kind of properties that other forms of intelligence and ability have. Another conception views emotional intelligence as part of the personality domain (Petrides and Furnham, 2003). In both cases, the assumption is that there is a general emotional intelligence factor that can be broken down into a number of more distinct competencies or skills.

Matthews thinks that neither conceptualization is useful. In his view, “emotional intelligence” is too vague a term to be of much use in either theory or practice (Roberts et al., 2007). He thinks it has become an “umbrella term” for a variety of separate attitudes, competencies, and skills that are only loosely interrelated, including basic temperament (e.g., positive and negative emotionality), information processing (e.g., emotion recognition), emotion-regulation (e.g., mood repair), and miscellaneous kinds of implicit and explicit skills.

Matthews talked about two commonly used strategies for assessing emotional intelligence—trait questionnaires and ability tests—though he cautioned that each strategy has drawbacks. He said many trait questionnaires are available and most are personality-like scales that provide scores for various emotional intelligence traits. He noted these are self-report assessments, which he thinks raises a paradox that undermines their validity. As Matthews put it, “If having good self-awareness of your emotional functioning is central to emotional intelligence, then if you lack emotional intelligence, how can your questionnaire responses be very meaningful?”

The Mayer-Salovey-Caruso Emotional Intelligence Test (the MSCEIT) is an example of an ability test for measuring emotional intelligence. The MSCEIT assesses the respondent’s ability to perceive, use, understand, and regulate emotions. The assessment uses scenarios drawn from everyday life situations to measure how well people perform tasks and solve emotional problems. For instance, the assessment includes the “Faces Subtest,” in which the test taker is presented with the face of a person showing an emotion, and the test taker rates the extent to which certain emotions are being expressed. Matthews showed an example of the face of a woman smiling. The test taker is asked to rate on a 5-point scale of “definitely not present” to “definitely present” the extent to which the face shows anger, disgust, sadness, happiness, fear, surprise, etc.

Matthews said one issue with the MSCEIT and other assessments like it is determining the “correct” response to an item. For the MSCEIT, the correct answers are determined through use of an expert panel and through collecting data from a normative sample. In Matthews’ view, neither approach is ideal, although he said that the assessment shows modest

correlations (.1 to .3) with a variety of criteria including life satisfaction, social skills and relationships, and coping.

Matthews and his colleagues Richard Roberts and others at ETS have been working on another assessment strategy that relies on situational judgment tests. The researchers are exploring the use of both text-based and video-based scenarios designed to evaluate how well individuals can judge the emotions of a situation. An example of a text-based scenario follows:

Clayton has been overseas for a long time and returns to visit his family. So much has changed that Clayton feels left out. What action would be the most effective for Clayton?

In the video-based format, a clip of an emotive situation is shown, and the test taker is presented with several response options. Matthews presented an example in which a person in a work situation is upset because her office is being moved around, and this has disrupted her work activities. The test taker is presented with four possible responses that the boss might make to address the employee's complaint. In one response, the boss becomes angry, tells her that the move is important for the firm's functioning, and that she should simply put up with it. In another, the boss is more empathetic with the employee, recognizes that the employee has some grounds for being upset, and explains the rationale behind the office move. The test taker is instructed to choose the best response. Matthews said that the work is in its early stages, but there seems to be some evidence that the results are predictive of high school GPA, well-being, and social support, even controlling for other factors.

Matthews closed by restating that emotional intelligence remains a nebulous and ill-defined construct. The field has not yet come to consensus on a definition or conceptualization of the construct, and findings from research examining its malleability—that is, the extent to which it is trainable—are inconclusive. While there are multiple strategies for assessing the construct, he thinks they are better suited for research than for any form of high-stakes testing.

5

Measurement Considerations

The assessments described in Chapters 2 through 4 have been designed for a variety of purposes. Some—such as the assessment components of Operation ARIES! Packet Tracer, and the scenario-based learning strategy described by Louise Yarnall—are designed primarily for formative purposes. That is, the assessment results are used to adapt instruction so that it best meets learners’ needs. Formative assessments are intended to provide feedback that can be used both by educators and by learners. Educators can use the results to gauge learning, monitor performance, and guide day-to-day instruction. Students can use the results to assist them in identifying their strengths and weaknesses and focusing their studying. A key characteristic of formative assessments is that they are conducted while students are in the process of learning the material.¹

Other assessments—referred to as summative—are conducted at the conclusion of a unit of instruction (e.g., course, semester, school year). Summative assessments provide information about students’ mastery of the material after instruction is completed. They are designed to yield information about the status of students’ achievement at a given point in time, and their purpose is primarily to categorize the performance of a student or a system. The PISA problem-solving assessment and the portfolio assessments used at Envision Schools are examples of summative

¹The reader is referred to Andrade and Cizek (2010) for further information about formative assessment and the difference between formative and summative assessment.

assessments, as are the annual state achievement tests administered for accountability purposes.

All assessments should be designed to be of high quality: to measure the intended constructs, provide useful and accurate information, and meet technical and psychometric standards. For assessments used to make decisions that have an important impact on test takers' lives, however, these issues are critical. When assessments are used to make high-stakes decisions, such as promotion or retention, high school graduation, college admissions, credentialing, job placement, and the like, they must meet accepted standards to ensure that they are reliable, valid, and fair to all the individuals who take them. A number of assessments used for high-stakes decisions were discussed by workshop presenters, including the Multistate Bar exam used to award certification to lawyers, the situational judgment test used for admitting Belgian students to medical school, the tests of integrity used for hiring job applicants, and some of the assessment center strategies used to make hiring and promotion decisions.

For the workshop, the committee arranged for two presentations to focus on technical measurement issues, particularly as they relate to high-stakes uses of summative assessments. Deirdre Knapp, vice president and director of the assessment, training, and policy studies division with HumRRO, spoke about the fundamentals of developing assessments. Steve Wise, vice-president for research and development with Northwest Evaluation Association, discussed the issues to consider in evaluating the extent to which the assessments validly measure the constructs they are intended to measure. This chapter summarizes their presentations and lays out the steps they highlighted as fundamental for ensuring that the assessments are of high quality and appropriate for their intended uses.² Where appropriate, the reader is referred to other sources for more in-depth technical information about test development procedures.

Defining the Construct

According to Knapp, assessment development should begin with a "needs analysis." A needs analysis is a systematic effort to determine exactly what information users want to obtain from the assessment and how they plan to use it. A needs assessment typically relies on information gathered from surveys, focus groups, and other types of discussions

²Knapp's presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Knapp.pdf [August 2011]. Wise's presentation is available at http://www7.national-academies.org/bota/21st_Century_Workshop_Wise.pdf [August 2011].

with stakeholders. Detailed information about conducting a needs analysis can be found at <http://www.needsassessment.org> [August 2011].

Knapp emphasized that it is important to have a clear articulation of the construct to be assessed: that is, the knowledge, skill, and/or behavior the stakeholders would like to have measured. The construct definition helps the test developer to determine how to measure it. She cautioned that for the skills covered in this workshop, developing a definition and operationalizing these definitions in order to produce test items can be challenging. For example, consider the variability in the definitions of critical thinking that Nathan Kuncel presented or the definitions of self-regulations that Rick Hoyle discussed. In order to develop an assessment that meets appropriate technical standards, the definition needs to be detailed and sufficiently precise to support the development of test items. Test development is less challenging when the construct is more concrete and discrete, such as specific subject-matter or job knowledge.

One of the more important issues to consider during the initial development stage, Knapp said, is whether the assessment needs to measure the skill itself or simply *illustrate* the skill. For instance, if the goal is to measure teamwork skills, is it necessary to observe the test takers actually performing their teamwork skills? Or is it sufficient that they simply answer questions that show they know how to collaborate with others and effectively work as a team? This is one of the issues that should be covered as part of the needs analysis.

Knapp highlighted the importance of considering which aspects of the construct can be measured by a standardized assessment and which aspects cannot. If the construct being assessed is particularly broad and the assessment cannot get at all components of it, what aspects of the construct are the most important to capture? There are always tradeoffs in assessment development, and careful prioritization of the most critical features can help with decision making about the construct. Knapp advised that once these decisions are made and the assessment is designed, the developer should be absolutely clear on which aspects of the construct are captured and which aspects are not.

Along with defining the construct, it is important to identify the context or situation in which the knowledge, skills, or behaviors are to be demonstrated. Identification of the specific way in which the construct is to be demonstrated helps to determine the type of assessment items to be used.

Determining the Item Types

As demonstrated by the examples discussed in Chapters 2, 3, and 4, there are many item types and assessment methods, ranging from

relatively straightforward multiple-choice items to more complex simulations and portfolio assessments. Knapp noted that some of the recent innovations in computer-based assessments allow for a variety of “glitzy” options, but she cautioned that while these options may be attractive, they may not be the best way to assess the targeted construct. The primary focus in deciding on the assessment method is to consider the knowledge, skill, and/or behavior that the test developer would like to elicit and then to consider the best—and most cost-effective—way to elicit it.

Knapp discussed two decisions to make with regard to constructing test items: the type of stimulus and the response mode. The stimulus is what is presented to the test taker, the task to which he/she is expected to respond. The stimulus can take a number of different forms such as a brief question, a description of a problem to solve, a prompt, a scenario or case study, or a simulation. The stimulus may be presented orally, on paper, or using technology, such as a computer.

The response mode is the mechanism by which the test taker responds to the item. Response modes might include choosing from among a set of provided options, providing a brief written answer, providing a longer written answer such as an essay, providing an oral answer, performing a task or demonstrating a skill, or assembling a portfolio of materials. Response modes are typically categorized as “selected response” or “constructed-response,” and constructed-response items are further categorized as “short-answer constructed-response,” “extended-answer constructed-response,” and “performance-based tasks.” Response modes also include behavior checklists, such as those described by Candice Odgers to assess conduct disorders, which may be completed by the test taker or by an observer. The response may be provided orally, on paper, through some type of performance or demonstration, or on a computer.

Knapp explained that choices about the stimulus type and the response mode need to consider the skill to be evaluated, the level of authenticity desired,³ how the assessment results will be used, and practical considerations. If the test is intended to measure knowledge of factual information, a paper-and-pencil test with brief questions and multiple-choice answer options may be all that is needed. If the test is intended to measure more complex skills, such as solving complex, multipart problems, a response mode that requires the examinee to construct an answer is likely to be more useful.

Layered on top of these considerations about the best ways to elicit the targeted skill are practical and technical constraints. Test questions

³Authenticity refers to how closely the assessment task resembles the real-life situation in which the test taker is required to use the skill being assessed. As described earlier, the level of authenticity desired is an issue that should be addressed as part of a needs analysis.

that use selected-response or short-answer constructed-response modes can usually be scored relatively quickly by machine. Test questions that use extended-answer constructed-response or performance-based tasks are more complicated to score. Some may be scored by machine, by programming the scoring criteria, but humans may need to score others. Scoring by humans is usually more expensive than scoring by machine, takes longer, and introduces subjectivity into the scoring process. Furthermore, constructed-response and performance-based tasks take longer to answer, and fewer can be included on a single test administration. They are more resource-intensive to develop and try out, and they usually present some challenges in meeting accepted measurement standards for technical quality. These practical and technical constraints are discussed in more detail below.

