BEST PRACTICES IN SURVEY DEVELOPMENT FOR THE EVALUATION OF STUDENT GAINS FROM ENTREPRENEURSHIP PROGRAMS AND CLASSES

Jessica Menold, Kathryn Jablokow, PhD, Sarah Zappe, PhD,
Phil Reeves, Liz Kisenwether
PENN STATE UNIVERSITY

Daniel Ferguson, PhD and Senay Purzer, PhD
PURDUE UNIVERSITY

Abstract

According to the Princeton Review, 2,000 entrepreneurship courses are currently offered at a wide variety of universities across the country (Juergen 2011). With so much effort focused on entrepreneurship education, evaluating the impact that these courses have on a student's entrepreneurial mindset or the abilities required to become a successful entrepreneur has garnered much interest in recent years (Kuratko 2005; Oosterbeek, van Praag, and Ijsselstein 2010). A common method used to measure the effect that a project, course, or program has on an individual's entrepreneurial mindset is the use of a class or program-wide survey (Kuratko 2005). Although surveys to measure gains in entrepreneurial mindset are in widespread use, they often lack a theoretical framework to guide their creation. This can lead to data and analysis that reflect the poor psychometric properties of the survey rather than the actual impact of entrepreneurship programs or courses on students (Kline 1986). This paper examines the current state of survey techniques within entrepreneurship education, as well as proposing a process for more robust assessment creation.

Introduction

As the number of entrepreneurship education programs continues to grow, the need to assess these programs' effectiveness grows as well. One of the most efficient methods for course and program assessment is the use of pre- and post-course/program surveys of students in order to evaluate their growth with respect to relevant constructs. The results of such surveys are used as formative evaluations of projects, courses, and programs; however, without full knowledge of the validity evidence of these surveys, it is inappropriate to use their empirical results as indicative of the success or failure of the relevant course or program.

Because gathering validity evidence can be a difficult endeavor, this paper advocates creating content valid surveys based on Messick's (1995) unified theory of validity, as well as the works of Downing and Haladyna (1997), in particular, their paper on test item development. We begin by proposing a survey development process based on the works of the previously mentioned psychometricians and propose a process to be used in the development of entrepreneurship survey instruments. We evaluate the current state of the assessment of entrepreneurship education by examining a selection of existing entrepreneurship education assessment instruments through the lens of our proposed survey development process.

Using the works of Messick and of Downing and Haladyna, we derived a set of nine metrics and used them to evaluate twelve entrepreneurship assessment instruments. These twelve instruments were selected based on the relevance of their underlying constructs to entrepreneurship and/or their frequent citation in the entrepreneurship literature. After reviewing each of the instruments, a short discussion of the general strengths and weaknesses found in entrepreneurship education assessment is presented, followed by recommendations for future work.

Theoretical Framework for Evaluation

In highlighting the importance of an instrument's validity in the context of score interpretation, Messick (1995) notes: "The construct validity of score interpretation comes to undergird *all* score-based inferences." Score interpretation – i.e., the interpretation of the results of an assessment instrument – can play a pivotal role in the assessment of entrepreneurship programs, which is why the validity of the instruments used to assess them is so important. In short, score interpretation is dependent upon the validity evidence collected for the instrument itself, making the rigor of the development process for instruments of critical importance.

Content relevance and representativeness are the first steps towards developing a sound instrument. Content relevance and representativeness refer to the range and limits of content coverage – i.e., the boundaries of the construct domain to be assessed. Test items are the building blocks of any assessment instrument, and by nature, they specify the content domain of the instrument. In other words, sound instruments are composed of sound items that generate support for the instrument in the collected body of validity evidence. Sound items are grounded in a theoretical framework and are representative of and relevant to the content domain of interest. In the following sections, we review Messick's unified theory of validity as it pertains specifically to content relevance and representativeness, along with a theoretical framework for instrument evaluation derived from the work of Downing and Haladyna (1997).

Messick's Unified Theory of Validity

Messick (1995) defined validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment." According to Messick's theory, validity can be differentiated into six aspects: content, substantive, structural, generalizability, external, and consequential. Definitions of these six aspects of content validity can be found in pages 16-17 (Messick 1995). To illustrate the structure of Messick's theoretical model, Purzer and Cardella (n.d.) transformed Messick's unified theory into a process diagram for instrument creation, as shown in Figure 1. The diagram in Figure 1 outlines a path that instrument developers should follow as they collect validity evidence while creating an instrument. For the purposes of this paper, we focus specifically on the first two aspects of Messick's model – i.e., the content aspect and the substantive aspect – which focus on content relevance and representativeness, or the characteristics of the content domain that is being assessed.

