RESEARCH & PRACTICE IN ASSESSMENT

VOLUME FOUR | WINTER 2009
www.RPAjournal.com
ISSN#2161-4210

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

Editors

Table of Contents

Comments from the Editor

Keston H. Fulcher
Director of Assessment, Evaluation, and Accreditation
James Madison University

The three articles in this installment of Research and Practice in Assessment address specific issues pertaining to assessment practice. Bresciani, Griffiths, and Rust suggest a theoretical framework of the stages faculty experience as they consider the role of assessment in their work. This piece has implications regarding how assessment consultants approach faculty.

The second piece by Flateby shares how the University of South Florida developed its renowned system for assessing writing and critical thinking. As with most successful programs, it evolved over years with careful thought and a mix of trial and error. Included are links to rubrics used by USF.

Before using students' test scores to make inferences or decisions about particular programs, faculty and administrators should be confident that these scores are valid for their intended purpose. To this end, Barry and Finney investigate how conditions during test piloting can affect psychometric properties. Indeed, they find that a test's factor structure varies depending upon the stakes of the test.

Assessment at North Carolina State University: Adapting to Change in the Workplace

Marilee J. Bresciani
San Diego State University

Jane H. Griffiths and Jon P. Rust
North Carolina State University

Abstract

Effectively introducing change in job responsibilities, particularly when dealing with tenured faculty, can be challenging. More often, additions or changes to work tasks, such as integrating assessment procedures into existing work tasks, requires employees to apply new and/or more complex knowledge, skill, and ability. When compared to organizations practicing contemporary-type work methods, institutions practicing traditional-type work methods, such as those common to traditional university settings, can find adaptation to change particularly onerous. For example, tenured faculty may perceive introductions of new concepts or new terminology as substantive changes in their practice, even though the change is an introduction of new labels to their current practice or a systematization of a former practice. Consequently, the integration of new assessment procedures, as in this instance, can have a significant impact on faculty when learning to accommodate that change. Therefore, understanding why long-tenured employees may be particularly resistant to change in the workplace is important when adding assessment procedures to existing work responsibilities. To better understand faculty resistance to change and to help facilitate the change process, one can apply the integration of work adaptation theory. This paper reviews concepts included in the theory of work adaptation, with a focus on work adaptation theory developed by Petrini and Hultman. Petrini and Hultman cite six common beliefs that lie at the root of employee resistance to change and provide strategies for addressing such resistance. The six common beliefs include the following: (a) One's needs are currently met by the traditional methods already in place, (b) The change will make it more difficult to meet one's needs, (c) The risks involved outweigh the possible benefits, (d) There is no basis for the change – it's just another plan to get more work out of us with fewer resources, (e) The organization is mishandling the change, and (f) The change will fail and go away. This paper addresses issues related to employee resistance when incorporating undergraduate assessment into the culture of a Research Extensive institution. Discussed are experiences in confronting Petrini and Hultman's six beliefs when working with tenured employees as well as the application of strategies they suggest when addressing employee resistance to change. Furthermore, the six beliefs and strategies are applied as a means to clarify key findings with regard to the institution's successful implementation of changes designed to improve student learning.

Background

The study institution is a state supported, research extensive, urban, and land-grant institution with an emphasis on science, engineering and technology. More than 29,000 students attend this institution, of which more than three quarters are undergraduates and almost nine of every ten are native state residents. Undergraduate assessment at this institution was initially a response to accreditation requirements and concerns for accountability from the state legislature. In its inception, assessment included a strong commitment to evidence-based decision-making with the intent to continuously improve programs. Assessment-based program review began at this institution in the early 1990's. At that time, program review was a process of reporting on the current state of a program, a "snap shot" of where the program was at some point in time.

In 1997, Vice Provosts endeavored to recreate the cumbersome program review process making it more meaningful and incorporating student learning outcomes assessment as the vehicle to creating an environment of continuous improvement associated with program review. Thus, an ad-hoc committee of faculty from across the campus was organized to establish guidelines for program review with the following set of requirements: (a) to focus the process on continuous improvement, (b) to make the process sensitive to outside accreditation, and (c) to respect program autonomy. Three years later, with guidelines set, a second faculty led ad-hoc committee, the Committee on Undergraduate Program Review (CUPR), was formed and given the charge of implementing the process.

In the spirit of maintaining program autonomy, CUPR respected and supported the notion that the faculty of a program should determine what the educational objectives and student learning outcomes should be for their program. Further, CUPR has worked to ensure that the faculty of a program should be the ones to decide which assessment methods are best able to measure the extent to which the graduates of a program meet the stated outcomes.

To begin this implementation process, CUPR first made sure every college was represented by interested, respected, and dedicated faculty. Next, they adopted a shared conceptualization (determined by the CUPR members) and a common vocabulary or set of definitions for key words associated with assessment. Finally, CUPR set out to transform the institution by changing the way faculty approached the process of evaluating undergraduate student learning and to imbed that process into the day-to-day activities of the institution. Introducing such a significant change requires that faculty work through a period of adjusting to new responsibilities and procedures. Applying the theory of work adaptation assisted the institution and its faculty with the change adjustment process.

