Empirical Research in Vocational Education and Training, Vol. 4(1), 2012, 77–90

Assessing business-planning competence using the Collegiate Learning Assessment as a prototype

Richard J. Shavelson

*Stanford University and SK Partners, LLC **

Abstract

The assessment of competence has emerged as a critical issue as nations seek to develop their education systems, work forces, and their citizens' capacity for life-long learning. While the assessment of competence may be critical to national well-being in the 21st century, as Hartig, Klieme and Leutner (2008, p. v) pointed out, «the theoretical modeling of competencies, their assessment, and the usage of assessment results in practice present new challenges for psychological and educational research». The purpose of this paper is to move the field forward by providing one possible model for assessing competence. The model presented here underlies the Collegiate Learning Assessment (CLA). Drawing on lessons from the CLA, the model was used to develop a concrete prototype for assessing business-planning competence.

1. Brief overview of the model of competence measurement

In broad terms, competence is a «…complex ability … that … [is] closely related to performance in real-life situations» (Hartig, Klieme & Leutner, 2008, p. v; see also McClelland, 1973; Weinert, 2001). In the approach about to be described, competence is further delineated by a set of seven facets. These facets carve out the domain in which measures of competence – their tasks, response formats, and scoring – might be developed. This approach is exemplified with the Collegiate Learning Assessment (CLA) and then applied to a business-planning task. While not the focus of this paper, but for completeness of setting forth the model (see Shavelson, 2010b; in press), assuming an indefinitely large number of possible forms of a competence measure, a particular competence test may be viewed as a sample of tasks and responses from a large domain. Under some reasonable assumptions the assessment

* Richard J. Shavelson, Stanford University and SK Partners, LLC, United States of America; richs@stanford.edu

This paper was written while I was a Humboldt Fellow at the Institute for Human Resource Education and Management, Munich School of Management, Ludwig-Maximillians-Universität. I wish to thank to both the Humboldt Foundation and the Institute, especially Prof. Dr. Susanne Weber, for their support.

tasks/responses[1] and the raters who score test-takers' performance can be considered as randomly sampled. With this assumption, a statistical theory for modeling the reliability and validity of competence scores, generalizability (G) theory (e.g. Shavelson & Webb, 1991) can be used to evaluate the quality of the competency measurement.

2. Sketch of the construct of competence and measurement model

The construct, competence, a complex ability closely related to performance in real-life settings, can be characterized by the following seven facets (cf. Shavelson, 2010, in press; Weinert, 2001):

(1) complexity: a complex physical and/or intellectual ability or skill
(2) performance: a capacity not just to «know» but also to be able to do or perform
(3) standardization: tasks, responses, scoring-rubric, testing conditions (etc.) are the same for all individuals
(4) fidelity: tasks provide a high fidelity representation of situations in which competence is to be demonstrated in the real world
(5) level: performance meets some level of «good enough» to be competent
(6) improvement: the abilities and skills measured can be improved over time by education, training, and deliberative practice
(7) disposition: personal and social characteristics such as identity, perspective taking, self-regulation, social responsibility that motivate high levels of learning and performance

When these facets are combined, a mapping sentence that defines a measure of competence can be written as follows:

> A measure of competence should tap complex physical and/or intellectual *abilities and skills* to produce *observable performance* on a common *standardized* set of tasks that simulate with high *fidelity* the performances that are expected to be enacted in the «real world» («criterion») situations to which inferences of competence are to be drawn, with scores reflecting the *level* of performance (mastery or continuous) on tasks where *improvement* can be made through *dispositions* for self-regulation, learning, and deliberative practice.

1 Task refers to a situation, problem, or decision to be made that is presented to the test taker. Response refers to action taken by the test taker as demanded by the task. The distinction is important because not only tasks but also responses can sample what is required in the criterion situation in which behavior is concerned. Or tasks and responses can be quite removed from reality. For example, in multiple-choice questions, typically the stem presents a very short, synoptic task and the response is to select among 4 or 5 alternatives. Neither is a high fidelity representation of most criterion situations in life. Because referring to task/response is awkward, I use task throughout the paper but in doing so to mean both task and response.