According to Messick, "The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality." In other words, the content aspect of construct validity serves to specify the boundaries of the construct domain, or the determination of the skills, traits, knowledge, and attitudes that are related to the relevant construct. The content aspect requires that the tasks or behaviors to be assessed are both relevant to the construct domain and representative of the domain. Typically, content relevance and representativeness are assessed by expert professional judgment.

With regard to the substantive aspect of construct validity, Messick notes: "The substantive aspect refers to theoretical and content of the items. This is followed in the process by a pilot test with a sample representative of the final population(s) for whom the instrument is intended. Content creation during the blueprint development stage can be a difficult task, and boiling that content down into a list of cohesive items can be challenging. Downing and Haladyna proposed an ideal process for test item development, as discussed in the following section.

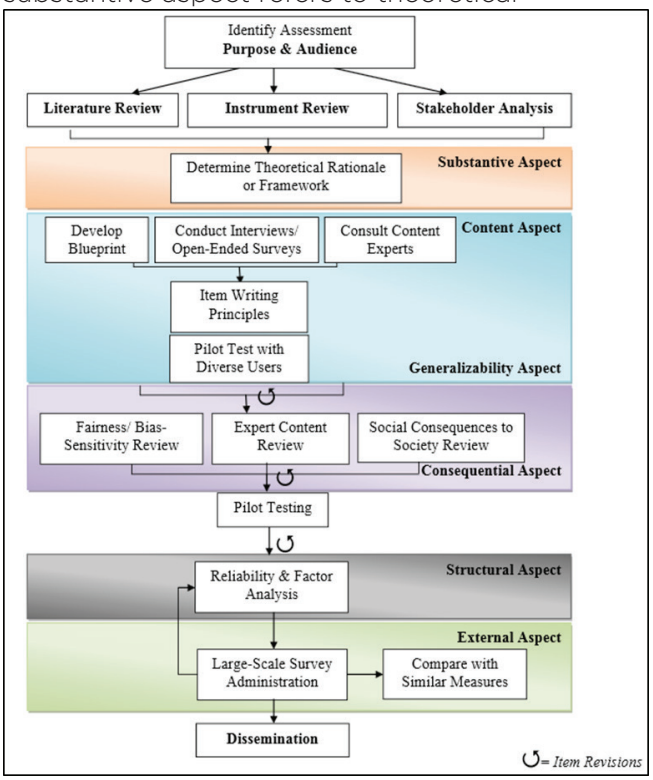

Figure 1. Purzer and Cardella's diagram for development of valid measurement instruments

Referring to Figure 1, the steps that relate to the content aspect refer exclusively to item development. Development of an instrument blueprint, open ended surveys, and consulting content experts are three different exercises that can be used to create and evaluate items for a survey or other type of instrument. These tasks lead to a review of basic item writing principles, followed by a pilot test with a representative sample.

Development of an instrument blueprint refers to the identification of behaviors, tasks, skills, and knowledge related to the construct to be tested; the open ended surveys and a consultation with content experts serves to confirm the blueprint. Once the behaviors, tasks, skills, and knowledge of the related construct are identified, item writing principles (Downing and Haladyna 1997; Kline 1986; Messick 1995) are used to guide the format

Downing and Haladyna's Ideal Process for Item Creation

Downing and Haladyna's (1997) ideal process for item creation uses Messick's (1995) unified theory of validity as a foundation and expands the substantive and content aspects. Downing and Haladyna argue that both qualitative and quantitative evidence are needed to support the use or deletion of test items in an instrument. Quantitative evidence gathering is well established and can be found in such methods as factor analysis and classical item analysis. Downing and Haladyna outline a precise method to gather qualitative evidence and list eleven types of qualitative evidence, accompanied by the specific activity needed to collect the evidence, as shown in Figure 2.

Of the eleven types of qualitative evidence presented in Figure 2, we selected four as relevant and implementable in entrepreneurship education assessment. These four are content definition, test specifications, item content verification, and item tryout and pretesting, which can be used to guide the selection, creation, and/or evaluation of an entrepreneurship survey or instrument.