Work Adaptation Theory

The introduction of innovative change to daily work tasks, such as adding embedded assessment procedures into faculty members' day-to-day academic responsibilities, can have a significant impact on the individuals learning to accommodate that change. Therefore, when a job changes, employees are required to adapt. The theory of work adaptation can be used to illustrate this process. Work adaptation is an outgrowth of over forty years of research by Dawis, Lofquist, and scholars in their development of a theory to explain how individuals adjust to changes in the workplace (Dawis & Lofquist, 1984; Sharf, 1997). The basic premise suggests that to maintain job satisfaction, individuals continually strive for a complementary relationship with their jobs. The level of job satisfaction is dependent upon the extent that the individual's needs, values, and interests are met in the workplace. Similarly, the employee must satisfy the knowledge, skill, and ability requirements of his or her job. As a result, a balance between individual needs and job requirements must be found to attain and maintain job satisfaction. However, when changes are introduced into the workplace, that balance or equilibrium is disturbed (Sharf, 1997; Yeatts, Folts, & Knapp, 2000). Typically, this disturbance will provoke feelings of anxiety and resistance within the employee, which in turn can lead to a reduction in job satisfaction. The employee will then begin the arduous task of reestablishing equilibrium in an effort to regain job satisfaction.

Frequently, changes in the workplace demand more complex knowledge, skill, and ability requirements than the traditional methods they replace (Hackman, 1990; Yeatts et al., 2000). As a result, organizations practicing traditional-type work methods, such as those common to a traditional university setting, will find adaptation to change especially challenging. Typically, in traditional work settings, work tasks are separated out; each employee performs a different task or focuses in on a specialty area, often developing expertise in that task or area over time. The nature of this environment invokes a propensity for individuals to protect intellectual and practical knowledge. Particularly for university faculty, the role of expert is highly regarded. Therefore, a global understanding of the organization and its work processes are limited because each employee closely guards his or her knowledge and skills (Yeatts et al., 2000). Further, this approach supports an employee's notion that if he or she is the only worker with the knowledge to perform a certain task then he or she will become indispensable to an employer, thus ensuring job security and/or status. Alternatively, organizations practicing contemporary-type work methods assemble teams that require every member to perform each individual task of an entire work process (Yeatts et al. 2000). This approach ensures the transfer of knowledge and skills among the participating employees and facilitates understanding and efficiency of the entire work process.

When addressing adaptation to change in a traditional university setting, one needs to consider the characteristics of those individuals in the presenting work environment. More often, university faculty and staff are long-tenured and this alone can make it more challenging for individuals to adjust to new practices and relinquish previously successful methods (Fossum, Arvey, Paradise, & Robbins, 1986). In a review of the literature, Yeatts et al., (2000) cite common difficulties in long-tenured employees adjustment to change in the workplace. This includes a propensity for long-tenured employees to lag behind in knowing how to apply new tools and techniques as well as an inability to see how their work performance can be improved through the implementation of new knowledge. Generally, long-tenured employees have more invested in traditional methods and may also doubt their ability to effectively learn new procedures. In addition, an employee that believes he or she has already attained long-term job security and, thus, has demonstrated his or her ability to perform current work requirements, may be resistant to take on more demanding work tasks. Moreover, employees holding positions of seniority may view sharing hard earned knowledge and experience (often a condition of contemporary job redesign) as a threat to status, privileges, or control of their work environment. Given these characteristics, altering the knowledge, skill, and ability requirements for long-tenured employees can significantly disrupt the balance between job requirements and individual needs. Naturally, this disequilibrium provokes high levels of anxiety and resistance within the employee resulting in almost inevitable job dissatisfaction.

Resistance to learning new job processes for the long-tenured employee can be likened to the middle aged individual who has never needed to learn to swim. Hass and Keeley (1998) provide a colorful metaphor to illustrate this: if an individual has managed quite well in life without knowing how to swim he or she may not be too enthusiastic about taking up swimming lessons. This individual must first be persuaded that there are very good reasons for learning how to swim. In addition, the individual needs to believe that swimming is a skill he or she is able to develop (and without drowning in the process). From this person's perspective, learning how to swim may have some possible advantages but it is not seen as a life necessity. Rather, it is viewed as a threatening (the thought of drowning) and physically demanding task.

Understanding why long-tenured employees may be particularly vulnerable to change in the workplace is important when asking faculty to assist with institutional change such as assessment. The process of integrating assessment procedures into existing work tasks provides an opportunity to address more common causes of resistance in reaction to change as well as to implement appropriate methods for reducing resistance. Petrini and Hultman (1995) cite six common beliefs that lie at the root of employee resistance to change and also suggest methods for overcoming resistance:

First Belief: One's needs are currently met by the traditional methods already in place. In higher education, many faculty believe that the processes they employ in their day-to-day work tasks already function quite well. Accordingly, the introduction of a perceived additional task, such as conducting assessment of student learning, is considered an "add-on" to traditional responsibilities. As a result, faculty may have a propensity to believe there is no legitimate need for the change.