This definition of competence and these seven facets circumscribe the domain of tasks, responses and scoring upon which inferences about competent performance are to be drawn. These tasks, responses, raters' scores and the like are sampled to form a particular measure of competence. Note that multiple assessments can be built by repeated sampling from the domain. Standards are set as to the level of performance at or above which a person would be considered to be competent in a domain such as business planning. Assuming sampling of tasks, raters, and the like is random, statistical models such as generalizability theory (e.g. Webb, Shavelson & Haertel, 2007) can be used to model and evaluate interpretations of the measurement, i.e. as a measure of, say, business-planning competence.

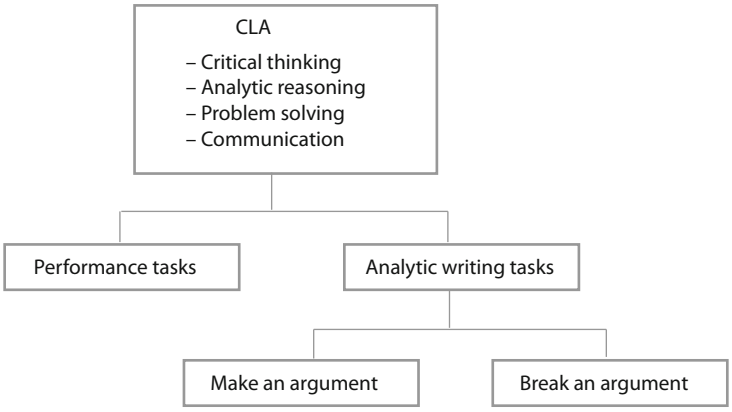
In addition, qualitative methods are needed to examine validity claims. These methods include, for example, experts' judgments of the representativeness and quality of tasks sampled (content validity; e.g. Wigdor & Green, 1991) and the extent to which thinking and reasoning processes underlying performance support inferences of competence (cognitive validity using, for example, a think-aloud method; e.g. Ericsson & Simon, 1993).

3. The Collegiate Learning Assessment (CLA): one possible prototype for competence measurement

The CLA was developed to measure college undergraduates' learning – in particular their learning to think critically, reason analytically, solve problems and communicate clearly (Shavelson, 2008, 2010a). The assessment focuses on the institution or on programs within an institution, not on individual students' performance (although students receive their scores, confidentially). Institution- or program-level scores are reported, both in terms of observed performance and as value added beyond what would be expected from entering students' ability (SAT or ACT college admissions test) scores.

The assessment has two major components: a set of performance tasks and as set of two different kinds of analytic writing prompts (Figure 1). Performance tasks pose a problem or decision to be addressed and students use a mix of information to solve the problem or recommend a course of action, based on evidence. One analytic writing prompt asks students to take a position on a topic and the other asks students to critique an argument. SK Partners' description of the CLA focuses on the performance task. This paper focuses on performance tasks.

Figure 1. Overview of the Collegiate Learning Assessment



The CLA, philosophically and theoretically, differs substantially from most college learning assessments, like ETS' Measure of Academic Proficiency and Progress (MAPP)[2] and ACT's Collegiate Assessment of Academic Progress (CAAP). Consequently, let's begin with the familiar. Most learning assessments grow out of an empiricist philosophy and a psychometric/behavioral tradition (Case, 1996; Shavelson, 2008a). From this stance, everyday complex tasks are divided into components. Then each component is analyzed to identify the abilities required for successful performance. Let's suppose that components such as critical thinking, problem solving, analytic reasoning and written communication are identified. A separate measure of each ability would then be constructed and students would take each subtest. At the end of testing, students' scores on the tests would be added up (or scaled) to construct a total score. The total score is used to describe students' performance not only on the assessment in hand but also on all possible assessments that might be created by sampling from the universe of complex tasks in the competence domain. That is, the generalization goes beyond the particular competence measure to all possible such measures that could have been created – we are interested in competence in the large domain and not just on a particular sample of tasks.

In contrast, the CLA is based on a combination of rationalist and socio-historical philosophies in the cognitive-constructivist and situated-in-context traditions (e.g. Case, 1996; Shavelson, 2008b). The CLA's conceptual underpinnings are embodied in what has been called by McClelland (1973) as a criterion sampling approach to measurement (Table 1). This approach assumes that the whole is greater than the sum of the parts and that complex tasks require the integration of abilities and skills that cannot be captured when divided into and measured as individual components and then added up.