Test Administration Issues

How will the assessment be administered to test takers? Where will it be administered? When will it be administered and how often? Who will administer it? There are numerous options for how the test may be delivered to examinees and how they respond to it. Choosing among these options requires consideration of practical constraints.

A small assessment program, with relatively few examinees and infrequent administrations, has many options for administration, Knapp advised. For example, performance-based tasks that involve role playing, live performances, or that are administered one-on-one (one test administrator to one examinee) are much more practical when the examinee volume is small and test administrations are infrequent. When the examinee volume is large, performance-based tasks may be impractical because of the resources they require. The resources required for performance-based tasks can be reduced if they can be presented and responded to via computer, particularly if the scoring can be programmed as well.

Despite the resource required, several currently operating large standardized testing programs make use of performance-based tasks. As described in Chapter 2, the Multistate Bar Exam includes a performance-based component with a written response and is administered to approximately 40,000 candidates each year. Test takers pay \$20 to take this assessment.

Another example is Step 2 of the United States Medical Licensing Examination (USMLE), which includes a performance-based component. The Clinical Skills portion of the exam evaluates medical students' ability to gather information from patients, perform physical examinations, and communicate their findings to patients and colleagues. The assessment uses standardized patients to accomplish this. Standardized patients are

humans who are trained to pose as patients. They are trained in how they should respond to the examinee's questions in order to portray certain symptoms and/or diseases, and they are trained to rate the examinee's skills in taking histories from patients with certain symptoms. (For more information, see http://www.usmle.org/examinations/step2/step2cs_content.html [August 2011].) Approximately, 33,300 individuals took the exam between July 1, 2009, and June 30, 2010 (see <http://www.nbme.org/PDF/Publications/Annual-Report.pdf> [August 2011]). This exam is expensive for test takers; the fee to take the test is \$1,100.

A third example is the portfolio component of the assessment used to award advanced level certification for teachers by the National Board for Professional Teaching Standards (NBPTS). This assessment evaluates teachers' ability to think critically and reflect on their teaching and requires that teachers assemble a portfolio of written materials and videotapes. Approximately 20,000 teachers take the assessment each year (Mary Dilworth, vice president for research and higher education with the NBPTS, personal communication, May 31, 2011), and scores are available within 6 to 7 months (see <http://www.nbpts.org/> [August 2011]). This assessment is also expensive; examinees pay \$2,500 to sit for the exam.

Scoring

Knapp noted that if the choice is to use extended constructed-response or performance-based tasks, decisions must be made about how to score them. These types of open-ended responses may be scored dichotomously or polytomously. Dichotomous scoring means the answer is scored either correct or incorrect. Polytomous scoring means a graded scale is used, and points are awarded depending on the quality of the response or the presence of certain attributes in the response. Either way, a scoring guide, or "rubric," must be developed to establish the criteria for earning a certain score. The scoring criteria may be programmed so a computer does the scoring, or humans may be trained to do the scoring. When humans do the scoring, substantial time must be spent on training them to apply the scoring criteria appropriately. Since scoring constructed-response and performance-based tasks requires that scorers make judgments about the quality of the answer, the scorers need detailed instructions on how to make these judgments systematically and in accord with the rubric. Likewise, when constructed-response items are scored by computer, the computer must be "trained" to score the responses correctly, and the accuracy of this scoring must be closely monitored.

For some purposes, it is useful to set "performance standards" for the assessment. This might mean determining the level of performance considered acceptable to pass the assessment. Or it may mean classifying

performance into three or more categories, such as “basic,” “proficient,” and “advanced.” Making these kinds of performance-standards decisions requires implementing a process called “standard setting.” For further information about setting standards, see Cizek and Bunch (2007) or Zeiky, Perie, and Livingston (2008).

Technical Measurement Standards

Any assessment used to make important decisions about the test takers should meet certain technical measurement standards. These technical guidelines are laid out in documents such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), hereafter referred to as the *Standards*. Knapp and Wise focused on three critical technical qualities particularly relevant for assessments of the kinds of skills covered in the workshop, given the challenges in developing these assessments: reliability, validity, and fairness.

Reliability

Reliability refers to the extent to which an examinee’s test score reflects his or her true proficiency with regard to the trait being measured. The concern of reliability is the precision of test scores, and, as explained in more detail later in this section, the level of precision needed depends on the intended uses of the scores and the consequences associated with these uses (see also American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, pp. 29-30).

Reliability is evaluated empirically using the test data, and several different strategies can be used for collecting the data needed to calculate the estimate of reliability. One strategy involves administering the same form⁴ of the test or parallel forms of the test to the same group of examinees at independent testing sessions. When multiple administrations are impractical or unavailable, an alternative strategy involves estimating reliability from a single test form given on a single occasion. For this type of reliability estimate, the test form is divided into two or more constituent parts, and the consistency across these parts is determined using an estimate such as coefficient alpha or a split-half reliability coefficient. Each of these strategies for estimating reliability examines the precision

⁴A form is the specific collection of items or tasks that are included on the test.

of scores in relation to specific sources of error. Additional information about estimating reliability is available in Haertel (2006) and Traub (1994).

For tests that are scored by humans, another type of reliability information is commonly reported. When humans score examinee responses, they must make subjective judgments based on comparing the scoring guide and criteria to a particular test taker's performance. This introduces the possibility of scoring error associated with human judgment, and it is important to estimate the impact of this source of error on test scores. One estimate of reliability when human scoring is used is "inter-rater agreement," which is obtained by having two raters score each response and calculating the correlation between these scores. Knapp indicated that an estimate of inter-rater agreement provides basic reliability information, but she cautioned that it is not the only type of reliability evidence that should be collected when responses are scored by humans. A more complete data collection strategy involves generalizability analysis, which can be designed to examine the precision of test scores in relation to multiple sources of error, such as testing occasion, test form, and rater. Additional information about generalizability analysis is available in Shavelson and Webb (1991).

Reliability is typically reported as a coefficient that ranges from 0 to 1. The level of reliability needed depends on the nature of the test and the intended use of the scores: there are no absolute levels of reliability that are considered acceptable. When test results are used for high-stakes purposes, such as with a high school exit exam, reliability coefficients in the range of .90, or higher are typically expected. Lower reliability coefficients may be acceptable for tests used for lower stakes purposes, such as to determine next steps for instruction.

Generally, all else being equal, the more items on a test, the higher the reliability. This is because longer tests obtain a more extensive sample of the knowledge, skills, and behaviors being assessed than do shorter tests. Tests that rely on open-ended questions, such as extended-answer constructed-response and performance-based tasks, tend to consist of fewer items because these types of questions take more time to answer than do multiple-choice items. For practical reasons, such as the amount of testing time available, and because of concerns about examinee fatigue, tests can only include a limited number of these types of questions. Thus, tests that make use of open-ended questions tend to be less reliable than tests that primarily use multiple-choice questions, in part, because they contain fewer test questions. In addition, tests that require that judgments be made about the quality of the response—either when humans do the scoring or when scoring is done by artificial intelligence—introduce error associated with these judgments, which also tends to reduce reliability levels. Knapp advised that these factors

should be considered in relation to the interpretations and uses of test scores in making decisions about the types of questions used on the test.

Two other measures of score precision to consider are the standard error of measurement and classification consistency. The standard error of measurement provides an estimate of precision that is on the same scale as the test scores (i.e., as opposed to the 0 to 1 scale of a reliability coefficient). The standard error of measurement can be used to calculate a confidence band for an individual's test score. Additional information on standard errors of measurement and confidence bands can be found in Anastasi (1988, pp. 133-137), Crocker and Algina (1986, pp. 122-124), and Popham (2000, pp. 135-138), and the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, pp. 28-31).

The third measure of precision—classification consistency—is most relevant when tests are used to classify the test takers into performance categories, such as “basic,” “proficient,” or “advanced,” or simply as “proficient” or “not proficient,” or “pass” and “fail.” When important consequences are tied to test results, classification consistency should be examined. Classification consistency estimates the proportion of test takers who would be placed in the same category upon repeated administrations of the test. In this case, the issue is the precision of measurement near the cut score (the score used to classify test takers into the performance categories). Additional information about classification consistency can be found in the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 30).

It is important to point out that for some of the more innovative assessments, these measures of precision cannot be estimated. As Knapp put it, “computer-based technology has gotten way ahead of the capabilities of psychometric tools.” For example, at present there is no practical way to estimate reliability for some of the computerized assessments, such as those that are part of Operation ARIES! or Packet Tracer.

Validity

Validity refers to the extent to which the assessment scores measure the skills that they purport to measure. As Steve Wise framed it, validity refers to the “trustworthiness of the scores as being true representations of a student's proficiency in the construct being assessed.” Validation involves the evaluation of the proposed interpretations and uses of the test results. Validity is evaluated based on evidence—both rational and empirical, qualitative and quantitative. This includes evidence based on the processes and theory used to design and develop the test as well as

a variety of kinds of empirical evidence, such as analyses of the internal structure of the test, analyses of the relationships between test results and other outcome measures, and other studies designed to evaluate the extent to which the intended interpretations of test results are justifiable and appropriate. Wise and Knapp both emphasized that evaluation of validity and collection of validity evidence is a continuing, ongoing process that should be regularly conducted as part of the testing program. See Messick (1989) and Kane (2006) for further information about validation.

Wise noted that many factors can affect the trustworthiness of the scores, but two are particularly relevant for the issues raised in the workshop: motivation to perform well and construct irrelevant variance. One of the most important influences on motivation to perform well is the ways in which the scores are used—the interpretations made of them, the decisions about actions to take based on those interpretations, and the consequences (or stakes) attached to these decisions. When the stakes are high, Wise explained, the incentive to perform well is strong. The more important the consequences attached to the test results, the higher the motivation to do well. Motivation to perform well is critical, Wise stressed, in obtaining test results that are trustworthy as true representations of a student's proficiency with the construct. If the test results do not matter or do not carry consequences for students, they may not try their best, and the test results may be a poor representation of their proficiency level.

Motivation to do well can also bring about perverse behaviors, Wise cautioned. When test results have important consequences for students, examinees may take a number of actions to improve their chances of doing well—some appropriate and some inappropriate. For example, some students may study extra hard and spend long hours preparing. Others may find inappropriate short cuts that work to invalidate the test results, such as finding out the test questions beforehand, copying from another test taker, or bringing disallowed materials, such as study notes, into the test administration. These types of behaviors can produce scores that are not accurate representations of the students' true skills.

For the kinds of skills discussed at this workshop, motivation to do well can introduce a second source of error, which Wise described as "fake-ability." Some of the constructs have clearly socially acceptable responses. For example, if the assessment is designed to measure constructs such as adaptability, teamwork, or integrity, examinees may be able to figure out the desired response and respond in the socially acceptable way, regardless of whether it is a true representation of their attitudes or behaviors. Another concern with these kinds of items is that they may be particularly "coachable." That is, those who are helping a test taker prepare for the assessment can teach the candidate strategies for scoring high on the assessment without having taught the candidate

the skill or construct being assessed. Thus, the score may be influenced more by the candidate's skill in test taking strategies than his or her proficiency on the construct of interest.