Content Definition

Content definition refers to the selection of the survey/instrument domain and associates the construct that is to be measured with the test specifications and items. This first form of evidence clearly defines the boundaries of the assessment. Job-task

analysis is often suggested as key in defining the content of a survey or instrument and simply involves evaluating the construct in its "natural environment." For example, if we are interested in evaluating risk-taking in the context of entrepreneurship, we would observe entrepreneurs in their work environments and evaluate how/if/when they take risks, what that risk-taking looks like, and how it is different from risk-taking in the general population. Content may also be defined from extant literature or existing theories that have been accepted by some panel of experts in the field. Based on the works of Messick, and Downing and Haladyna, we show below a condensed list of three metrics that will aid in the creation and evaluation of entrepreneurship instruments through the lens of content definition:

1. Clearly defined boundary of the construct(s) of the instrument or survey

2. Definition of the construct(s) based on literature, job task analysis, and/or expert review

3. Documentation of methods used to select the content presented (Crant 1996)

Test Specifications

Test specifications link the content domain with the test items and help to define the portions of the content domain that will be evaluated or sampled. Referring back to Messick's unified theory, the test specifications lay out a guide to ensure that items are both relevant and representative of the content domain being assessed – i.e., that the items are related to the knowledge, skills, attitudes, and behaviors to be assessed. The test specifications define the type of content within the assessment, as well as specifying the size of each content category to be present in the survey instrument. For example, a list of test specifications for entrepreneurship may include: risk-taking, business savvy, and creativity, with a breakdown of 20%, 50%, and 30%, depending on the underlying theoretical framework. Based on the works of Messick, and Downing and Haladyna, we show below a condensed list of three metrics that will aid in the creation and evaluation of entrepreneurship instruments through the lens of test specifications:

Model of Item Validity Evidence: Qualitative Evidence

Type of Evidence	*Activity*	*Evidence Needed*
Content definition	Role delineation, job-task analysis; practice analysis completed	Documentation of the method(s) used to select item content
Test specifications	Table of specifications or test blueprint created	Documentation of systematic link of test content to test specifications or test blueprint
Item writer training	Develop training materials and methods; train item writers	Documentation of methods, principles, written materials, and sample items
Adherence to item-writing principles	Standard item-writing rules adopted	Evidence of compliance with rules and documentation of process used to review items
Cognitive behavior	Cognitive classification system used to classify items	Documentation of system used and its rationale; reports of any research using system
Item content verification	Content experts review and judge items	Content experts' credentials; records of content-expert review process
Item editing	Review items and professionally edit	Credentials and experience of editors; editorial and style guidelines, documentation of edit and review cycle
Bias–sensitivity review	Bias–sensitivity review policies and procedures developed	Documentation of bias–sensitivity review; rationale for policies; credentials of reviewers
Item tryout and pretesting	Pretest, pilot test, or field test items; item performance data; examinee interviews	Documentation of examinee pilot test data; examinee and item characteristics
Key validation and verification	Correctness of keyed answer verified by panel of content experts	Policy and procedures for key verification; documentation of key validation results
Test security plan	A test security policy and set of procedures are developed	Copy of policy and procedures manual that specifies how items are protected from security lapses

Figure 2. Downing and Haladyna's Eleven Types of Qualitative Validity Evidence

1. Detailed rationale or process for breakdown of proportions of content domain to be represented by test items

2. Content areas and more specific content within these areas clearly stated

3. Agreement on proportional breakdown by panel of researchers or content experts

Item Verification

When creating items for an instrument, the content is typically based on expert experience, textbooks, or a thorough literature review. However, in order to verify that the content reflected in the items is representative and relevant to the domain of assessment, it is commonplace to organize a panel of experts to evaluate the item set. Typically, this panel of experts is composed of professionals with

significant knowledge and/or experience in the domain being assessed. The panel is also briefed on what is to be assessed – i.e., the panel is explicitly told the definition of the construct(s) being assessed and is instructed to evaluate the items with reference to this definition. This helps to avoid dissension amongst the experts on such things as content definition or theoretical framework, and helps maintains focus on the relevance and representativeness of the items. Based on the works of Messick, and Downing and Haladyna, we show below a condensed list of two metrics that will aid in the creation and evaluation of entrepreneurship instruments through the lens of item verification:

1. Initial items based on literature review, previous instruments/survey, field experience

2. Items verified through panel of content experts on both relevance and representativeness to the content domain relative to the pre-defined construct definition