2 ETS changes the name of its college learning assessment frequently enough so that it has become difficult to keep up with the current name. The MAPP may now be called the ETS Proficiency Profile.

Table 1. CLA's criterion sampling approach

Criterion sampling approach	Collegiate Learning Assessment (CLA)
Sample tasks from «real-world» domains	Samples holistic, real-world tasks drawn from life experiences
Sampe operant as well as respondent responses	Samples constructes responses (not multiple-choice)
Elicits complex abstract thinking («operant through patterns»)	Elicits critical thinking, analytic reasoning, problem solving and communication
Provides information on how improve on tasks («cheating» is not possible if can do criterion task)	Provides instructors with tasks for teaching as well as assessment

The Collegiate Learning Assessment's criterion-sampling approach

The criterion-sampling notion is simple and goes like this. If you want to know what a person knows and can do, sample tasks from the domain in which that person is to act, observe her performance, and infer competence and learning. For example, if you want to know if a person not only knows the laws for driving a car but also whether he can drive a car, don't give him only a multiple-choice test. That works fine for testing his knowledge of driving laws. Rather, also put him behind the wheel of a car and give him a driving test. The test would sample tasks from the general driving domain such as starting a car, pulling into traffic, turning right and left in traffic, backing up, and parking. Based on this sample of performance it is possible to draw inferences about his driving performance more generally.

The CLA followed the criterion-sampling approach by defining a domain of real-world tasks that are holistic, and drawn from life situations (as will be seen). It samples tasks and collects students' *operant responses*. Operant responses are student-generated responses that get modified with feedback as the task is carried out. These responses parallel those expected in the real world. There are no multiple-choice items in the assessment; life does not present itself as a set of alternatives with only one correct course of action. Finally, the CLA helps college professors create CLA-like tasks so the instructors can «teach to the test». With the criterion-sampling approach, «cheating» by teaching to the test is not a bad thing. If a person «cheats» by learning and practicing to solve complex, holistic, real-world problems she has demonstrated the knowledge and skills we seek as educators to develop in students. That is she has learned to think critically, reason analytically, solve problems and communicate clearly. Note the contrast with traditional learning assessments where practicing isolated skills and learning strategies for scoring high on these tests may lead to improvement on the test but such improvement is unlikely to generalize to a broad, complex domain.

CLA performance tasks

CLA performance tasks follow from the definition of the construct of competence. The domain from which tasks might be sampled focuses on everyday situations such as reading a newspaper, or engaging in civil discourse, or performing at work or in school, and the like (Table 2).

Table 2. Characteristics of CLA performance tasks

Task format	Response format
Real-world problem	Make recommendation or decision, reach a conclusion or solve a problem
Holistic, complex problem	Minimally structured to support line of argument
Provides information that may	Written and not selected
– be relevant or irrelevant to problem	Requires evidence
– be reliable or unreliable	Requires evaluation of possible alternatives
– lead to know judgmental errors (e.g. correlation is not causality, representativeness)	

Information to be used in working through the task is found in a document library. Documents in the library, regardless of the topic of the task, must be comprehensible to any college graduate. That is, humanities students should be able to read, comprehend and use information from a science oriented task that any college graduate would be expected to understand. Moreover, these documents contain information that may or may not be:

– reliable, that is trustworthy
– valid, that is relevant to the particular task at hand
– susceptible to errors, that is errors in judgment when cognitive shortcuts that reduce mental strain are used («judgmental heuristics»; e.g. correlation is not causality)

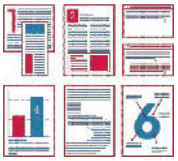
DynaTech is an example of a performance task found on the CLA (Figure 2). The student is told that DynaTech is a company that makes instruments for aircraft. The company's president is about to approve the acquisition of a SwiftAir 235 for the sales force when the aircraft was involved in an accident. As the president's assistant you [the student] have been asked to evaluate the contention that the SwiftAir is accident prone.

Students are provided an «in-basket» of information that might be useful in advising the president (Figure 2). They must weigh the evidence – some relevant, some irrelevant; some reliable, some not, some susceptible to judgmental errors, some not – and use this evidence to support a recommendation to the president. DynaTech exemplifies the kind of performance tasks found on the CLA and their complex, real-world nature.