A related issue is construct irrelevant variance. Problems with construct irrelevant variance occur when something about the test questions or administration procedures interferes with examinees' ability to assess the intended construct. For instance, if an assessment of teamwork is presented in English to students who are not fluent in English, the assessment will measure comprehension of English as well as teamwork skills. This may be acceptable if the test is intended to be an assessment of teamwork skills in English. If not, it will be impossible to obtain a precise estimate of the examinee's ability on the intended construct because another factor (facility with English) will interfere with demonstration of the true skill level. This can be a particular concern with some of the more innovative item types, such as those that are computer based or involve strategies such as simulations or role-playing, Wise noted. If familiarity with the item type or assessment strategy gives students an advantage that is not related to the construct, the assessment will give a flawed portrayal of the examinee's skills. This influences the validity of the inferences being made about the test scores.

Fairness

Fairness in testing means the assessment should be designed so that test takers can demonstrate their proficiency on the targeted skill without irrelevant factors interfering with their performance. As such, fairness is an essential component of validity. Many attributes of test items can contribute to construct irrelevant variance, as described above, and thus require skills that are not the focus of the assessment. For instance, suppose an assessment is intended to measure skill in mathematical problem solving, but the test items are presented as word problems. Besides assessing math skill, the items also require a certain level of reading skill. Examinees who do not have sufficient reading skills will not be able to read the items and thus will not be able to accurately demonstrate their proficiency in mathematical problem solving. Likewise, if the word problems are in English, examinees that do not have sufficient command of the English language will not be able to demonstrate their proficiency in the math skills that are the focus of the test.

Additional considerations about fairness may arise in relation to cultural, racial, and gender issues. Test items should be written so that they do not in some way disadvantage the test taker based on his or her racial/ethnic identification or gender. For example, if the math word problem discussed above uses an example more familiar or accessible to boys than

girls (e.g., an example drawn from sports), it may give the boys an unfair advantage. The same may happen if the example is more familiar to students from a white Anglo-Saxon culture than to racial/ethnic minority students. Many of the skills covered in the workshop present considerable challenges with regard to fairness. For example, cultural issues may cause differential performance on assessments of skills in communication, collaboration, or other interpersonal characteristics. Social inequities related to income, family background, and home environment may also cause differential performance on assessments. Students may not have equal opportunities to learn these skills.

The measurement field has a number of ways to evaluate fairness with assessments. The *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, pp. 71-106) provides a more complete discussion.

The Relationship Between Test Uses and Technical Qualities

Knapp and Wise both emphasized that when test results are used for summative purposes and high-stakes decisions are based on the results, the tests are expected to meet high technical standards to ensure decisions are based on accurate and fair information. For example, if a test is used for pass/fail decisions to determine who graduates from high school and who does not, the measurement accuracy of the scores needs to be high. Meeting high technical standards can be challenging and expensive because it requires a number of actions to be taken during the test development, administration, and scoring stages. For example, when tests are used for high-stakes purposes, reliability and classification consistency should be high. Test items will need to be kept secure. They cannot be reused multiple times because students remember them and pass the information on to others. Having to continually replenish the item pool is expensive and resource intensive, and it requires developing multiple forms of the test.

If different forms of the test are used, efforts have to be made to create test forms that are as comparable as possible. When tests are comprised of selected-response items or short constructed-response items, quantitative methods can be used to ensure that the scores from different test forms are equivalent. Statistical procedures—referred to as “equating” or “linking”—can be used to put the scores from different forms on the same scale and achieve this equivalence. For a number of reasons, linking or equating is usually not possible when tests are comprised solely of extended constructed-response items. In this situation, there is no straightforward way to ensure that the test forms are strictly comparable

and test scores equivalent across different forms. See Kolen and Brennan (2004) or Holland and Dorans (2006) for additional explanation of linking.

Thus, the test developer is often faced with a number of dilemmas. Constructed-response and performance-based tasks may be the most authentic way to assess 21st century skills. However, achieving high technical standards with these item types is challenging. When tests do not meet high technical standards, the results should not be used for high-stakes decisions with important consequences for students. But, when the results do not impact students' lives in important ways (i.e., "they do not count"), students may not try their best. Raising the stakes means increasing the technical quality of the tests. Test developers must face these issues and set priorities as to the most important aspects of the assessment. Is it more important to have authentic test items or to meet high reliability standards? Test developers are often faced with competing priorities and will need to make tradeoffs. Decisions about these tradeoffs will need to be guided by the goals and purposes of the assessment as well as practical constraints, such as the resources available.

6

Synthesis and Policy Implications

Joan Herman, chair of the steering committee and discussant at the workshop, posed two questions: Should we assess 21st century skills? If so, do we know how to do it?

In response to the first question, she said her answer was a whole-hearted “yes.” In her view, all of the workshop presentations demonstrated the importance of these skills. Beginning with Richard Murnane’s presentation that highlighted the critical relationships between these skills and labor market outcomes to presentations by Nathan Kuncel, Stephen Fiore, and Rick Hoyle, speakers emphasized the need for these skills to function well in today’s society. One after another, each presenter made a case for the need for students to be well rounded in their abilities to think critically; problem solve; interact effectively with others; and manage their own learning, emotions, and development. To Herman, it would be a disservice to students and society at large to focus schooling solely on narrow academic content while neglecting the broader aspects of development.¹ More important than simply assessing the skills, Herman noted, we should be integrating the assessment and teaching of 21st century skills with academic content. As she put it, “This should not be something added on to what teachers are already required to do, but should be part of their routine practice for building academic knowledge.”

But, do we know how to assess 21st century skills? Herman’s answer to this question was that it depends on the kinds of skills. With respect to

¹For additional discussion about breadth of instruction, see Bok (2006) and Lewis (2006).

cognitive skills, Herman thinks we know how to assess problem solving embedded in content, as Kuncel was arguing for. She noted that we also know how to develop assessments that require students to apply their knowledge, to evaluate evidence, and to perform other critical thinking and analytical reasoning tasks. There appear to be rich learning models on which to base these assessments, she added, but evaluating higher-order thinking skills has not received the attention it might have over the past few years.

With respect to some of the interpersonal and intrapersonal skills discussed at the workshop, she was somewhat more hesitant, but she said her hesitancy was in relation to the purposes and uses for the assessments, not the relative importance of the skills. She noted these days the word “assessment” has come to mean only large-scale, summative, accountability assessment, and, in her judgment, many of the measures of interpersonal and intrapersonal skills are clearly not ready to be used for this purpose. As she put it, “The long research histories in each area give rise to any number of measures for assessing individual constructs, but measures that are suitable for summative accountability purposes are few and far between.” Assessments can serve many purposes, however. For teachers, she pointed out, assessments are most useful if they provide information that can be used for formative purposes, to help make instructional decisions on a day-to-day basis. Some of the measures of interpersonal and intrapersonal skills seem to be well suited for this purpose or for purposes that involve small-scale administration.

As part of this discussion session, presenters and audience members raised a number of issues with regard to strategies for assessing 21st century skills, particularly the skills classified as interpersonal and intrapersonal. This chapter provides a synthesis of some of the main points raised by steering committee members and workshop participants and closes with a discussion of the implications for policy and strategies for moving forward.

REFLECTIONS ON ASSESSMENT STRATEGIES

Naming the Skill, Defining the Constructs

One point that arose repeatedly over the course of the workshop was the issue of labeling and defining the skills—from the name given to 21st century skills in general to the specific definitions of the constructs. Together, the collection of 21st century skills are sometimes referred to as “noncognitive” skills, a term to which several participants objected because all of the skills require some sort of cognition. These skills are sometimes referred to as “soft skills,” a term that some participants dislike

because it seems to downplay their importance. Others quibbled with the term “21st century skills” because it implies the skills were not needed in the 20th century and appears not to recognize that more than a decade of the 21st century has already passed. Thus, there is an issue with terminology at the broadest level.

There were also concerns expressed about placing these skills into three clusters (cognitive, interpersonal, and intrapersonal), as the committee had done. Some workshop participants pointed out it is misleading to imply the clusters of skills are independent and mutually exclusive. For instance, all of the skills included within the interpersonal and intrapersonal skills require cognition. That is, it is impossible to perform skills such as collaboration, complex communication, or self-regulation without using cognition. Likewise, intrapersonal skills and interpersonal skills are interdependent. For instance, self-management skills certainly come into play when participating in a collaborative task. The committee’s classifications were useful for the purposes of structuring the workshop, but there are issues with implying that the clusters are discrete and unrelated.

At a finer level, there are also issues with defining the constructs subsumed under the three broad categories identified by the committee. Stephen Fiore addressed this in his remarks in relation to interpersonal skills, noting “there is a proliferation of concepts associated with interpersonal skills, and it is problematic because we have different labels that may be describing the same construct, and we have the same label that may be describing a different construct.” For example, with regard to interpersonal skills, terms like social competence, soft skills, social self-efficacy, and social intelligence may all be used to refer to the same skills, or they may each refer to a different set of capabilities. Likewise, in discussing intrapersonal skills, Rick Hoyle pointed out the lack of consensus in the field with regard to defining skills like self-regulation. There is little agreement among researchers, he said, and sometimes the same researcher defines it differently within a single paper.

Settling on terminology for this set of skills and definitions for the constructs needs to be done before assessments can be developed. As Hoyle described this need in relation to self-regulation, “the current state of the conceptualization of self-regulation is the primary obstacle to producing assessments of it.” Defining the skills in a clear and precise way is fundamental to development of assessment tasks and essential for ensuring that the resulting scores support the intended inferences.

Validity, Reliability, and Authenticity

Another issue highlighted by workshop participants was the extent to which assessments of these skills are trustworthy and have fidelity. This

concern is essentially about reliability and validity: that is, do the assessments provide accurate results that support the intended inferences? The discussion centered around a number of issues related to reliability and validity, such as if the assessments measure what they are intended to measure; how susceptible they are to faking; how well they capture the actual processes involved in demonstrating the skill; and how reliable they are. The summary below elaborates on these issues in relation to each cluster of skills.

Cognitive Skills

With regard to skills in the cognitive cluster, such as critical thinking and problem solving, Kuncel pointed out, “We have a good understanding of these constructs when they are considered from a domain-specific perspective.” As he described, “we know what it means to think critically in certain contexts, such as when considering a physics problem or evaluating a study in cognitive psychology, and we have a good understanding of how to assess these skills from a domain-specific perspective.” The example assessments of cognitive skills presented at the workshop were all set within a context. For the PISA problem-solving test, each task specifies the context, which all come from situations encountered in daily life. The Multistate Bar exam poses critical thinking questions within the context of the situations lawyers encounter. Operation ARIES! focuses on evaluating scientific evidence, and Packet Tracer focuses on solving problems with computer networking.

According to Kuncel, the problems arise with domain-general conceptions of these skills. In his view, focusing on broad critical thinking skills, such as understanding the law of large numbers, and training students to apply these skills, is not a useful endeavor. In his work, he has found no evidence that learning these sorts of skills improves critical thinking in general or in ways that can be transferred from one domain to another. Further, he finds little evidence that a domain-general concept of critical thinking is distinct from general cognitive ability.

Interpersonal Skills

With regard to interpersonal skills, Fiore reminded the audience of the complexity of interpersonal interactions. Interpersonal skills involve a mix of attitudinal, behavioral, and cognitive factors, all of which are used to read the person in the context of the interaction and determine the most appropriate way to respond. Designing assessments to measure these processes is challenging. One issue that Fiore described is the fidelity of the assessment: that is, the extent to which the assessment involves observa-

tions of actual interactions and actual emotional responses to the interactions. He noted the scenario-based learning examples described by Louise Yarnall and the portfolio assessments described by Bob Lenz represent real-life interactions with authentic exchanges. With the scenario-based learning examples, the students are introduced to a problem through a real-life mechanism, such as an online letter from a manager. The students have to work in teams to address the situation and collaborate to figure out how to solve a complex work-related problem. These assessments integrate technical and social skills.