Item Tryout

Item "tryout" or pretesting is a method used to evaluate the cognitive processes that items evoke from test takers. For example, one approach to pretesting items is the "think-aloud" method, in which test takers are asked to say out loud their thoughts and feelings while completing the survey or assessment. This aids the reviewer in gauging how well an item matches the cognitive task being assessed, as well as flagging any confusing or misleading items on which test takers may stumble. Based on the works of Messick, and Downing and Haladyna, we show below one metric that will aid in the creation and evaluation of entrepreneurship instruments through the lens of item tryout:

1. Pretesting of items using think aloud or other protocols to ensure the test takers engage in the expected cognitive processes

The four forms of qualitative evidence discussed above can be used to aid in the selection and/or creation of valid entrepreneurship education assessment instruments. In the following section, we use the metrics we defined for these four forms of evidence to evaluate the current state of the art in entrepreneurship assessments through a selection of entrepreneurship surveys and instruments.

Evaluation of Entrepreneurship Instruments

This brief review examines twelve instruments used for the evaluation of entrepreneurial self-efficacy, entrepreneurial orientation, and/or entrepreneurial behaviors or traits of students. Most often, these student evaluations are used as measures of performance for entrepreneurship programs at a variety of universities by surveying students across a semester, year, or program. To be included in this review, the instruments or surveys had to explicitly measure constructs related to entrepreneurship or be frequently used in entrepreneurship evaluations. Below is a list of the twelve instruments included in this review:

1. Entrepreneurial self-efficacy (Taatila and Down 2012)

2. NCIIA Entrepreneurship Inventory (Shartrand et al. 2008)

3. Proactive Personality Scale (Bateman and Crant 1993)

4. Entrepreneurial self-efficacy scale (Chen, Greene, and Crick 1998)

5. General enterprising tendency (Caird 1991)

6. Individual entrepreneurial orientation (Bolton and Lane 2012)

7. Student entrepreneurial orientation (Taatila and Down 2012)

8. Entrepreneurial behavior inventory (Lau et al. 2012)

9. Tolerance for Ambiguity Scale (Herman et al. 2010)

10. Engineering Entrepreneurship Survey (Duval-Couetil, Reed-Rhoads, and Haghighi 2011)

11. Entrepreneurial mindset index (Shartrand et al. 2008)

Instrument Evaluation

The instruments listed above (numbered 1 to 12) and the metrics used for evaluation (as previously outlined in the previous section) were used to construct an evaluation matrix in Table 1. Each metric was treated as a dichotomous variable, then each instrument was assigned an "X" when a metric was met or a dash when that metric was not met (to the best of our knowledge). The articles used to review each instrument are listed in the references (Bateman and Crant 1993; Chen, Greene, and Crick 1998; Duval-Couetil, Reed-Rhoads, and Haghighi 2011; Herman et al. 2010; Kussmaul et al. 2006; Purzer and Cardella n.d.; Taatila and Down 2012).

Prior to reviewing our findings across all twelve instruments, we will examine one instrument in depth in order to demonstrate how the metrics were applied. We chose the NCIIA Entrepreneurship Inventory (Shartrand et al. 2008) for this more detailed review, as it performed best across all metrics. The first group of metrics (content definition) are evaluations of the depth to which each instrument defined and set boundaries to the content domain being evaluated. The theoretical framework for the NCIIA Entrepreneurship Inventory was based on previous work, specifically [], and this previous research helped shape the boundaries of the domain. The development of the instrument began with an evaluation of stakeholder needs, including which content was of interest and the best ways to evaluate this content. This adds further justification to the definition of the content domain, as well as the boundary of assessment. All of the methods used to define content were well documented in Kussmaul et al. (2006) and Shartrand et al. (2008). For these reasons, the NCIIA Entrepreneurship Inventory satisfied all three metrics within the content definition group.

The NCIIA Entrepreneurship Inventory also excelled in the next two grouping of metrics, i.e., test specifications/blueprint creation and item verification, respectively. Within Shartrand, there is a clear breakdown of content areas, as well as a list of specific factors in each area. Also, "the items were initially based on a pre-existing taxonomy," and the "list of items [was circulated] to ten additional leaders of entrepreneurship education," highlighting both item verification and agreement on proportional breakdown. The only metric we could not confirm for the NCIIA Entrepreneurship Inventory was the item-tryout; to the best of our knowledge, this inventory was not tested with a smaller pilot sample utilizing a think-aloud or similar protocol.