Resolution: Clearly explain why the change is essential and explain why and how the change will help faculty better meet their needs. In addition, Blank (1990) suggests helping employees examine exactly what they do in the workplace and why they do it. This presents the opportunity for individuals to identify loopholes or inefficiencies in their work tasks and opens up the possibility for them to see how they might do things differently.

Second Belief: The change will make it more difficult to meet one's needs. In an environment where faculty are primarily rewarded for their research and grant-writing, adding on an expectation to evaluate student learning is perceived to detract from the ability to meet the institutional expectations that exist for research. If expectations for research are not met, then faculty members may not receive the money they need to operate at their desired level. Thus, their needs are not met.

Resolution: Help diminish this threat to job satisfaction by evaluating whether their facts are complete and accurate. Determine whether their assumptions are founded on accurate information. In other words, is it true that engaging in assessment of student learning will detract from their research efforts or will the residual effects of change inevitably enhance research efforts?

When correcting inaccurate perceptions, provide viable information to support those corrections. Ask for, or even suggest, ways you might be able to assist in helping them (faculty) adjust to the changes. Demonstrate a willingness in a collaborative effort to help them adjust to the changes while finding ways for them to meet their needs.

Third Belief: The risks involved outweigh the possible benefits. Many faculty believe that the risk of the time spent on evaluating student learning is not worth the benefit of learning how to do it.

Resolution: Establish what grounds the faculty have to support this belief. Assess whether their facts are correct and that their interpretations of those facts provide a realistic assessment of the risks. In addition, Blank (1990) suggests supporting the rationale for future benefits with theory, research, and evidence. This is something that university faculty can easily relate to and that will help build confidence in the credibility of future changes. For example, using supportive data from a comparable university that had implemented similar assessment techniques with favorable results, would offer tangible support for the proposed change.

Fourth Belief: There is no basis for the change—it's just another plan to get more work out of us with fewer resources. Understandably, many faculty are skeptical about whether assessment is "here to stay." Higher education, as well as K-12, has been inundated with "quality assurance movements" and it has caused many to question the validity of learning about another process. This is further exhausted by the fact that some faculty believe that assessment is a way to reduce resources currently assigned to educational initiatives.

Resolution: Help employees understand the necessity for change. First, listen to their concerns or problems and be careful to address each while also explaining the consequences of continuing with the current methods. Identify ways the change will improve the university, college, or department. Be specific. For example, you may explain that having assessment data, which demonstrates the efficacy of a program, can be used to laud successes as well as to identify opportunities for improvement.

Fifth Belief: The organization is mishandling the change. Many faculty believe that the implementation of the assessment process or the way in which assessment is being conducted is not efficient. This judgment could be based on sound criteria or could be simply based on the appearance that doing assessment is taking too much time away from other projects that are valued more by the faculty or by the institution.

Resolution: Ask employees to identify their concerns then listen carefully. If mistakes have been made, apologize. Do not use excuses, rather accept accountability and provide the necessary information to explain what is being done to correct past mistakes and prevent future ones. Ask for their ideas in preventing future mishaps, but be honest about which suggestions are viable and provide a rationale. Give your employees a straight answer. Building employee support requires standing behind your promises and following through with your agreements. Failing to do this will heighten employee resistance.

Sixth Belief: The change will fail and go away. This is similar to the example given for the Fourth Belief in that many faculty have seen improvement initiatives in higher education prior to student learning outcomes assessment. At the start-up of each initiative, institutional support has been evident. Yet, as the initiative continues, institutional support lags and therefore faculty remain cautious about investing their time in anything "new" and different.

Resolution: Be firm in your conviction that the change is here to stay and state the reasons for this, however, explain that the process of that change is open to discussion and collaboration. Again, listen to concerns, determine if they are basing their beliefs on accurate and complete information, correct any inaccuracies, and provide information to support your corrections. Accept accountability for your mistakes and involve employees in brainstorming ideas for making a successful change and how you might help them better implement that change. Follow through with the final decisions for making improvements. Successful implementation of collaborative solutions will help build credibility and reduce employee resistance.

As with any major project, successful implementation is highly dependent on thorough preparation, smart planning, and logical execution (Blank, 1990). Before presenting a proposal for change, Blank recommends that you have all the facts clear, accurate, and complete. That means have every angle covered and anticipate possible challenges or doubts posed by your employees. Identify possible problems as well as the far-reaching effects of the proposed changes then determine how they will be managed. Therefore, be prepared to discuss those details when questioned.

It is well and good to endorse effectual communication as one of the keys to implementing successful change but exactly how does one communicate with a resistant employee? Fortunately, Petrini and Hultman (1995) provide guidelines for communication with the resistant employee. The key lies in understanding the individual's state of mind. Obviously one cannot know for certain what a person is thinking but one can observe an individual's behavior and ask: What fact, belief, feeling, or value is being conveyed by what this individual is saying or doing? However, Petrini and Hultman state that the most effective method for determining the source of an individual's resistance is to ask specific yet non-threatening questions. The questions they suggest using are listed in four categories: (a) verify the facts, (b) challenge their beliefs, (c) acknowledge their feelings, and (d) relate the change to their values.