Fiore views the portfolio examples as somewhat less authentic. While the portfolios are structured collections of student work in which students have documented the application of knowledge in a particular classroom context, the evaluation of interpersonal skills is based on self-, peer, and teacher ratings. Although these ratings are drawn from actual situations the student was involved in, there is no control over the context or the nature of the interaction. For instance, the situations may or may not have involved conflict in the context of the collaborative projects. The type of communication on which the student is evaluated may differ from one student to the other. These variations interfere with both reliability and validity, Fiore commented, in that the sampling of behavior and performance included in the portfolio may not be consistent from year to year or even from student to student.

The other two examples—situational judgment tests and assessment center tasks—assess interpersonal skills in more contrived, controlled situations, Fiore said. The assessor sets up the situation to which the test taker is responding or in which the test taker is interacting. This guarantees that certain samplings of behavior are observed, but they are not as authentic as the other approaches. For instance, assessment centers obtain simulated examples of behavior; the observers see how job candidates perform in the situation simulated at the assessment center but not how the candidate performs when he or she actually encounters that situation in real life.

Fiore thinks situational judgment tests are even more removed from real-world situations in that the test taker simply chooses what he or she judges to be the best response. The candidate does not have to perform the skill or demonstrate the capability. Fiore characterizes these assessments as low in fidelity—low in enactive fidelity (the amount of true interaction that takes place) and low in affective fidelity (the extent to which the experience elicits an emotion response). He also highlighted certain problems that have arisen with these assessments. First, there is some complexity in understanding why a candidate may have responded incorrectly. To respond to the problem, the test taker has to choose the appropriate response to the situation, but he or she also has to interpret the situation.

When the test taker responds incorrectly, it is impossible to discern if he or she did not know the appropriate response or did not understand the situation. Fiore said situational judgment tests are also susceptible to faking in that test takers can make guesses about the most socially acceptable response. To address this concern, some assessments ask the test taker to choose the best and worst response, not just the best response. These issues present potential threats to the validity of the test results.

Intrapersonal Skills

Assessment of intrapersonal skills is also challenging because of the complexity of the processes involved. Hoyle reminded audience members that intrapersonal skills involve planfulness, self-discipline, delay of gratification, dealing with distractions, and adjusting the course when things do not go as planned—all characteristics of self-regulation or, put another way, the management of goal pursuit. The examples presented involved assessments of integrity, conduct disorders/antisocial behavior, self-regulated learning, and emotional intelligence. While these are all skills involved with self-regulation, Hoyle said one of the first things to consider is whether these skills are separable from personality. For instance, with regard to integrity, is there a certain personality profile associated with people who are prone to engage in dishonest behavior, or conversely people who are likely to operate with integrity in the workplace? Similar issues were raised by Gerald Matthews with respect to the distinction between emotional intelligence and personality.

The examples included a variety of strategies for assessing these skills. For tests of integrity, the strategies include both direct measures, such as self-report in which the test taker clearly knows the purpose of the assessment, and indirect measures, where the purpose is masked from the test taker. With regard to self-report measures of integrity, Hoyle questioned their utility, asking “How useful is it to ask a person who is dishonest to tell you if they are dishonest?” Nevertheless, he pointed out, considerable evidence documents their reliability, validity, and usefulness in employee selection. It is important to remember, however, that these assessments are used to reduce the prevalence of counterproductive behaviors in the aggregate, and test takers never receive their scores or any feedback on their performance. This is an important distinction from the type of testing done in the K-12 setting, where the focus is on reporting and interpreting scores in order to improve performance.

For evaluating antisocial behaviors and conduct disorders, a single assessment strategy has been adopted by the field—the childhood behavior checklist (or Achenbach system). In this case, there is broad consensus in the field about the characteristics of the disorder, and the

construct is well defined. The checklist includes permutations that allow it to be administered and scored from the point of view of the child or adolescent, the parents, or the teachers, which permits multiple sources of information in making a diagnosis. Hoyle noted it has been shown to be both valid and reliable. He highlighted Odgers' research documenting that early identification and intervention can vastly improve outcomes for people with these disorders. Several participants also called attention to the recently skyrocketing problems with bullying in schools and noted that early identification of conduct disorders may help reduce the incidence of this behavior.

The other two examples were of assessments still used for research purposes. Hoyle found the assessments of self-regulated learning that Tim Cleary is exploring to be both intriguing and promising. The assessment strategies allow the researchers to directly observe someone engaged in the activity of learning, and one of the alternatives that Cleary discussed is having children report online as they actually proceed through the learning process. Hoyle commented on the multitude of insights that can be obtained by having children report on what they are doing before they begin an activity, while they are engaged in the activity, and then reflecting on it afterward. Preliminary work suggests these measures are predictive of course grades. With regard to the assessments of emotional intelligence, Hoyle tended to agree with Matthews that the construct is not yet well defined, and questions remain about its distinction from personality. As Hoyle put it, the measures Matthews discussed tend to be highly correlated with personality to the extent that "one wonders if one really needs separate measures of emotional intelligence or if, in fact, one is able to capture that variability in standard personality measurement."

Fairness and Accessibility

A third issue discussed throughout the workshop was fairness. As explained in Chapter 5, in a testing context, fairness means the assessment should be designed so that test takers can demonstrate their proficiency on the targeted skill without irrelevant factors interfering with their performance. Fairness is an essential component of validity. Some of the constructs discussed during the workshop raised considerable concern about fairness and possible sources of bias. One issue alluded to previously in this chapter is whether the assessments are measuring the skills they purport to measure or are actually measuring personality traits or intelligence. To what extent is a domain-general conception of critical thinking distinct from general cognitive ability (intelligence)? To what extent are emotional intelligence and integrity distinct from personality?

There is some research to help answer these questions, but it is important to be clear on what exactly is being assessed.

Related to this is the notion of trainability or malleability: that is, that proficiency on the particular skill can increase as a consequence of training and practice. To what extent can a person learn to have more integrity, to become more self-regulated, or to have better social skills? Some students may come to school better prepared to collaborate with others or to manage their own learning. This may occur as a result of family background characteristics, home environment, or other out-of-school experiences. To what extent would assessments be measuring skills that can be learned in school versus family background? There is some research on these issues as well, but as Greg Duncan, professor of education with the University of California, Irvine, noted, the findings are not definitive. Related to this issue is the notion of opportunity to learn. If these skills are indeed trainable, to what extent will all students have equal exposure to instruction in the skills? If students are expected to acquire these skills and teachers are held accountable for teaching them, instructional programs will be needed so that students have the opportunity to learn them. This issue has direct bearing on fairness and ultimately on the validity of assessments. Workshop participants noted that these issues will need to be investigated and understood before moving into wide use of assessments of these skills, particularly if the results are used to make important decisions about students.

There were also considerable concerns about the issue of construct irrelevant variance, particularly as it relates to English language learners. Patrick Kyllonen, director of the Center for Academic and Workplace Readiness and Success at the Educational Testing Service, cited statistics that in the state of California, 25 percent of all public school students are English language learners, with the numbers increasing rapidly in other states as well (e.g., see National Research Council, 2011). For an assessment like the situational judgment test that presents a verbally dense description of a situation, language skills are critically important. For students with weak English language skills, the assessments would be a reading test, not a measure of interpersonal skills.

IMPLICATIONS FOR POLICY

Herman posed two additional questions to the group during the discussion session. If 21st century skills were included in assessments, what would the assessment system look like? And how would we go about implementing such a system?

In responding to the first question, she returned to her point about the many types of assessments and the many ways of using the results. She

highlighted the fact that throughout the workshop, participants repeatedly raised questions about the purposes of the assessments and the levels at which they would be used. In her view, the full spectrum of assessment purposes should be explored in determining ways to incorporate these skills into K-12 schooling. She said she would advocate for a system that included a variety of formative components intended both to guide instructional decision making and to enable early identification of potential problems. These might be combined with assessments used for a variety of summative purposes, including accountability for schools, teachers, and students, under the goal of ensuring students receive the exposure and engagement they need to develop the skills that are critical for college and workforce readiness.

In addressing the second question, she called for work to identify the constructs on which to focus. Throughout the workshop, a variety of skills and constructs were discussed, but as Herman put it, “we cannot do everything at once.” The initial work would be to identify the most critical skills and predispositions for students to learn, set priorities on what is most important, and then develop strategies for teaching and assessing them.

She referred to the Race to the Top (RTTT) assessment consortia² as one vehicle for moving this work forward. She said the changes enacted through the RTTT efforts provide a timely opportunity for bringing attention to new skills. The cognitive skills of critical thinking and problem solving, she noted, are already incorporated into the common core standards. The next step would be to make sure these skills are included in the curriculum and the assessments and then to encourage focus on some of the interpersonal and intrapersonal skills.

As part of this discussion, Patrick Kyllonen commented about the idea of “consequential validity” or the social/educational consequences of having the assessment in place and making use of the test results. There are many examples, he noted, of tests inserted into testing systems, not necessarily because they will improve psychometric properties, but because of the consequences they might bring about. An example would be the inclusion of writing assessments in many standardized assessments—such as the SAT, GRE, MCAT, and LSAT—despite the fact that they may not significantly improve the predictive validity of the assessment. In this case, the notion is that including an assessment of writing, and attaching stakes to it, should bring about an increased focus on developing writing skills, both by teachers in their instruction and by potential test takers as they prepare for the assessment. Currently, in K-12 education, Kyllonen continued, accountability systems revolve almost entirely

²See <http://www2.ed.gov/programs/racetothetop-assessment/index.html> [May 2011].

around the ability of students to take reading and math tests. Thus, one consequence of incorporating 21st century skills into the assessment or the accountability system would be to encourage teachers and students to spend more time on these skills. As characterized by one workshop participant, what is tested is taught, and what is not tested is not taught.

Herman also spoke about teacher and teaching capacity. She summarized comments from workshop participants who pointed out that the development of 21st century skills and their integration with academic content is not a regular feature of curriculum or instruction; in some school systems, there may be some focus on the cognitive skills, but this is certainly not the case for the interpersonal and intrapersonal skills. While some teachers may have experience with assigning grades for effort, attitude, and behavior, the interpersonal and intrapersonal skills discussed at the workshop go far beyond these measures. This means the teaching and assessment of 21st century skills will require changes in curriculum and teacher practices that will require a substantial amount of teacher development. As emphasis on these skills takes on new meaning, teachers would need a good deal of assistance both to understand the nature of these constructs and to learn how to develop them in their students so that all students have the opportunity to learn them. This has implications both for teacher preparation programs and for teacher inservice professional development.

Herman also called for transparency. She noted the changes required in curriculum, instruction, teacher training, and assessment can be made more smoothly by transparency. Being transparent will help teachers and students understand the skills that are being emphasized and will help the assessment developers better understand the skills that are to be measured.