In reviewing the instruments listed in Table 1 overall, we see that some metrics were strong across all twelve instruments, while other metrics were rarely satisfied. In general, the selected twelve instruments appear to be strongest in content definition and weakest in test specifications or blueprint creation, respectively. In terms of content definition, each of the instruments based their construct definition on previous literature, job task analysis, or expert opinion. For example, the entrepreneurial self-efficacy scale (Taatila and Down 2012), NCIIA Entrepreneurship Inventory (Shartrand et al. 2008), Engineering Entrepreneurship Survey (Duval-Couetil, Reed-Rhoads, and Haghighi 2011), and student entrepreneurial orientation scale (Taatila and Down 2012) all derived their constructs and domain boundaries from thorough literature reviews and expert opinion.

The detailed breakdown of test proportions metric was only met by the NCIIA Entrepreneurship Inventory (to the best of our knowledge), which provided a detailed breakdown of the content areas covered by its items. The NCIIA Entrepreneurship Inventory was also the only instrument to document that its items were reviewed and agreed on by a panel of content experts.

Another often-missed metric in our evaluation of the twelve instruments was the pretest or tryout method used to ensure correct cognitive processes, which refers to the use of small pilot tests and think-aloud protocols. These pretests act as a last measure to ensure that the items within the instrument are evoking the proper responses from respondents. Think-aloud protocols can also help instrument developers identify problem words or phrases that may be unfamiliar or awkward for the respondent. These practices can flag items that may be especially prone to social biasing (Duval-Couetil, Reed-Rhoads, and Haghighi 2010).

A final key observation is that the practice of cutting and pasting certain items from a variety of instruments into one single instrument was common across multiple instruments we evaluated. For example, the individual entrepreneurial orientation scale and the student entrepreneurial orientation scale both selected items from other instruments and reworded them slightly to make them more appropriate for their audience. Unfortunately, while this practice is convenient, instrument developers using it run the risk of altering the construct being measured. In effect, by changing the items and only utilizing portions of an instrument, survey developers can no longer be certain that the construct measured by and validated for the original instrument is still being measured by the new instrument (Duval-Couetil, Reed-Rhoads, and Haghighi 2010; Duval-Couetil, Reed-Rhoads, and Haghighi 2011; Herman et al. 2010).

Summary and Recommendations for Future Work

In summary, the instruments currently being used in entrepreneurship education are strong in certain areas of validity evidence, such as content definition; however, many appear to be weak in other areas, such as test specifications, item verification, and item tryout. We also noticed an alarming trend toward cutting and pasting items from a variety of instruments in order to create a single instrument. This practice is not advised and goes against many principles of sound instrument development (Downing and Haladyna 1997; Duval-Couetil, Reed-Rhoads, and Haghighi 2010).

Instrument development can sometimes take years to complete, and many psychometricians argue that collecting validity evidence is a lifelong endeavor that is never really finished (Kline 1986; Messick 1995). Further, following every rule and procedure outlined by Messick and Downing and Haladyna (1997) is unwieldy and may not be practical in all situations. However, the metrics proposed in Table 1 provide an efficient protocol for evaluating surveys or instruments to use in entrepreneurship research, including course and program assessments. Survey creation can be aided by the use of these metrics as a stage gate model for survey and instrument development.

Key areas for improvement in entrepreneurship program and course assessment instruments include the adoption and use of test specifications, and the use of content experts to verify item representativeness and relevance. The adoption of these practices and methods will support entrepreneurship education as a whole by providing more reliable and accurate instruments and thus give a more accurate view of the state of entrepreneurship education.

	INSTRUMENTS/METRICS	A	B	C	D	E	F	G	H	I	J	K
CONTENT DEFINITION	Clearly defined boundary of construct	-	Yes	Yes	Yes	-	Yes	Yes	Yes	-	Yes	Yes
	Definition of construct based on literature, job task analysis, or expert review	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Documentation of methods used to select content presented (Crant 1996)	Yes	Yes	-	Yes	-	-	Yes	Yes	-	Yes	Yes
TEST SPECIFICATIONS/ BLUEPRINT	Detailed breakdown of test proportions	-	Yes	-	-	-	-	-	-	-	-	-
	Content areas and more specific content within areas clearly stated	-	Yes	-	Yes	-	Yes	-	Yes	-	Yes	-
	Agreement on proportional breakdown by panel	-	Yes	-	-	-	-	-	-	-	-	-
ITEM VERIFICATION	Initial items based on literature review, previous instruments/survey, field experience	Yes	Yes	-	Yes	Yes	Yes	-	Yes	Yes	Yes	Yes
	Items verified through panel of content experts	Yes	Yes	-	-	-	-	-	Yes	-	Yes	-
ITEM TRYOUT	Pretest or tryout method utilized to ensure correct cognitive processes	-	-	-	Yes	-	-	-	Yes	-	Yes	Yes