Figure 2. CLA's DynaTech performance task

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

1: Newspaper articles about the accident
2: Federal Accident Report on in-flight breakups in single engine planes
3: Pat's e-mail to you & Sally's e-mail to Pat
4: Charts on SwiftAir's performance characteristics
5: Amateur Pilot article comparing SwiftAir 235 to similar planes
6: Pictures and description of SwiftAir Models 180 and 235



Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane.

To get a better understanding of what might be contained in a performance task's document library, consider the CLA's «crime» performance task. You are now a consultant to the incumbent mayor who is up for re-election. The main election issue is the rising number of crimes in the city and their association with drug trafficking. The mayor has proposed increasing the number of police to address crime. His opponent is a city council member who has proposed an alternative to police – increased drug education. The proposal, the council person argues, addresses the cause and is based on research studies. You are given an in-basket of information regarding crime rates, drug usage, the relationship between the number of police and the number of robberies, research studies and newspaper articles – some relevant, some not, some reliable, some not, some sensitive to judgmental error, some not (Figure 3). Your task is to advise the mayor, based on the evidence, as to whether his opponent is right about both drug education and the causal interpretation of the positive relationship between number of police and number of crimes.

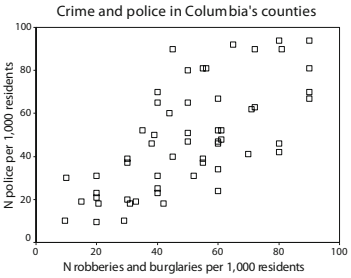
Figure 3. CLA in-basket items from the «crime» performance task

Jefferson Daily Press

Smart-Shop Robbery Suspect Caught

Drug-Related Crime on the Rise in Jefferson

By PIERRA SEROC



Crime rate and drug use in Jefferson: by zip code

Zip code	Percent of population using drugs	Number of crimes in 1999
11510	1	10
11511	3	20
11512	5	90
11520	8	50
11522	10	55

CLA scoring

Students' performance is scored with a rubric (see Table 3). More specifically, for all performance tasks, students' performance is scored on four dimensions, each with a 6-point scale: (1) analytic reasoning and evaluation, (2) problem solving, (3) writing effectiveness, and (4) writing mechanics. While this is a generic rubric, it is instantiated in each and every performance task. That is, the rubric specifies for DynaTech, for example, which information is reliable and which is not; which information is valid and which is not; and which judgmental heuristics are embedded in the task. In this way, raters can score students' performance with full knowledge where errors in judgment and argumentation might be made. While human scorers evaluate a sample of students' responses, this is done to create benchmarked scores (exemplars of scores of 0, 1 … 6). Benchmarked papers are then given as input to natural language processing software and computers are «taught» to score performance. Computers can score as reliably as or more so than human raters (e.g. Klein, 2007).

Table 3. CLA scoring rubric

Analytic reasoning – identifies strenghts and weaknesses of alternative arguments – accurately judges quality of evidence avoiding unreliable, invalid or erroneous information
Problem solving – provides decision and solid rationale based on credible evidence – acknowledges uncertainty and need for further information
Writing effectiveness – organizes «advice» in logically cohesive and easy-to-follow way – provides valid and comprehensive details supporting each argument and information source on which based
Writing mechanics – writes well constructed complex sentences – shows outstanding control of grammar conventions – demonstrates adept use of vocabulary

CLA technology

Many of the ideas underlying the CLA are not new. If you look at the history of learning assessment (e.g. Shavelson & Huang, 2003; Shavelson 2007a, b) as far back as the late 1930s, assessments like the CLA were being built in the United States. At the end of the 1970s ETS was experimenting with constructed-response tasks, ACT had created the College Outcomes Measurement Project (COMP), and New Jersey had created tasks in critical thinking to assess undergraduates' learning. These assessments had marvelous performance tasks but in all cases the attempts to bring these assessments to large-scale testing failed. They were costly, logistically challenging, and time consuming to score.