Feasibility and Moving Forward

As one workshop participant pointed out, students in U.S. schools already spend considerable time taking tests. Many educators would not readily welcome the idea of adding more tests to the school day. However, this idea assumes the assessments would be something put upon students rather than an integrated part of the curriculum. The view of the assessments endorsed by Herman and other workshop participants was that the various constructs would be incorporated into the academic curriculum so that their teaching would be an integral part of the instructional program. For instance, it is not difficult to imagine incorporating a team project into the regular science, social studies, language arts, or mathematics program. Incorporating activities in which students must problem solve, think creatively, and communicate their work to others using multiple types of mediums seems natural in academic settings.

Adding ongoing formative assessments that help to guide instruction of these skills does not seem like a heavy burden to place on teachers and students. As John Behrens noted in describing the Packet Tracer, the system relies on “stealth assessments”; often students do not even realize they are being tested.

At the same time, other workshop participants stressed it is important not to lose sight of the need to ensure that students in the United States learn the basic academics. As Paul Sackett put it, “If we were at a different conference, we would be spending time lamenting the fact that students in the U.S. are not up to par on some fundamental academic skills.” Likewise, Deirdre Knapp noted all 21st century skills are not equal—some are clearly more important for students to learn than others, and we are further along in knowing how to assess some skills than others. Thus, it is critical to set priorities for where and how to spend the limited time, money, and resources.

Kyllonen also emphasized the importance of considering the cost tradeoffs. He noted the various examples of assessments included some “ingenious low-cost assessments and some dazzling high-cost assessments.” He encouraged work to study the differences in order to figure out where high-cost investment is cost-effective and where it might not make a difference. He and others pointed to examples other than those presented at the workshop that might be important resources and models. For instance, Herman mentioned the work that David Conley, with the Educational Policy Improvement Center (EPIC), has been doing to identify critical components of college and career readiness, as well as similar efforts by the National Assessment of Educational Progress (NAEP) to focus the 12th grade assessment on these skills. Kyllonen also spoke of the exams used to assess critical-thinking skills at the college level, such as the Collegiate Learning Assessment (CLA), the ACT CAP Test, and the ETS Proficiency Profile Test. They are all operational programs, he pointed out, that may serve as models. Knapp noted the work the military has been doing to evaluate temperament, persistence, and stamina. Others commented that while the Envision High School was featured at the workshop, a number of such high schools throughout the country are working to incorporate instruction and assessment of 21st century skills into the curriculum in innovative ways.

Defining the overall purpose of the assessments was an issue raised repeatedly in deciding on a path for moving forward. Sackett framed the issue as deciding between a focus on individual results or group-level results. He asked, “Do we want students to leave school with an individualized certificate that documents their level of competence in each skill? Or do we want to document how the nation is doing in aggregate?” He cautioned obtaining precise and reliable assessment at the individual level is difficult,

costly, and time consuming. On the other hand, Steve Wise questioned how best to address the different aspirations that students have. While there is currently a heavy emphasis on ensuring all students pursue higher education, in reality, that is not likely to occur. Students have different goals. Do we design a system that is a “one size fits all plan,” he asked, do we focus on minimal competency across the board, or do we design a system that attends to the specific needs of the individual?

Several workshop participants spoke of the types of research needed in order to move forward with assessments of these skills. Deirdre Knapp pointed out many assessments are “pushing the envelope” as far as psychometric capabilities. For example, how does one evaluate the reliability of assessments such as those used by Art Graesser’s Auto Tutor? Greg Duncan called for research in two areas. First, he noted, if we are to relate these skills to training in school, we need to know what it takes to change these skills. That is, how malleable are they and what is involved in improving them? Second, he called for more in-depth study of the predictive power of the various skills, noting that what is needed is not simply correlations among the variables but well-controlled analyses to demonstrate that improvement in these skills results in improvement in academic and labor market outcomes. Finally, Juan Sanchez, professor of management and international business at Florida International University, called for increased levels of cross-disciplinary efforts, stressing that successfully tackling these issues will require the collaboration of expertise from many disciplines including measurement, cognitive psychology, and information technology.

References

- Adams, M.H., Stover, L.M., and Whitlow, J.F. (1999). A longitudinal evaluation of baccalaureate nursing students' critical-thinking abilities. *Journal of Nursing Education*, 38, 139-141.
- Adey, P., Csapo, B., Demetriou, A., Hautamaki, J., and Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educational Research Review*, 2, 75-97.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Authors.
- Anastasi, A. (1988). *Psychological Testing* (6th Ed.). New York: Macmillan.
- Anderson, R.F. (2002, September 16). Deaths: Leading causes for 2000. *National Vital Statistics Report*, 50(16). Available: http://www.cdc.gov/nchs/data/nvsr/nvsr50/nvsr50_16.pdf [July 2011].
- Andrade, H.L., and Cizek, G.J. (2010). *Handbook of Formative Assessment*. New York: Routledge.
- Autor, D.H., Levy, F., and Murnane, R.J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279-1333.
- Bangert-Drowns, R.L., and Bankert, E. (1990). *Meta-analysis of Effects of Explicit Instruction for Critical Thinking*. Paper presented at the meeting of the American Educational Research Association, Boston, MA.
- Bauer, K.W., and Liang, Q. (2003). The effect of personality and precollege characteristics on first-year activities and academic performance. *Journal of College Student Development*, 44, 277-290.
- Behrens, P.J. (1996). The Watson-Glaser critical-thinking appraisal and academic performance of diploma school students. *Journal of Nurse Education*, 35, 34-36.
- Behrens, J.T., Mislevy, R.J., DiCerbo, K.E., Rutstein, D.W., and Levy, R. (in press). Evidence-centered design for learning and assessment in the digital e world. In M. Mayrath, J. Clarke-Midura, and D. Robinson (Eds.), *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. New York: Springer-Verlag.

- Berry, C.M., Sackett, P.R., and Wiemann, S.A. (2007). A review of recent developments in integrity test research. *Personnel Psychology*, 60, 271-301.
- Best, J.R., and Miller, P.H. (2010). A developmental perspective on executive function. *Child Development*, 81, 1641-1660.
- Bok, D. (2006). *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More*. Princeton, NJ: Princeton University Press.
- Bondy, K.N., Koenigseder, L.A., Ishee, J.H., and Williams, B.G. (2001). Psychometric properties of the California critical-thinking tests. *Journal of Nursing Measurement*, 9, 309-328.
- Brown, J.M. (1998). Self-regulation and the addictive behaviors. In W.R. Miller and N. Heather (Eds.), *Treating Addictive Behaviors* (2nd ed., pp. 61-73). New York: Plenum Press.
- Byham, B.C. (2010). *Rethinking Assessment Centers: Multiple Variations to Meet Multiple Needs*. Paper presented at the 35th International Congress on Assessment Center Methods, Singapore.
- Cano, J., and Martinez, C. (1991). The relationship between cognitive performance and critical-thinking abilities among selected agricultural education students. *Journal of Agricultural Education*, 32, 24-29.
- Caspi, A., Gegg, D., Dickson, N., Harrington, H., Langley, J., Moffitt, T.E., and Silva, P.A. (1997). Personality differences predict health-risk behaviors in young adulthood: Evidence from a longitudinal study. *Journal of Personality and Social Psychology*, 73, 1052-1063.
- Chamberlain, S.R., Mütter, U., Blackwell, A.D., Clark, L., Robbins, T.W., and Sahakian, B.J. (2006, February). Neurochemical modulation of response inhibition and probabilistic learning in humans. *Science*, 311(10).
- Cizek, G., and Bunch, M. (2007). *Standard Settings: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- Cleary, T.J., and Zimmerman, B.J. (2001). Self-regulation differences during athletic practice by experts, nonexperts, and novices. *Journal of Applied Sport Psychology*, 13, 185-206.
- Cleary, T.J., Zimmerman, B.J., and Keating, T. (2006). Teaching physical education students to self-regulate during basketball free throw practice. *Research Quarterly for Exercise and Sport*, 72, 452-463.
- Cleary, T.J., Peterson, J., Adams, T., and Callan, G. (2010). *Development and Validation of the Self-Regulation Microanalytic Interview*. Unpublished raw data.
- Connell, A., Dishion, T.J., Yasui, M., and Kavanagh, K. (2007). An adaptive approach to family intervention: Linking engagement in family-centered intervention to reductions in adolescent problem behavior. *Journal of Consulting and Clinical Psychology*, 75, 568-579.
- Costa, P.T., Jr., and McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and the NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart, and Winston.
- De Jong, D.F., and Das-Smaal, E.A. (1990). The start counting test: An attention test for children. *Personality and Individual Differences*, 11(6), 597-604.
- Dishion, T.J., and Connell, A. (2006). Adolescents' resilience as a self-regulatory process. *Annals of the New York Academy of Sciences*, 1094, 125-138.
- Edwards, T.B. (1950). Measurement of some aspects of critical thinking. *Journal of Experimental Education*, 18, 263-278.
- Ellis, L.K., Rothbart, M.K., and Posner, M.I. (2004). Individual differences in executive attention predict self-regulation and adolescent psychosocial behaviors. *Annals of the New York Academy of Sciences*, 1021, 337-340.
- Ennis, R.H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 43, 44-48.

- Facione, N.C., and Facione, P.A. (1996). Externalizing the critical thinking in knowledge and clinical judgment. *Nursing Outlook*, 44, 129-136.
- Facione, N.C., and Facione, P.A. (1997). *Critical-Thinking Assessment in Nursing Education Programs: An Aggregate Data-Analysis*. Millbrae: California Academic Press.
- Facione, P.A. (1990). *The California Critical-Thinking Skills Tests—College Level*. Technical report #2, Factors Predictive of CT Skills. Millbrae: California Academic Press.
- Facione, P.A., Giancarlo, C.A., Facione, N.C., and Gainen J. (1995). The disposition toward critical thinking. *Journal of General Education*, 44, 1-25.
- Facione, P.A., Facione, N.C., and Giancarlo, C.A. (1998). *Professional Judgment and the Disposition Toward Critical Thinking*. Millbrae: California Academic Press.
- Ferris, G.R., Witt, L.A., and Hochwarter, W.A. (2001). Interaction of social skill and general mental ability on job performance and salary. *Journal of Applied Psychology*, 86(6), 1075.
- Fitzsimons, G.M., and Bargh, J.A. (2004). Automatic self-regulation. In R.F. Baumeister and K.D. Vohs (Eds.), *Handbook of Self-Regulation: Research, Theory, and Applications* (pp. 151-170). New York: Guilford Press.
- Foster, E.M., and Jones, D.E. (2005). The high costs of aggression: Public expenditures resulting from conduct disorder. *American Journal of Public Health*, 95, 1767-1772.
- Gadzella, B.M., Baloglu, M., and Stephens, R. (2002). Prediction of GPA with educational psychology grades and critical-thinking scores. *Journal of Educational Psychology*, 3, 618-623.
- Gadzella, B.M., Stacks, J.M., and Stephens, R. (2004). *College Students Assess Their Stressors and Reactions to Stressors*. Paper presented at the Texas A&M University Assessment Conference, College Station, TX.
- Galagan, P. (2010). *Bridging the Skills Gap*. Alexandria, VA: American Society of Training and Development.
- Garon, N., Bryson, S.E., and Smith, I.M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, 134, 31-60.
- Giancarlo, C.A. (1996). *Critical Thinking, Culture, and Personality: Predicting Latinos' Academic Success*. Unpublished doctoral dissertation, Department of Psychology, University of California, Riverside.
- Gick, M.L., and Holyoak, K.J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15(1), 1-38.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8, 53-96.
- Goldin, C., and Katz, L.F. (2008). *The Race Between Education and Technology*. Cambridge, MA: Harvard University Press.
- Goldman-Rakic, P.S. (1987). Circuitry of primate prefrontal cortex and regulation of behaviour by representational memory. In F. Plum (Ed.), *Handbook of Physiology, Section 1: The Nervous System* (vol. 5, pp. 373-417). Bethesda, MD: American Physiological Society.
- Goleman, D. (1998). *Working with Emotional Intelligence*. New York: Bantam Books.
- Graesser, A., Britt, A., Millis, K., Wallace, P., Halpern, D., Cai, Z., Kopp, K., and Forsyth, C. (2010). Critiquing media reports with flawed scientific findings: Operation ARIES! A game with animated agents and natural language dialogues. In S. Alevan, J. Kay, and J. Mostow (Eds.), *ITS 2010 Part II*. Berlin, Heidelberg: Springer-Verlag.
- Hadwin, A.F., Winne, P.H., Stockley, D.B., Nesbit, J.C., and Woszczyzna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology*, 93, 477-487.
- Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.). *Educational Measurement, 4th Edition*. Westport, CT: American Council on Education/Prager.