Table 1. Evaluation of Instruments Using Metrics for Qualitative Validity Evidence

References

Bateman, Thomas S., and J. Michael Crant. 1993. "The Proactive Component of Organizational Behavior: A Measure and Correlates." *Journal of Organizational Behavior* 14(2): 103-118.

Bolton, Dawn Langkamp, and Michelle D. Lane. 2012. "Individual Entrepreneurial Orientation: Development of a Measurement Instrument." Education and Training 54: 219-233.

Caird, Sally. 1991. "The Enterprising Tendency of Occupational Groups." International Small Business Journal 9(4): 75-81.

Chen, Chao C., Patricia Gene Greene, and Ann Crick. 1998. "Does Entrepreneurial Self-efficacy Distinguish Entrepreneurs From Managers?" Journal of Business Venturing 13(4): 295-316.

Crant, J. Michael. 1996. "The Proactive Personality Scale as a Predictor of Entrepreneurial Intentions." Management 29(3): 62-74.

Downing, Steven M., and Thomas M. Haladyna. 1997. "Test Item Development: Validity Evidence from Quality Assurance Procedures." Applied Measurement in Education 10(1): 61-82.

Duval-Couetil, Nathalie, Teri Reed-Rhoads, and Shiva Haghighi. 2010. "Development of an Assessment Instrument to Examine Outcomes of Entrepreneurship Education on Engineering Students." 2010 Frontiers in Education Conference (FIE). IEEE.

---. 2011. "The Engineering Entrepreneurship Survey: An Assessment Instrument to Examine Engineering Student Involvement in Entrepreneurship Education." *The Journal of Engineering Entrepreneurship* 2(2): 35-56.

Herman, Jeffrey L., et al. 2010. "The Tolerance for Ambiguity Scale: Towards a More Refined Measure for International Management Research." *International Journal of Intercultural Relations* 34(1): 58-65.

Juergen, Michelle. 2011 "The Top 50 Entrepreneurship Programs." *Entrepreneur.* 1 Oct. 2011. Accessed 10 Sept. 2014. http://www.entrepreneur.com/article/220327.

Kline, Paul. 1986. *A Handbook of Test Construction: Introduction to Psychometric Design*. London: Methuen.

Kuratko, Donald F. 2005. "The Emergence of Entrepreneurship Education: Development, Trends, and Challenges." *Entrepreneurship Theory and Practice* 29(5): 577-598.

Kussmaul, Clifton, et al. 2006. "Institutionalizing Entrepreneurship at Primarily Undergraduate Institutions." Presented at the 10th Annual Conference of the National Collegiate Inventors and Innovators Alliance, Portland, Oregon.

Lau, Theresa L. M., et al. 2012. "The Entrepreneurial Behaviour Inventory: A Simulated Incident Method to Assess Corporate Entrepreneurship." *International Journal of Entrepreneurial Behaviour & Research* 18(6): 673-696.

Messick, Samuel. 1995. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50(9): 741.

Oosterbeek, Hessel, Mirjam van Praag, and Auke Ijsselstein. 2010. "The Impact of Entrepreneurship Education on Entrepreneurship Skills and Motivation." *European Economic Review* 54(3): 442-454.

Purzer, Senay, and Monica Cardella. n.d. *Instrument Development Model: A Map based on Messick's Unified Theory of Validity*. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Shartrand, Angela, et al. 2008. "Assessing Student Learning in Technology Entrepreneurship." 38th Annual Frontiers in Education Conference. IEEE.

Taatila, Vesa, and Samuel Down. 2012. "Measuring Entrepreneurial Orientation of University Students." *Education+ Training* 54(8/9): 744-760.

Wilson, Fiona, Jill Kickul, and Deborah Marlino. 2007. "Gender, Entrepreneurial Self-Efficacy, and Entrepreneurial Career Intentions: Implications for Entrepreneurship Education." Entrepreneurship Theory and Practice 31(3): 387-406.