When dealing with employee resistance, the keys to successful change comprise several principal factors. Kirkpatrick (1993) provides a good summary of suggestions that fall in line with the contents this article. First, he recommends that one understand those individuals that will have to adjust to the change. Second, he emphasizes the importance of clear communication: provide all the facts—what, why, who, and when—and answer all employee questions thoroughly. Third, he advocates employee involvement. Ask employees to assist, to be part of the solutions, and to help identify resources that may help them or their colleagues. In addition, we should include accountability and follow through with agreements as important factors in building credibility and trust with one's employees. It is our intention that the methods derived from the work of Blank (1990) and Petrini and Hultman (1995), described above, can be used as practical applications for helping one understand, involve, and communicate with resistant employees.

Finally, one element that must be included as part of implementing successful change is to provide effective learning opportunities for employees. Effective and accessible training and education can help restore the balance between individual needs and job requirements. Increasing appropriate skills and knowledge helps employees meet the demands of a redesigned job and, thus, can help restore job satisfaction (Yeatts et al., 2000). Typically, anxiety levels heighten when employees are required to learn something new. However, employee anxieties can be tempered when (a) training is offered well in advance of the scheduled changes, (b) individuals can learn at their own pace, (c) supplementary learning opportunities are provided for those who want them, and (d) a safe learning environment is ensured. A safe learning environment should be supportive, encouraging (i.e., providing feedback, rewards, and praise to reinforce learning), and non-judgmental (Yeatts et al., 2000). With a safe learning environment in place, employees will not be so afraid to make mistakes or to ask questions but rather feel free to explore new approaches and thus be empowered to learn for themselves. In the case of assessing undergraduate education, that equates to an empowered learner taking ownership of developing the assessment process and becoming a key driver in implementing those new processes in their specific program or department.

CUPR's Implementation and Evaluation of the Transformation Process

Using the aforementioned advice of Petrini and Hultman (1995) as well as (Yeatts et al., 2000), and Kirkpatrick (1993), CUPR began the process of implementing undergraduate student learning outcomes assessment into traditional academic program review. The implementation process for transformation to assessment – based program review encompassed many of the criteria addressed in the theory of work adaptation. Furthermore, in order to understand the institution's evaluation of its ability to address the transformation of the institution through assessment in accordance with work adaptation theory, the institution conducted a survey (Bresciani, 2004). The following implementation steps and subsequent survey findings are organized by the six common work adaptation beliefs identified by Petrini and Hultman (1995).

First Belief: One's needs are currently met by the traditional methods already in place. Resolution in implementation: As faculty involved were primarily motivated by the improvement of student learning, the evaluation process was re-designed to emphasize the gathering of information in order to improve student learning. Specific on-campus examples were used to illustrate how programs on the whole could benefit from the assessment of student learning. In addition, faculty were alerted to the fact that outside accreditation required a focus on student learning (SACS, 2000), and thus inclusion of student learning assessment into the revised program review process was inevitable.

To facilitate communication of the refined program review process, CUPR held informational presentations to the Provost, Deans, Department Heads, the Faculty Senate, the Council on Undergraduate Education (which establishes and maintains the general education requirements), and groups of faculty from individual colleges and departments. Notices went to the university community addressing the timetable for reviews as well as pre – review "assignments" designed to encourage programs to get started. Many people in the Division of Undergraduate Affairs (UGA) and CUPR generated an on-line "toolkit." This "toolkit," which included many online resources for those getting started in assessment, was publicized to the university community as an available resource.

Survey findings: In the survey that was administered to all full-time faculty in the spring of 2004 (Bresciani, 2004), faculty reported not clearly understanding why the change in the process had been made. Faculty believed that the formalized reporting of student learning (something the majority of faculty reported already doing) was not understood clearly. Where understanding was reported, it was due to the linkage to regional accreditation.

Clearly more communication is needed in order to promote the value of formalized reporting of student learning. Thus, varying frameworks for the dissemination of information and feedback have been organized and implemented. Follow-up surveys will be conducted to see if these changes in communication processes will prove effective.

S*econd Belief: The change will make it more difficult to meet one's needs.* Resolution in implementation: CUPR has worked continuously to make the process as manageable as possible. For instance, CUPR members have demonstrated with examples that grading coursework can easily and effectively be combined with course and even program outcomes assessment. Furthermore, CUPR provided additional on-line examples of ways in which assessment has helped programs improve.

The most challenging conversation has been in regards to the concern that research and grant writing will suffer as a result of engaging in outcomes assessment of student learning. CUPR has held many conversations around this topic and there has been no consensus as to whether assessment should become a part of the promotion and tenure consideration. And if consensus was reached, would the evaluation of student learning be as valued as other types of research? Conversations continue.

Survey findings: Faculty reported that the assessment of student learning takes a great deal of time, particularly the documentation of the assessment results. While they find benefits in the evaluation of student learning and can provide examples of how the process has helped improve student learning, faculty remain concerned that student learning assessment will go un-rewarded and unrecognized by senior administration (Bresciani, 2004).

In order to address this concern, more conversations need to be held at the senior administrative and faculty leadership levels, especially if the needs of the faculty are defined by the expectations of the administration through the rewards of promotion and tenure. If the needs of the faculty are identified by the faculty as being those along the lines of making the process simpler, then further information is needed to get the faculty's perspectives on how the processes can further be streamlined.