- Halpern, D.F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53, 449-455.
- Heckman, J.J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312, 1900-1902.
- Herman, J.L., Aschbacher, P.R., and Winters, L. (1992). *A Practical Guide to Alternative Assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Hochwarter, W.A., Witt, L.A., Treadway, D.C., and Ferris, G.R. (2006). The interaction of social skill and organizational support on job performance. *Journal of Applied Psychology*, 91(2), 482-489.
- Hoffman, R.R. (1998). How can expertise be defined? Implications of research from cognitive psychology. In R. Williams, W. Faulkner, and J. Fleck (Eds.), *Exploring Expertise* (pp. 81-100). New York: Macmillan.
- Holland, P.W., and Dorans, N.J. (2006). *Linking and Equating*. In R.L. Brennan (Ed.), *Educational Measurement*, 4th Edition. Westport, CT: American Council on Education/Prager.
- Houston, J. (2007). *Future Skill Demands, from a Corporate Consultant Perspective*. Presentation at the Workshop on Research Evidence Related to Future Skill Demands, National Research Council. Available: http://www7.nationalacademies.org/cfe/Future_Skill_Demands_Presentations.html [March 2009].
- Ilkowska, M., and Engle, R.W. (2010). Working memory capacity and self-regulation. In R.H. Hoyle (Ed.), *Handbook of Personality and Self-Regulation* (pp. 265-290). Malden, MA: Blackwell.
- Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64. Westport, CT: American Council on Education and Praeger.
- Kantrowitz, T.M. (2005). *Development and Construct Validation of a Measure of Soft Skills Performance*. Unpublished dissertation, Georgia Institute of Technology, Atlanta, GA.
- Kitsantas, A., and Zimmerman, B.J. (2002). Comparing self-regulatory processes among novice, nonexpert, and expert volleyball players: A microanalytic study. *Journal of Applied Sport Psychology*, 14, 91-105.
- Klauer, K., and Phye, G. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, 78(1), 85-123.
- Klein, C., DeRouin, R.E., and Salas, E. (2006). Uncovering workplace interpersonal skills: A review, framework, and research agenda. In G.P. Hodgkinson and J.K. Ford (Eds.), *International Review of Industrial and Organizational Psychology* (vol. 21, pp. 80-126). New York: Wiley and Sons.
- Kobrin, J.L., Patterson, B.F., Shaw, E.J., Mattern, K.D., and Barbuti, S.M. (2008). *The Validity of the SAT for Predicting First-Year College Grade Average*. Research Report 2008-5. New York: College Board.
- Kolen, M., and Brennan, B. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. New York: Springer.
- Kowalski, P., and Taylor, A.K. (2004). Ability and critical thinking as predictors of change in students' psychological misconceptions. *Journal of Instructional Psychology*, 31, 297-303.
- Kuncel, N.R., and Grossbach, A. (2007). The predictive validity of nursing admission measures for performance on the national council licensure examination: A meta-analysis. *Journal of Professional Nursing: Official Journal of the American Association of Colleges of Nursing*, 27(2), 124-128.
- Kuncel, N.R., and Hezlett, S.A. (2007). Standardized test predict graduate students' success. *Science*, 315, 1080-1081.
- Kuncel, N.R., and Hezlett, S.A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, 19(6), 339-345.
- Levy, F., and Murnane, R.J. (2004). *The New Division of Labor: How Computers Are Creating the Next Job Market*. Princeton, NJ: Princeton University Press.

- Lewis, H. (2006). *Excellence without a soul: How a great university forgot education*. New York: Public Affairs/Perseus Books.
- Lievens, F., and Sackett, P.R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, 91, 1181-1188.
- Lievens, F., and Sackett, P.R. (2007). Situational judgment tests in high stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology*, 92, 1043-1055.
- Lievens, F., and Sackett, P. (2011). *The Validity of Interpersonal Skills Assessment via Situational Judgment Tests for Predicting Academic Success and Job Performance*. Manuscript submitted for publication.
- Lievens, F., Buyse, T., and Sackett, P.R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, 90, 442-452.
- Lievens, F., Buyse, T., and Sackett, P.R. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, 94, 1095-1101.
- Lindqvist, E., and Westman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, 3(1), 101-128.
- Logan, G.D., and Cowan, W.B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, 91, 295-327.
- Loughry, M.L., Ohland, M.W., and Moore, D.D. (2007). Development of a theory-based assessment of team member effectiveness. *Educational and Psychological Measurement*, 67(3), 505-524.
- Mayer, J.D., Salovey, P., and Caruso, D. (2000). Models of emotional intelligence. In R.J. Sternberg (Ed.), *Handbook of Intelligence* (pp. 396-420). New York: Cambridge University Press.
- Mayer, R.E. (1990). Problem solving. In M.S. Eysenck (Ed.), *The Blackwell Dictionary of Cognitive Psychology* (pp. 284-288). Oxford: Basil Blackwell.
- Mayer, R.E., and Wittrock, M.C. (1996). Problem-solving transfer. In D.C. Berliner and R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 47-62). New York: Simon & Schuster Macmillan.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). New York: Macmillan.
- Millis, K.K., Cai, Z., Graesser, A.C., Halpern, D., and Wallace, P. (2009). Learning scientific inquiry by asking questions in an educational game. In T. Bastiaens (Ed.), *Proceedings of World Conference on E-learning in Corporate, Government, Healthcare, and Higher Education* (pp. 2951-2956). Chesapeake, VA: AACE.
- Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., and Halpern, D. (in press). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, and J. Lakhmi (Eds.), *Serious Games and Edutainment Applications*. London: Springer-Verlag.
- Mischel, W. (1958). Preference for delayed reinforcement: An experimental study of a cultural observation. *Journal of Abnormal and Social Psychology*, 56, 57-71.
- Mischel, W., Ayduk, O., Berman, M., et al. (in press). "Willpower" over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*.
- Mislevy, R.J., and Rensvante, M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing and T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum.
- Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., and Wagner, T.D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49-100.

- Moffitt, T.E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, 100(4), 674-701.
- Moffitt, T.E., Caspi, A., Harrington, H., et al. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and Psychopathology*, 14, 179-207.
- Motowidlo, S.J., Dunnette, M.D., and Carter, G.W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75(6), 640-647.
- Nash, J., Collins, B., Loughlin, S., Solbrig, M., Harvey, R., Krishnan-Sarin, S., et al. (2003). Training the transdisciplinary scientist: A general framework applied to tobacco use behavior. *Nicotine and Tobacco Research*, 5(Suppl. 1), S41-S53.
- National Research Council. (2008). *Research on Future Skills Demands: A Workshop Summary*. M. Hilton, Rapporteur. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- National Research Council. (2010). *Exploring the Intersection of Science Education and 21st Century Skills: A Workshop Summary*. M. Hilton, Rapporteur. Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- National Research Council. (2011). *Allocating Federal Funds for State Programs for English Language Learners*. Panel to Review Alternative Data Sources for the Limited English Proficiency Allocation Formula under Title III, Part A., Elementary and Secondary Educational Act, Committee on National Statistics and Board on Testing and Assessment. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Nisbett, R.E., and Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Nock, M.K., Joiner Jr., T.E., Gordon, K., Lloyd-Richardson, E., and Prinstein, M.J. (2006.) Nonsuicidal self-injury among adolescents: diagnostic correlates and relation to suicide attempts. *Psychiatry Research* 144, 65-72.
- Odgers, C.L., Caspi, A., Broadbent, J.M., Dickson, N.P., Hancox, R., Harrington, H., et al. (2007). Prediction of differential adult health burden by conduct problem subtypes in males. *Archives of General Psychiatry*, 64, 1-9.
- Odgers, C.L., Milne, B., Caspi, A., Crump, R., Poulton, R., and Moffitt, T.E. (2007). Predicting prognosis for the conduct-problem boy: Can family history help? *Journal of the American Academy of Child and Adolescent Psychiatry*, 46, 1240-1249.
- Odgers, C.L., Moffitt, T.E., Broadbent, J.M., et al. (2008). Female and male antisocial trajectories: From childhood origins to adult outcomes. *Development and Psychopathology*, 20, 673-716.
- Organisation for Economic Co-operation and Development (2010, September). *PISA 2012 Problem-Solving Framework*. Draft for field trial. Available: <http://www.oecd.org/dataoecd/8/42/46962005.pdf> [August 2011].
- Peterson, N., Mumford, M., Borman, W., Jeanneret, P., and Fleishman, E. (1999). *An Occupational Information System for the 21st Century: The Development of O*NET*. Washington, DC: American Psychological Association.
- Petrides, K.V., and Furnham, A. (2003). Trait emotional intelligence: Behavioural validation in two studies of emotion recognition and reactivity to mood induction. *European Journal of Personality*, 17(1), 39-57.
- Popham, W.J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders*. Boston: Allyn and Bacon.
- Pulakos, E.D., Arad, S., Donovan, M.A., and Plamondon, K.E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology*, 85, 612-624.