What makes the CLA new is that it solves past problems of time, cost and scoring by capitalizing on internet, computer, and statistical-sampling technologies (Table

4). Only with the advent of these technologies is it not practicable to follow in the tradition of the criterion-sampling approach. Complex performance tasks and critical writing tasks can be presented, and students' performance on the tasks can be scored by natural language processing software without compromising reliability or validity (e.g. Klein et al., 2005, 2007, 2008). Moreover, the CLA uses matrix sampling so that not all students take all questions, to reduce testing time. (Nevertheless, even with this technology, it takes a fair amount of time – 90 minutes – to answer subsets of questions. Finally, reports can be produced rather quickly because of the technology used.

Table 4. CLA technology and reporting

Characteristic	Attributes
Open-ended tasks	– tap critical thinking, analytic reasoning, problem solving and written communication – realistic work samples – engaging task as suggested by alluring titles such as «brain boost», «catfish» «lakes to rivers» – applicable to different academic majors
Computer technology	– interactive internet platform – paperless administration – natural language processing software for scoring students' written communication – online rater scoring and calibration of performance tasks – report institution's (and subdivision's) performance (and individual student performance confidentially to student)
Focus	– institution and school/department/program within institutions – not on individual student performance (although their performance is reported to them confidentially)
Sampling	– samples students so that not all students perform all tasks – samples tasks for random subsets of students – creates scores at institution or subdivision/program level as desired (depending on sample sizes)
Reporting	– controls for students' ability so that «similarly situated» benchmark campuses can be compared – provides value added estimates – from freshman to senior year or with measures on a sample of freshmen and seniors – provides percentiles – provides benchmark institutions

4. CLA approach to assessing business-planning competence

If we follow the CLA approach to construct assessment of business-planning competence, we need to begin to define what is meant by competence in this domain. I will leave the details of such a definition to my colleagues in the field. Here I sketch one possible working definition for the sake of proceeding. In this way the universe of tasks that might be sampled to form a business-planning competence assessment can be roughly delimited.

Competence in business planning

Business planning is a multifaceted complex of abilities and skills that is closely related to, that is enacted in performance in real-life business-planning situations. The closely interrelated facets of business planning include the coordination of the following macro-level task areas: (1) product or service ideas and planning; (2) market and competition research; (3) sales, marketing, distribution and public relations planning; (4) business model and organization specification; (5) time line for launching enterprise; (6) risk assessment; and (7) financial planning.

A mapping sentence for a business planning assessment might go like this:

> A measure of business-planning competence should tap *complex* business-planning abilities and skills to produce observable *performance* on a common *standardized* set of business-planning tasks that simulate with high *fidelity* the performances that are expected to be enacted when engaged in business-planning in «real world» («criterion») situations to which inferences of business-planning competence are to be drawn, with scores reflecting the *level* of performance (mastery or continuous) on business-planning tasks where *improvement* in planning can be made through *dispositions* for learning, self-regulation and deliberative practice.

Business-planning performance task

To be concrete, I draw upon an example of a business-planning task presented by Mosler, Hofknecht and Holten (April 15, 2011).[3] The task involves deciding which of several textile vendors to use for a fashion-industry enterprise. The task involves several facets of business planning: product and service ideas, model of business system, risk assessment and financial planning. In reality, an assessment of business-planning competence would be comprised of many such tasks (and others; see below). I have taken the liberty to embellish bits and pieces of the Mosler-Hofknecht-Holten performance task.

For the textile-selection task, instructions would set the context and indicate the nature of the response expected of the student. The student might receive the following instructions:

> You are planning a start-up enterprise in the fashion industry. You have to decide on a supplier for textiles. Your criteria are low price and high reliability. You have narrowed the choice of suppliers down to the two that seem most competitive. Given the following information, choose one supplier and justify your choice. In justifying your choice, be sure to provide the evidence and rationale that led you to the choice. Moreover provide evidence and rationale for not choosing the alternative supplier.

3 Mosler, Hofknecht and Holten are masters students who participated in a seminar on business-planning competence assessment run by Professor Susanne Weber at the Institute for Human Resource Education and Management, Munich School of Management, Ludwig-Maximillians-Universität.