Third Belief: The risks involved outweigh the possible benefits. Resolution in implementation: CUPR's use of teams of faculty in writing the "rules of engagement" for the assessment process has been key in making the process guidelines less threatening to faculty. Furthermore, it has been faculty feedback, which has advised the revisement of the guidelines. Indeed, using examples of how assessment has helped programs improve is also a key factor in defusing fear of potential risks since the examples themselves are of success and programmatic in nature, and thus not personally threatening. Finally however, the question of whether the cost is worth the benefit is elusive at this point as start-up costs for any venture are often higher than the revenue generated. Conducting cost-analysis studies too early in the implementation process can lead to misinterpretations of both costs and benefit.

Survey findings: The majority of faculty surveyed identified value in the assessment process as it relates to the improvement of student learning (Bresciani, 2004). Many faculty remain concerned that the formalization of the process has taken too much of their time and that that time may be better spent on their research and grant activities. These concerns may be due to the actual amount of time being spent on the assessment of student learning or they may be due to the amount of time devoted to learning the assessment process.

Clearly, more information is needed to clarify the core of the concern. In addition, the benefits gained from assessing student learning (e.g., improvements made to student learning) should be better represented to the campus community so that they can readily see assessment's value.

Fourth Belief: There is no basis for the change – it's just another plan to get more work out of us with fewer resources. Resolution in implementation: In addition to the items mentioned above, providing resources can facilitate the process and make it clear that university administration is solidly behind the effort. Therefore, UGA provided financial support in the form of "mini-grants" to assist programs with well-thought implementation plans. Software designed to facilitate assessment efforts (e.g., TracDat) was purchased for any program that indicated they desired it. Additionally, many workshops were conducted by UGA and CUPR to educate faculty and assist them in: writing learning outcomes (Fall 2001 – present), identifying assessment methods (Spring 2002 – present), how to use TracDat (Fall 2002 – present), and how to make assessment meaningful and manageable (Fall 2002 – present).

These workshops were set so as to help faculty meet yearly requests from CUPR and the Vice

Provost's office. The requests or "assignments" included asking faculty to develop and submit educational objectives and student learning outcomes in August 2001, develop and submit assessment plans including identifying assessment methods in August 2002, and collect data and submit a small report on assessment of at least one outcome in August 2003. This process promotes faculty involvement and encourages personal investment in the new assessment procedures.

Survey findings: While the survey indicated that faculty would prefer that their department devote specific resources to this endeavor, many colleges have done so and there are several resources that are made centrally available as well. The faculty has had a mixed review as to whether these resources were meaningful to faculty, let alone desired by them (Bresciani, 2004).

When faculty are not positing their belief that the regional re-accreditor is the primary motivator for this fourth belief, they are either lamenting its creation or singing it praises in how well it has encouraged them to be more reflective in their practice. The largest concern appears to be one of a reallocation of time. In other words, the needed resource is time or a reallocation of existing duties so that meaningful reflection of what is being done is in fact occurring.

Fifth Belief: The organization is mishandling the change. Resolution in implementation: In this instance, mistakes were made and accountability was accepted. It is possible that some mistakes could have been avoided. In other cases, perceived mistakes were not mistakes per se, but opportunities to learn how to make the process more efficient and effective. Faculty are regularly asked to provide insight into how we might improve the process. As we move the process forward, we find additional opportunities for improvement and work with faculty to generate solutions that are faculty friendly, yet programmatically accountable and effective. The committee in control of the process continues to grow in size and through each membership growth spurt; new ideas emerge in how to make the process better.

Survey findings: As previously mentioned, the faculty's greatest concerns, as expressed in the survey, revolved around finding efficiencies in the formalized process (Bresciani, 2004). While this was a consistent concern, no specific means to make the assessment of student learning process were identified. Further exploration of explicit means of refining the assessment process must be sought, as not doing assessment is just not an option.

Sixth Belief: The change will fail and go away. Resolution in implementation: CUPR reiterated to faculty that outside accrediting agencies are articulating their expectation for assessment of student learning. Further, the assessment movement has been gaining momentum in all areas of institutional performance for over twenty years leading to various state governments, such as Virginia, Florida, and Texas to apply oversight on learning outcomes to public institutions. With this kind of committed governmental structure, one would assume that assessment will not disappear soon. Add these to a firm reminder from the Provost that assessment is here to stay and will be used in program planning and the conclusion is inevitable.

Now, the challenge is in communicating the value of continuing the process. Some faculty have begun to raise the question of whether the improvements in student learning gained through assessment would have been made without assessment. Having no data on this simply means that this argument can only be theoretical in nature, yet it remains a key belief. However, keeping with this belief, the point is that the time that is asked of faculty to engage in the assessment of student learning is an expressed value and concern of faculty and therefore should be addressed as such.