- Riggio, R.E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology*, 51(3), 649-660.
- Roberts, B.W., and DelVecchio, W.F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, 126, 3-25.
- Roberts, R.D., Zeidner, M., and Matthews, G. (2007). Emotional intelligence: Knowns and unknowns. In G. Matthews, M. Zeidner, and R.D. Roberts (Eds.), *The Science of Emotional Intelligence: Knowns and Unknowns* (pp. 419-474). New York: Oxford University Press.
- Rothbart, M.K., and Bates, J.E. (2006). Temperament. In W. Damon, R.L. Lerner (Series Eds.), and N. Eisenberg (Vol. Ed.), *Handbook of Child Psychology: Vol. 3: Social, Emotional, and Personality Development* (6th ed., pp. 99-166). New York: Wiley.
- Rothbart, M.K., and Hwang, J. (2002). Measuring infant temperament. *Infant Behavior and Development*, 25, 113-116.
- Sackett, P.R., and Decker, P.J. (1979). Detection of deception in the employment context: A review and critical analysis. *Personnel Psychology*, 32, 487-506.
- Sackett, P.R., and Harris, M.M. (1984). Honesty testing for personnel selection: A review and critique. *Personnel Psychology*, 37, 221-245.
- Sackett, P.R., and Wanek, J.E. (1996). New developments in the use of measures of honesty integrity, conscientiousness, dependability, trustworthiness, and reliability for personnel selection. *Personnel Psychology*, 49, 787-829.
- Sackett, P.R., Burris, L.R., and Callahan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology*, 42, 491-529.
- Salomon, G., and Perkins, D.N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, 24, 113-142.
- SCANS. (1999). *Skills and Tasks for Jobs*. Washington, DC: U.S. Department of Labor.
- Schneider, R.J., Ackerman, P.L., and Kanfer, R. (1996). To "act wisely in human relations": Exploring the dimensions of social competence. *Personality and Individual Differences*, 21, 469-481.
- SH&A/Fenestra. (2007). *E-evaluation™ Technology-enhanced Assessment Centers*. Available: <http://www.fenestrainc.net/> [August 2011].
- Shavelson, R.J., and Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.
- Sherer, M., Maddux, J.E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., and Rogers, R.W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports*, 51, 663-671.
- Smith-Jentsch, K.A., Salas, E., and Baker, D.P. (1996). Training team performance-related assertiveness. *Personnel Psychology*, 49, 909-936.
- Sonnentag, S., and Lange, I. (2002). The relationship between high performance and knowledge about how to master cooperation situations. *Applied Cognitive Psychology*, 16, 491-508.
- Spector, P.E., Schneider, J.R., Vance, C.A., and Hezlett, S.A. (2000). The relation of cognitive ability and personality traits to assessment center performance. *Journal of Applied Social Psychology*, 30, 1474-1491.
- Sternberg, R.J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology*, 49(3), 607-627.
- Stevens, M.J., and Campion, M.A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management*, 25(2), 207-228.
- Taggar, S., and Brown, T.C. (2001). Problem-solving team behaviors. *Small Group Research*, 32(6), 698-726.
- Taube, K.T. (1997). Critical-thinking ability and disposition as factors of performance on a written critical-thinking test. *Journal of General Education*, 46, 129-164.
- Thorndike, R.K. (1920). Intelligence and its uses. *Harper's Magazine*, 140, 227-335.

- Traub, R. (1994). *Reliability for the Social Sciences*. Thousand Oaks, CA: Sage.
- Unsworth, N., and Engle, R.W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, 114, 104-132.
- Watson, G.B., and Glaser, E.M. (1994). *Watson-Glaser Critical-Thinking Appraisal Form-S Manual*. San Antonio, TX: Harcourt Brace and Psychological Corporation.
- Williams, R.L. (2003). *Critical Thinking as a Predictor and Outcome Measure in a Large Undergraduate Educational Psychology Course*. University of Tennessee. Available: <http://eric.ed.gov/PDFS/ED478075.pdf> [August 2011].
- Winne, P.H., and Jamieson-Noel, D.L. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology*, 28, 259-276.
- Winne, P.H., and Perry, N.E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich, and M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 532-568). Orlando, FL: Academic Press.
- Yarnall, L., and Ostrander, J. (in press). The assessment of 21st century skills. In C. Secolsky (Ed.), *Measurement, Assessment, and Evaluation in Higher Education*. New York: Routledge.
- Zeiky, M., Perie, M., and Livingston, S. (2008). *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupations Tests*. Princeton, NJ: Educational Testing Service.
- Zimmerman, B.J. (2000). Attaining self-regulation: A social-cognitive perspective. In M. Boekaerts, P. Pintrich, and M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13-39). Orlando FL: Academic Press.

Appendix A

Agenda and Participants for the January Workshop

Assessment of 21st Century Skills
Workshop
January 12–13, 2011

University of California, Irvine
Beckman Conference Center
Huntington Room

AGENDA

Wednesday, January 12

- 9:30-9:40 Opening Remarks**
Welcome
 Stuart Elliott (Director, Board on Testing and
 Assessment)
 Bruce Fuchs (National Institutes of Health, cosponsor
 of the project)
- Overview of Workshop**
 Joan Herman (CRESST, Chair of Workshop Steering
 Committee)
- 9:40-12:15 Session 1: Background Information**
 Moderators: Joan Herman and Pat Kyllonen (ETS and
 Workshop Steering Committee)
- (9:40-10:00) Why Are 21st Century Skills Important?**
 Richard Murnane (Harvard University)

This presentation will address the following issues:

- What is unique in the 21st century that makes these skills especially valuable in the labor market and/or in other life domains (learning, family life, civic engagement)?
- How does the growing use of computers and technology affect the labor market and the demand for 21st century skills?
- What does more recent research suggest about the skills needed to be successful in the 21st century?

(10:00-10:15) How Will You Know If Your Students Are 21st Century Ready?

Deborah Boisvert (Boston Area Advanced Technical Education)

The presenter will respond to the opening presentation, reflecting her work with employers to define, teach, and assess 21st century skills of computer technicians.

(10:15-10:45) The Teaching and Learning of 21st Century Skills

Eric Anderman (Ohio State University)

This presentation will address the following issues:

- What is known about the extent to which the three skill clusters and/or the skills within them can be taught and learned?
- To what extent are learning, teaching, and assessment of the three skill clusters domain specific or domain general?

(10:45-11:00) Discussion

Moderators will lead a question-and-answer session with the presenters and audience members.

11:00-11:15 Break

(11:15-11:35) Approaches to Developing Assessments of 21st Century Skills

Deirdre Knapp (HumRRO, Workshop Steering Committee)

This presentation will address the following issues:

- What are the different approaches to assessment of these skills and what steps are involved in carrying out these approaches?
- What processes are used for identifying the skills to be measured, operationalizing the skills through the test blueprint, and creating assessment tasks and scoring procedures?
- How should the intended uses of the assessment results guide the test development process?
- What steps should be taken to ensure that the assessments are reliable and valid?

(11:35-11:55) Unique Challenges and Opportunities in the Assessment of 21st Century Skills

Steven Wise (Northwest Evaluation Association, Workshop Steering Committee)

This presentation will address the following issues:

- What are the unique challenges and opportunities for defining and measuring these constructs, when compared to more traditional academic skills and knowledge?
- How might the results of these assessments be used? Should they be used for high-stakes purposes?
- What issues may arise in relation to the validity, reliability, and fairness of assessments of these skills?

(11:55-12:15) Questions and Discussion

Moderators will lead a question-and-answer session with the presenters and audience members.

12:15-1:15 Lunch in the Beckman Center Cafeteria

Continued discussion of ideas presented during the morning sessions

1:15-3:45 Session 2: Assessing Cognitive Skills

Moderators: Greg Duncan (University of California, Irvine, Workshop Steering Committee) and Paul Sackett (University of Minnesota, Workshop Steering Committee)

(1:15-1:45) Defining and Measuring Cognitive Skills

Nathan Kuncel (University of Minnesota)

This presentation will address the following issues:

- What are 21st century cognitive skills? To what extent do they differ from each other and from general cognitive ability? What are the conceptual differences that are proposed to exist between these constructs?
- What are the existing measures of these constructs, and to what extent do these existing measures match their conceptual specifications?
- What are the relationships between the existing measures of these constructs?

(1:45-2:00) Questions and Discussion

Moderators will lead a question-and-answer session with the presenters and audience members.

(2:00-3:15) Panel Discussion: Examples of Assessments of Cognitive Skills

For each example, the panelists will address the following issues:

- What skill or skills are measured? Why are these skills important?
- What is the purpose of the assessment?
- What strategies were used to develop the assessment and why were these selected?
- What assessment methods are used and why were these selected?
- How is the assessment scored? What data are available on the technical quality of the assessment, including validity, reliability, fairness, and comparability across administrations?
- What data are available on the cost and practical feasibility of the assessment?

(2:00-2:20) Interactive Problem Solving for PISA 2012

Joachim Funke (University of Heidelberg), by video conference

(2:20-2:40) Operation ARIES!: Learning Critical Thinking about Science with Intelligent Conversational Agents in a Game Environment

Art Graesser (University of Memphis) and Heather Butler (Claremont McKenna College)

(2:40-3:00) Intrusive and Unobtrusive Assessment of Entrepreneurial and Technical Skills through Simulation and Gaming

John Behrens (Cisco Systems)

(3:00-3:20) Assessment of Critical Thinking and Problem Solving on the Multistate Bar Exam

Susan Case (National Conference of Bar Examiners)

3:20-3:30 Break**(3:30-4:00) Moderated Discussion**

Moderators will explore the following issues with panelists and audience members:

- What are the implications of the presentations (and examples) for the design of 21st century assessments for K-12 and higher education?
- Do common themes or approaches emerge from the examples? How might the noneducation examples generalize to education?
- How might 21st century assessments be incorporated into current research efforts, such as the development of assessment systems by the two-state consortia? What functions can/should the assessments serve? How might the results be used?
- What equity and accessibility challenges do these assessments raise?
- What barriers might slow development and/or use of assessments of 21st century skills? How might they be overcome?

4:00-5:00 Synthesis of Key Ideas
Moderator: Joan Herman

(4:00-4:20) Discussion: Eva Baker (CRESST)

(4:20-4:40) Discussion: Richard Murnane (Harvard University)

Discussants will reflect on the day's discussions and offer their synthesis of the ideas presented. Audience members will be invited to ask questions and share their ideas as well.

5:00 Conclude Formal Agenda for Day 1
Joan Herman

5:30 Working Group Dinner (at Beckman Center)
Plan for the second day of the workshop

Thursday, January 13

9:00-11:45 Session 3: Assessing Interpersonal Skills

Moderators: Deirdre Knapp and Juan Sanchez (Florida International University, Workshop Steering Committee)

(9:00-9:30) Defining and Measuring Interpersonal Skills
Steve Fiore (University of Central Florida)

This presentation will address the following issues:

- What are 21st century interpersonal skills and why are they important?
- How are these skills typically assessed? What are the challenges in assessing them?
- What types of assessments are currently available to evaluate these skills?

(9:30-9:40) Questions and Discussion

Moderators will lead a question-and-answer session with the presenters and audience members.

(9:40-11:00) Panel Discussion: Examples of Assessments of Interpersonal Skills

For each example, the panelists will address the following issues:

- What skill or skills are measured? Why are these skills important?
- What is the purpose of the assessment?
- What strategies were used to develop the assessment and why were they selected?
- What assessment methods are used and why were these selected?
- How is the assessment scored? What data are available on the technical quality of the assessment, including validity, reliability, fairness, and comparability across administrations?
- What data are available on the cost and practical feasibility of the assessment?

(9:40-10:00) Online Portfolio Assessments of the 4 Cs
Bob Lenz (Envision Schools)

(10:00-10:20) 21st Century Skills in STEM Workforce Training Assessments
Louise Yarnall (SRI)

(10:20-10:40) Using Situational Judgment Tests for Medical School Admissions
Filip Lievens (Ghent University, Belgium), by video conference

(10:40-11:00) Assessment Centers 2011: Fifty Years of Best Practice and Today's Innovations
Lynn Gracin Collins (SH&A/Fenestra)

11:00-11:10 Break

(11:10-11:45) Moderated Discussion

Moderators will explore the following issues with panelists and audience members:

- What are the implications of the presentations (and examples) for the design of 21st century assessments for K-12 and higher education?
- Do common themes or approaches emerge from the examples? How might the noneducation examples generalize to education?
- How might 21st century assessments be incorporated into current research efforts, such as the development of assessment systems by the two-state consortia)? What functions can/should the assessments serve? How might the results be used?
- What equity and accessibility challenges do these assessments raise?
- What barriers might slow development and/or use of assessments of 21st century skills? How might they be overcome?