The student is then given a library of documents to review. These documents form the basis for choosing and justifying the choice. The library might contain, for example, the following documents:

> *Document 1:* Offer from supplier 1 – 100% organic flannel cotton (white) – first choice fabric – width 160cm – price Euro 4.90 per meter, VAT excluded – incentive to purchase immediately
> *Document 2:* Offer from supplier 2 – 100% organic flannel cotton fabrics – top quality – price Euro 8.90/m, VAT included – transport costs Euro 7.50 – incentive for initial order
> *Document 3:* E-mail from Uncle John recommending supplier 1 as most reliable
> *Document 4:* Delivery history going back three years on both suppliers
> *Document 5:* Financial plan elements
> *Document 6:* A clipping from a fashion magazine displaying a dress made with supplier 2's material

Some of the information in the library is reliable, some not. For example, the student learns that Uncle John used to be a major figure in the textile industry. However upon retirement Uncle John has become somewhat of a gadfly and while his family respects him, experts in the industry question whether his judgment has held up from the old days. Moreover, the student is provided statistical data on the past three years of the performance of both suppliers with respect to the question of reliability of product shipments. It shows supplier 2 to be, on average, considerably more reliable than supplier 1. Nevertheless there was a period of three months where supplier 1 was more reliable than supplier 2. In a footnote supplier 2's inconsistency is explained as follows: a storm damaged supplier 2's property and impacted the supply of electricity. Even in that case, supplier 2 was only a few days tardier than supplier 1.

Some of the information in the document library is relevant, some not. The offers from suppliers 1 and 2 and the data on supplier reliability is relevant information. In contrast, however compellingly attractive the dress in the fashion magazine is, this information does not bear on the decision.

And some information might incorporate well-known misconceptions or errors in business planning or in judgment in general. The financial planning information for the enterprise does not include salary for the principal person leading the start up – a common mistake in financial planning. Moreover, information from Uncle John is vivid and compelling («vividness heuristic»), the statistics on delivery are detailed and boring. However, an opinion of one informant does not outweigh statistical information based on large samples over time.

Other tasks in a business-planning assessment

The textile task would be one of a sample of perhaps 20 tasks in a measure of business-planning competence. In order to get reliable estimates of individual students'

competence, about 10–20 tasks would be needed (e.g. Shavelson, Baxter & Gao, 1993). Consequently, at least one or two more performance tasks would be needed. These tasks would be about the same size as the textile task. I estimate that it would take about 30 minutes for students to complete the textile task without rushing. But, of course, this is a hypothetical task at this point; it might take longer.

To increase the sample of tasks or items, different types of knowledge and skills would also be tapped in a business-planning assessment. Following Li, Ruiz-Primo and Shavelson (2006) the following might be measured:

- Concepts and facts (declarative knowledge and reasoning: «knowing that»).
- Procedures and skills (procedural knowledge and reasoning: «knowing how»)
- Schematic or «mental» or causal models of business as a tightly interconnected system (schematic knowledge and reasoning: «knowing why»)
- Adaptability to new or «tricky» situations (strategic knowledge and reasoning: knowing what and when knowledge and skills apply and regulating their application)

Multiple-choice, short answer and spreadsheet items could be used, in addition to performance tasks, to examine students' understanding of the materials in the document library, procedural skills with finances and risk assessment, and the like. For example, multiple-choice and short-answer items might assess declarative knowledge and procedural (perhaps with spreadsheet) skills. Open-ended items might ask for an evaluation of the utility of a document in the library with the student providing a rationale and justification for the utility claim. Or such items might ask how variation in sales and marketing information might be affected by choice of supplier (schematic-modeling knowledge).

5. Concluding comments

The model of competence presented here is but one possible model that might be used to build assessments. The model leads to a different approach to assessment construction, building on McClelland's (1973) criterion-sampling approach to competence measurement. The tasks sampled on the assessment instrument would be high-fidelity simulations of real-world business-planning tasks. The tasks would evoke observable performance; the level of competence could be set (see Shavelson, in press, for cautions with performance-level setting) in order to reach a decision about whether an individual is competent in business planning.

The tasks on the assessment would make good teaching activities as well. Indeed, teaching to the test would be looked upon positively as an outcome of using such a measure of competence. By teaching the abilities and skills needed to perform well on the test the student would be learning how to perform on actual tasks that he or she might encounter «on the job» of business planning.