Survey findings: By many metrics, the response to assessment of student learning among the faculty has been strong, yet there is still a ways to go. While a high percentage of programs have been submitting required assessment documents, more work still needs to be done to improve the quality of assessment. To further facilitate communication and involvement, CUPR has reviewed each assessment plan turned in and has responded to every program as they made submissions. CUPR also has solicited feedback from the programs on how to improve the process of implementing assessment of undergraduate education. At the same time, NCSU has developed a large group of involved faculty, and excellent resources for training as well as tools.

Many programs have developed their assessment efforts to an advanced state and CUPR accepted and analyzed the first assessment-based program review document in the 2002 – 2003 academic year. With all these successes, the question still looms: How do you know your institution has developed undergraduate assessment to the point where it is self-perpetuating? The answer will likely be related to the degree to which administrators and faculty have considered and addressed key characteristics associated with institutional transformations.

Conclusion

Petrini and Hultman's (1995) six common beliefs provide a framework in which to organize and implement a meta-analysis of your assessment process. Doing so may provide the administrators and faculties with solutions to challenges that may not have been so obvious before. In addition, it helps one analyze the extent to which the assessment process has been of value to improving student learning. One case study was presented here. The authors encourage readers to attempt to adopt this model at their institutions as they move toward a culture of accountability.

References

Blank, R. E. (1990). Gaining acceptance: The effective presentation of new ideas. *Total Quality Management, 1*, 69-73.

Dawis, R. V., & Lofquist, L. (1984). *A psychological theory of work adjustment.* New York, NY: Appleton-Century-Crofts.

Eckel, P., Green, M., & Hill, B. (2001). *On Change V—Riding the Waves of Change: Insights from Transform ing Institutions.* Washington D.C.: American Council of Education.

Ewell, P. T. (1997). Identifying indicators of curricular quality. In G. J. Gaff, L. J. Ratcliff and Associates, (eds) *Handbook of the undergraduate curriculum: A comprehensive guide to purposes, structures, practices, and change.* San Francisco, CA: Jossey-Bass.

Ewell, P. T. (2003). *Specific Roles of Assessment within this Larger Vision.* Presentation given at the Assessment Institute at IUPUI. Indiana University-Purdue University- Indianapolis.

Fossum, J., Arvey, R., Paradise, C. C., & Robbins, N. E. (1986). Modeling the skills obsolescence process: A psychological/economic integration. *Academy of Management Review, 11*, 362-374.

Haas, P. F., & Keeley, S. M. (1998). Coping with faculty resistance to teaching critical thinking. *College Teaching*, *46*, 63-67.

Hackman, J. R.(1990). *Groups that work (and those that don't).* San Francisco, CA: Jossey—Bass.

Kirkpatrick, D. L. (1993). Riding the winds of change. *Training & Development, 47,* 28-32.

Lopez, C. (2002). Levels of Implementation. Higher Learning Commission of the North Central Association: Chicago, IL. Retrieved December 18, 2004 from http://www.ncahigherlearningcommission.org/resources/assessment/index.html

Maki, P from Anderson, J.A., Maki, P., & Bresciani, M.J. (2002a, July). *Expanding Faculty Involvement in Assessment-Based Undergraduate Academic Program Review: A Case Study.* Paper presented at the meeting of the American Association of Higher Education Assessment Conference. Boston, MA.

Maki, P from Maki, P & Bresciani, M.J. (2002b, July). *Integrating Student Outcomes Assessment into a University's Culture.* Paper presented at the meeting of the American Association of Higher Education Faculty Forum on Roles and Rewards Conference. Phoenix, AZ.

Palomba, C.A. & Banta, T.W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education.* San Francisco, CA: Jossey-Bass.

Petrini, C., & Hultman, K. E. (1995). Scaling the wall of resistance. *Training & Development,* 49, 15-18.

Southern Association of Colleges and Schools. (2000). Guidelines for accreditation. Retrieved December 18, 2004 from http://www.sacscoc.org/

Sharf, R. S. (1997). *Applying career development theory to counseling.* Pacific Grove, CA: Brooks/Cole Publishing Company.

Yeatts, D. E., Folts, W. E., & Knapp, J. (2000). Older workers' adaptation to a changing workplace: Employment issues for the 21st Century. *Educational Gerontology, 26,* 565-582.

Developments and Changes Resulting from Writing and Thinking Assessment

Teresa Flateby
University of South Florida

Abstract

This article chronicles the evolution of a large research extensive institution's General Education writing assessment efforts from an initial summative focus to a formative, improvement focus. The methods of assessment, which changed as the assessment purpose evolved, are described. As more data were collected, the measurement tool was transformed into a system of assessment. Additionally, challenges encountered are discussed.

Introduction

Ten years ago the General Education assessment team at the University of South Florida (USF) used a holistic scale to evaluate student writing in the General Education curriculum. Student writing samples were collected at three points in the curriculum: when students (a) entered as first-year students, (b) completed their first year and, finally, (c) completed all general education courses. Raters who also scored the State's "rising junior" essay tests assigned scores of one through six connoting proficiency levels from "below" to "exceeds" expectations. While the results confirmed anecdotal evidence that some students were more than acceptable writers, they also indicated that many students were not proficient. Although we used this approach for several years (and collected summative data), we lacked formative data to identify specific student writing strengths and weaknesses that could inform instruction or the curriculum. Data confirmed writing deficiencies, but were not valuable for making changes and improvements, one essential purpose of assessment. As a result, the assessment team suggested evaluating the usefulness of an analytic rubric designed at USF for the classroom to address program assessment purposes.