11:45-12:45 Working Lunch in the Beckman Center Cafeteria

Continued discussion of ideas presented during the morning sessions

12:45-3:30 Session 4: Assessing Intrapersonal Skills

Moderators: Pat Kyllonen and Steven Wise

(12:45-1:15) Assessment of Self-Regulation and Related Constructs: Prospects and Challenges

Rick Hoyle (Duke University)

This presentation will address the following issues:

- What are 21st century intrapersonal skills and why are they important?
- How are these skills typically assessed? What are the challenges in assessing them?
- What types of assessments are currently available to evaluate these skills?

(1:15-1:30) Discussion

(1:30-3:30) Panel Discussion: Examples of Assessments of Intrapersonal Skills

For each example, the panelists will address the following issues:

- What skill or skills are measured? Why are these skills important?
- What is the purpose of the assessment?
- What strategies were used to develop the assessment and why were these selected?
- What assessment methods are used and why were these selected?
- How is the assessment scored? What data are available on the technical quality of the assessment, including validity, reliability, fairness, and comparability across administrations?
- What data are available on the cost and practical feasibility of the assessment?

(1:30-1:50) Integrity Testing for Employee Selection

Paul Sackett (University of Minnesota, Workshop Steering Committee)

(1:50-2:10) Targeting Context-Specific Self-Regulated Learning (SRL) Processes: An Overview and Illustration of SRL Microanalysis

Tim Cleary (University of Wisconsin–Milwaukee)

(2:10-2:30) Assessing Behavioral Problems That Predict Poor Educational and Life Outcomes

Candice Odgers (University of California, Irvine)

(2:30-2:50) Out of the Maze? In Search of Skills for Emotional Intelligence

Gerald Matthews (University of Cincinnati)

2:50-3:00 Break

(3:00-3:30) Moderated Discussion

Moderators will explore the following issues with panelists and audience members:

- What are the implications of the presentations (and examples) for the design of 21st century assessments for K-12 and higher education?
- Do common themes or approaches emerge from the examples? How might the noneducation examples generalize to education?
- How might 21st century assessments be incorporated into current research efforts, such as the development of assessment systems by the two-state consortia? What functions can/should the assessments serve? How might the results be used?
- What equity and accessibility challenges do these assessments raise?
- What barriers might slow development and/or use of assessments of 21st century skills? How might they be overcome?

3:30-4:00 Session 5: Reflection and Synthesis
Moderated discussion led by workshop steering committee

4:00 Closing Remarks, Adjourn
Joan Herman

PARTICIPANTS

Eric Anderman, Ohio State University
 John Behrens, Cisco Systems
 Lola Berber-Jimenez, California Polytechnic Science Project
 Paul Bloomberg, Transformative Inquiry Design for Effective Schools and Systems
 Deborah Boisvert, University of Massachusetts, Boston
 Liane Brouillette, University of California, Irvine
 Christopher Brown, Pearson Foundation
 Peggy Burke, Transformative Inquiry Design for Effective Schools and Systems
 Heather Butler, Claremont McKenna College

Susan Case, National Conference of Bar Examiners
 Tim Cleary, University of Wisconsin–Milwaukee
 Sara Clough, ACT, Inc.
 Lynn Gracin Collins, Sandra Hartog & Associates/Fenestra, Inc.
 Emily Dalton Smith, Gates Foundation
 Tran Dang, University of California, Irvine
 Greg J. Duncan, University of California, Irvine
 Steve Fiore, University of Central Florida
 Dennis Frezzo, Cisco Systems
 Bruce Fuchs, National Institutes of Health
 Joachim Funke, University of Heidelberg
 Tracy Gardner, General Educational Development Testing Service
 Nicole Gerardi, University of California, Los Angeles
 Art Graesser, University of Memphis
 Valerie Greenhill, e-luminate
 Erika Hall, Pearson Foundation
 Joan Herman, National Center for Research on Evaluation, Standards,
 and Student Testing, University of California, Los Angeles
 (committee chair)
 Rick Hoyle, Duke University
 John Jackson, National Science Foundation
 Stuart Kahl, Measured Progress
 Deirdre J. Knapp, HumRRO (committee member)
 Art Kramer, University of Illinois
 Brandi Kujala, Educational Policy Improvement Center
 Nathan Kuncel, University of Minnesota
 Patrick Kyllonen, Educational Testing Service (committee member)
 Robert Lenz, Envision Schools
 Filip Lievens, University of Ghent
 María Alicia López Freeman, California Science Project
 Tim Magner, Partnership for 21st Century Skills
 Michael Martinez, University of California, Irvine
 Gerald Matthews, University of Cincinnati
 Mick McManus, University of Queensland
 Beth Miller, Nellie Mae Education Foundation
 Julia Rankin Morandi, Los Angeles Education Partnership
 Richard Murnane, Harvard University (committee member)
 Suzanne Nakashima, California Science Project
 Paul Nichols, National Center for the Improvement of Educational
 Assessment
 Candice Odgers, University of California, Irvine
 Cornelia Orr, National Assessment Governing Board

Pamela Paek, National Center for the Improvement of Educational
Assessment

Jason Ravitz, Buck Institute for Education

Michael Russell, University of California, Irvine

Paul Sackett, University of Minnesota (committee member)

Andrea Saenz, U.S. Department of Education

Juan I. Sanchez, Florida International University (committee member)

Mary Seburn, Educational Policy Improvement Center

Brian Stecher, RAND

Christine Tell, Achieve

Cathy Tran, University of California, Irvine

Bernie Trilling, Oracle Education Foundation

Jerry Valadez, California State University, Fresno

Marjorie Wine, General Educational Development Testing Service

Steven Wise, Northwest Evaluation Association (committee member)

Louise Yarnall, SRI

Raymond Yeagley, Northwest Evaluation Association

Linda Zimmerman, Pearson

Doron Zinger, Olive Crest Academy

Appendix B

Agenda and Participants for the May Workshop

Assessment of 21st Century Skills
Workshop Follow-Up Symposium
May 4, 2011

Keck Center, Room 100
500 Fifth St., NW
Washington, DC

AGENDA

Wednesday, May 4

1:00-1:10 Introductions, Overview of Plans
Stuart Elliott, BOTA director
Gerhard Salinger, National Science Foundation (cosponsor
of project)

1:10-1:30 Brief Review of the Workshop in January
Joan Herman, CRESST (steering committee chair)

- What are 21st century skills and why are they important?
- How do the skills relate to college and career readiness/preparedness?
- What are the challenges in assessing these skills?

For information about the January workshop, see http://www7.nationalacademies.org/bota/Assessment_of_21st_Century_Skills_Homepage.html.

1:30-2:30 Panel Discussion of Sample Assessments of 21st Century Skills (20 minutes each)

Assessments of Cognitive Skills

Nathan Kuncel, University of Minnesota

Assessments of Interpersonal Skills

Stephen M. Fiore, University of Central Florida

Assessments of Intrapersonal Skills

Rick Hoyle, Duke University

Each panelist will

- Briefly describe/define the kind of skills grouped within their cluster.
- Give a quick overview of why the skills are important skills.
- Give a synthetic overview of the assessment examples that were presented at the January workshop and provide a critique/reaction to them.
- Discuss the extent to which the example assessments (or assessment strategies) are likely to provide reliable and valid information about the intended skill.

2:30-2:45 **Moderated Discussion**

Discussion Leader: Joan Herman

2:45-3:15 **Response: Measurement Guidance** (15 minutes each)

Deirdre Knapp, HumRRO (steering committee member)

Patrick Kyllonen, ETS (steering committee member)

Speakers will respond to the panelists and address the following:

- From a measurement/technical perspective, what is the feasibility of implementing assessments of these kinds of skills in the K-12 setting?
- What purposes can they feasibly serve? How might the results be used?
- What factors might complicate implementation of the assessment or assessment strategy?
- What fairness and equal access issues should be considered?

3:15-3:45 Response: Policy Guidance (15 minutes each)

Joan Herman

Steven Wise, Northwest Evaluation Association (steering committee member)

Speakers will respond to the panelists and address the following:

- What are the key messages for policy makers with regard to implementing assessments of these skills in the K-12 setting?
- What would you like policy makers to know about the assessments, the assessment strategies, and/or implementing measures of these constructs?

3:45-4:15 Moderated Discussion

Joan Herman

4:15 Adjourn

PARTICIPANTS

Geri Anderson-Nielsen, Gender Equity for Mathematics and Science

Nancy Smith Brooks, U.S. Department of Education

Christopher Brown, Pearson Foundation

Rex Clemmenssen, ACT, Inc.

Sara Clough, ACT, Inc.

Debbie Cole

Christopher Coro, U.S. Department of Education

Roman Czujko, American Institute of Physics

Emily Dalton Smith, Bill & Melinda Gates Foundation

George DeBoer, American Association for the Advancement of Science

Mary E. Dilworth, National Board for Professional Teaching Standards

Nancy Doorey, Education Testing Service, Center for K-12 Assessment and Performance Management

Emerson Elliott, National Council for Accreditation of Teacher Education

Maria Ferguson, Alliance for Excellent Education

Steven Fiore, University of Central Florida

Bruce Fuchs, National Institutes of Health

Gavin Fulmer, National Science Foundation

Peirce Hammond, U.S. Department of Education

Mark Heidorn, CTB/McGraw-Hill
 Andres Henriquez, Carnegie Corporation of New York
 Monica Herk, National Board on Education Sciences
 Joan Herman, National Center for Research on Evaluation, Standards,
 and Student Testing, University of California, Los Angeles
 (committee chair)
 Ricardo Hernandez, U.S. Department of Education
 Jeffrey Heyck-Williams, Two Rivers Public Charter School
 Rick Hoyle, Duke University
 Tom Keller, National Research Council
 Bill Kelly, American Society for Engineering Education
 Dana Kelly, National Center for Education Statistics
 Arthur Kendall, Social Research Consultants
 Jonathan King, National Institute on Aging
 Deidre Knapp, HumRRO (committee member)
 Ken Krehbiel, National Council of Teachers of Mathematics
 Pat Kyllonen, Educational Testing Service (committee member)
 Emily Lai, Pearson Foundation
 Natalie Nielsen, National Research Council
 Cornelia Orr, National Assessment Governing Board
 Stephen Provasnik, National Center for Education Statistics
 Taslima Rahman, U.S. Department of Education
 Laura Rasmussen, MPR Associates, Inc.
 Patrick Rooney, U.S. Department of Education
 Gerhard Salinger, National Science Foundation
 Gretchen Schultz, CTB/McGraw-Hill
 Elena Silva, Education Sector
 Candace Simon, Council of Great City Schools
 Grace Solares, U.S. Department of Education
 Gerald Sroufe, American Education Research Association
 Barbara Stein, National Education Association
 James Stone, National Research Center for Career and Technical
 Education
 Peter Swerdzewski, Regents Research Fund
 Marjorie Wine, General Educational Development Testing Service
 Steven Wise, Northwest Evaluation Association (committee member)