Finally, I suspect that dispositions so important to competent performance would be evoked as the student worked through the tasks on the assessment. Self-regulation, professional identity, affect and volition would all come into play while performing the complex, ever-changing real-world like tasks. If dispositions do come into play as students work through performance tasks[4] the need for self-report of dispositions with all of its limitations might be avoided. Or self-reports of dispositons might be used to augment, if predicatively valid in a potentially high-stakes testing situation, the competence-assessment information.

This paper has presented but one possible conceptual model and hypothetical business-planning task. The intent was to show how the criterion-sampling approach to competence assessment, embodied in the CLA approach, might be applied to the assessment of business-planning competence. This said, the devil is in the details. The next step would be to formally build the business-planning competence assessment and examine its reliability, validity and utility. Evidence from the CLA (Shavelson, 2010a; Klein et al., 2005, 2007, 2008) suggests that such an assessment could be built with positive measurement qualities.

References

Case, R. (1996). Changing views of knowledge and their impact on educational research and practice. In: D. R. Olson & N. Torrance (Eds.): Handbook of education and human development: New models of learning, teaching, and schooling. Oxford: Blackwell

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Cambridge MA: MIT Press

Hartig, J.; Klieme, E. & Leutner, D. (Eds.) (2008). Assessment of Competencies in Educational Contexts: State of the Art and Future Prospects. Göttingen: Hogrefe & Huber

Klein, S. (2007). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In: D. Nolan & T. Speed (Eds.): Probability and Statistics: Essays in Honor of David A. Freedman. IMS Collections, Vol. 2. Beachwood, OH: Institute for Mathematical Statistics

Klein, S.; Benjamin, R.; Shavelson, R. et al. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review,* 31(5), 415–439

Klein, S.; Freedman, D.; Shavelson, R. et al. (2008). Assessing school effectiveness. *Evaluation Review,* 32, 511–525

Klein, S. P.; Kuh, G. D.; Chun, M. et al. (2005). An approach to measuring cognitive outcomes across higher-education institutions. *Journal of Higher Education,* 46(3), 251–276

Li, M.; Ruiz-Primo, M. A. & Shavelson, R. J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In: S. Howie & T. Plomp (Eds.): Contexts of learning mathematics and science: Lessons learned from TIMSS. London: Routledge

McClelland, D. C. (1973). Testing for competence rather than testing for «intelligence». *American Psychologist,* 28(1), 1–14

Shavelson, R. J. (2007a). Assessing student learning responsibly: From history to an audacious proposal. *Change* (January/February), 26–33

Shavelson, R. J. (2007b). A Brief History of Student Learning: How We Got Where We Are and a Proposal for Where to Go Next. Washington, DC: Association of American Colleges and Universities

Shavelson, R. J. (2008a). Reflections on quantitative reasoning: An assessment perspective. In: B. L. Madison & L. A. Steen (Eds.): Calculation vs. context: Quantitative literacy and its implications for teacher education. Washington, DC: Mathematical Association of America.

4 Empirical evidence through think-aloud and interview methods could test this assumption.

Shavelson, R. J. (2008b) The Collegiate Learning Assessment. Forum for the Future of Higher Education: Ford Policy Forum. Cambridge, MA

Shavelson, R. J. (2010a). Measuring college learning responsibly: Accountability in a new era. Stanford, CA: Stanford University Press

Shavelson, R. J. (2010b). On the measurement of competency. *Empirical Research in Vocational Education and Training,* 2(1), 43–65

Shavelson, R. J. (in press). An approach to testing and modeling competence. In: O. Troitschanskaia & S. Bloemeke (Eds.): Modeling and Measurement of Competencies in Higher Education. Rotterdam: Sense

Shavelson, R. J.; Baxter, G. P. & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement,* 30(3), 215–232

Shavelson, R. J. & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change,* 35(1), 10–19

Shavelson, R. J.; Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement,* 36 (1), 61–71

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

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

Webb, N. M.; Shavelson, R. J. & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. Handbook of Statistics, 26, 81–124

Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In: D. S. Rychen & L. H. Salganik (Eds.): Defining and Selecting Key Competencies. Göttingen: Hogrefe & Huber

Wigdor, A. K. & Green, B. F. Jr. (Eds.) (1991). Performance assessment for the workplace (Vol. I). Washington, DC: National Academy Press