The classroom rubric, developed before the formal assessment of General Education occurred, was initiated to address needs identified in a two-year, team-taught writing-intensive learning community program at the University of South Florida. One of the goals of this program was to encourage the deeper learning often associated with writing. Two discoveries led to the development of the rubric. The program coordinator, who is also a faculty member of the English department, and I (the external evaluator of the program) determined through interviews and surveys that grading of students' writing assignments varied widely among faculty. Also early and throughout the two-year program, we observed complex thinking through classroom observations, reflecting the upper levels of Bloom's Taxonomy of Educational Objectives—Cognitive Domain (1956). Responding to these two findings, we recognized the need for a tool that enables the consistent evaluation of students' writing and thinking skills by faculty from diverse disciplines. We reviewed existing performance-based measures, but did not ind any that fulfilled the identified needs. Thus, we began the development of the Cognitive Level and Quality of Writing Assessment (CLAQWA) rubric.

Based upon commonly used writing handbooks, such as *St. Martin's Handbook, Harbrace College Handbook*, and *Scott Foresman Handbook for Writers,* the initial writing rubric included a five point scale with only levels one, three, and five defined. The sixteen trait analytic rubric was organized into categories, which were modified after meeting with teams of faculty and applying the rubric to papers. Due to a writing style often observed in beginning students' essays, the single category "Organization and Development" was divided into two: one pertaining to structure and another reflecting reasoning and evidence supplied. We realized that while many beginning students' essays had an appealing structure (five paragraph essays that students learn to produce for standardized testing), the quality of content and quality of reasoning exhibited were often weak. These and other results were used to refine the rubric to represent the full range of writing – qualities associated not just with learning to write, but also writing to learn.

When searching for a framework for the thinking portion of the resulting two-part scale, we chose Bloom's Taxonomy of Educational Objectives-Cognitive Domain (1956). In addition to its accessibility, the taxonomy reflects the type of thinking faculty typically advocate, such as analysis, synthesis, and evaluation. Moreover, several authors have recommended this taxonomy to assess writing. In 1983 Spear advocated the use of Bloom and his colleagues' work for writing evaluation, and Olson (1992) developed a writing curriculum around Bloom's cognitive levels. In 1997 Steele, in his rationale for the development of American College Testing's Critical Thinking Assessment Battery, (which required writing) maintained that "Bloom's Taxonomy remains useful as a means of analyzing and classifying the levels of intellectual demands in cognitive activities" (p. 19).

The work of Madaus and his colleagues (1973) provided the basis for USF's cognitive scale. Their work showed a branching at the higher end of the taxonomy, thus transforming it into a four-level taxonomy (instead of the original six-levels). We subdivided these four taxonomy levels into low, medium, and high categories. Unlike the writing scale, we chose not to define the categories within levels, because when using the cognitive scale to assess levels reached in student texts, we found little variation in instructors' judgments.

When first applying the rubric for program assessment purposes, we used the initial iteration of the scale (five levels, with levels one, three, and five defined). It soon became evident, however, that all five levels needed clear definitions to achieve acceptable inter-rater reliability. Indeed, if raters within the institution cannot agree on ratings of essays then it is impossible to make defensible statements about students' performance levels or to make comparisons over time, across years, or within groups. Thus, we began the laborious task of clearly describing all five levels of the sixteen element analytic scale.

This continuing phase of development underscores the evolutionary nature of rubric development and use. As data were gathered, variations and perceptions of the definitions surfaced. Because rubrics are based upon language, users' experience and biases, these factors impacted the interpretation of levels of the traits. As calculated by the percent of adjacent-rater agreement, acceptable inter-rater reliability values (.89-.93) were achieved following clarification of the rubric (Micceri, unpublished institutional document, http://usf.edu/assessment).

As we proceeded with the assessment of writing and thinking, we continued to collect data at the same points in the curriculum: the beginning of Composition 1, the completion of Composition 2, and in liberal arts "exit" classes that are completed in the junior and senior years. With this data collection plan we were attempting to ascertain if students were reaching expected writing levels and on which of the components of the writing rubric needed the most improvement. In collecting data, we randomly selected sections from Composition 1 and 2 classes and used essays from all students in those sections.. The data collection for exit classes was less structured; faculty volunteered to provide their sections' essays. Because the interest was in students' performance after completing the General Education curriculum, and not growth in these exit classes, this type of sample selection seemed defensible. We attempted, however, to ensure that students in the sample were representative of the relevant demographics of the USF student population.

Using Results

After scoring our students' essays with the analytic rubric for approximately three years, we made valuable discoveries, which were used to suggest instructional and curricular changes. For example, when we began measuring the cognitive levels reached in our junior and senior undergraduate students' texts, we developed a standard prompt within courses and allowed students a week to complete the assignment. Although written to elicit Level Four on the Cognitive Scale, results showed that student performance was lower than desired. This finding was consistent with the "Reasoning" and "Quality of Evidence" performance levels of the writing rubric. We were uncertain however, if students' performance was truly reflective of their achievement levels or if it was adversely affected by the prompt, which was only minimally tied to course content.

Due to this concern, we changed our assessment strategy to include assignments on instructors' syllabi, if they targeted sufficiently high cognitive levels. With this approach, we hoped to determine if connecting the prompt more specifically to class assignments would elicit higher thinking skills. Although not systematically researched, we made a significant discovery: the importance of the prompt. Faculty routinely thought they were asking students to write at higher cognitive levels than their prompt reflected, and often the expectations were unclear to students. In addition, after evaluating hundreds of students' papers written to address many different prompts, scoring teams found the prompts to be critical, not only for eliciting a specific cognitive level, but also clarifying expectations for students. More open-ended or ambiguous assignments produced lower performance than assignments with clear expectations. This finding has had broad-based instructional and faculty development relevance.

Our data and process revealed that even if faculty and assessment teams do not evaluate students' cognitive levels reflected in their writing, the conscious selection of appropriate cognitive levels and careful construction of the assignments to reflect these levels are important to eliciting desired writing. Also, attention to the cognitive levels helps ensure compatible results if comparisons are to be made. Our data support composition literature suggesting that when students begin writing at higher cognitive levels, often their writing skills deteriorate (Schwalm, 1985). This finding has both pedagogical and assessment implications. If a goal is for students to clearly communicate higher order thinking, they must be given adequate opportunities in multiple classes to develop these more advanced thinking skills. Also, for assessment purposes, an institution or program must decide which cognitive levels should be addressed in assignments, especially if comparisons are made; this too underscores the importance of carefully planning the assignment's cognitive level.

Another finding was used to make curricular changes. Results confirmed that many of our students were not writing at the level expected; more importantly, we discovered that the weakest areas pertained to thinking, such as providing supporting evidence, and developing and organizing ideas. Writing skills such as grammar and mechanics, while below desired levels, were stronger than critical thinking skills.

After assessing general education learning outcomes for several years, general education became the focus of our Quality Enhancement Plan, a plan required by the Southern Association for Colleges and Schools for improving student learning outcomes. The assessment data helped guide revisions to the general education curriculum, resulting in specific changes to address weaknesses discovered in students' writing and thinking. Process writing (encouraging revisions facilitated by feedback) is now required in four of the twelve general education courses. Central to the writing emphasis is the development of ideas, inclusion of supporting evidence, logical progression of ideas and cohesiveness of texts. In addition, the plan promoted graduate and undergraduate student training to assist with writing assessment and to provide feedback to larger classes. Another change introduced is a capstone course in which writing in students' disciplines is emphasized. Equally important, the general education curriculum now emphasizes critical and higher order thinking, as well as inquiry-based learning approaches.

In addition to the direct evidence collected, we gathered indirect survey data. These results indicated that some faculty were concerned about students' writing performance levels, felt ill-equipped to provide adequate feedback, were concerned about class sizes prohibiting the ability to give feedback, and were unsure if sufficient resources were available to help students with writing deficiencies.

To address some of these concerns, we have transformed our classroom and program assessment rubric into an online system (CLAQWA Online). This online system assists faculty, students, and assessment professionals to evaluate student writing and thinking across the curriculum and helps close the assessment loop. Faculty or assessment teams are able to select writing and thinking components appropriate for a particular assignment. The instructor or the team evaluates students' writing and thinking by indicating directly on students' online texts which of the five levels described for each element reflects the text and by providing additional comments, if desired. Students are then able to access their work, which have the weak or strong writing element levels embedded in their texts. Students are able to review online instructional examples written for all levels of each trait, with feedback explaining why each example represents a specific level. This review helps them understand performance at each level and improve their writing on a trait (thus closing the loop). Designed to aggregate results, faculty and assessment teams are easily able to determine problem areas to address in their classes or in programs, again helping to improve students' writing and thinking (i.e. close the assessment loop). Through the online system students are able to give feedback to each other, thus further engaging them in the writing and improvement process (http://www.usf.edu/assessment/CLAQWA/Online).

Also through our assessment processes we discovered another method for improving student writing, which has become valued by faculty. Several members of the scoring team who were teaching composition decided to modify the paper version of the CLAQWA rubric for peer review use in the classroom. Although peer review was already part of their classes, they found that the modified rubric produced improved writing as compared to the peer review process they had been using. The success experienced with peer review in composition classes led to questioning its applicability in classes from different disciplines. We have conducted several peer review studies to determine if improvement could be measured. In electrical engineering and literature classes, improvement was observed with paper or online approaches. In the most recent studies, focusing on peer review through the online system, measurable improvements were found in varying degrees in all sixteen of the rubric's elements.

Challenges and Conclusions

Several challenges associated with the writing assessment are currently being addressed at the University. Although we made changes in the General Education curriculum in response to the assessment data, the actual instructional changes are not as widespread. Because the use of assessment results and faculty development opportunities are interdependent, identifying the person(s) or unit(s) responsible for coordinating results with development is critical. Without this coordination, the optimal use of assessment data may not be realized, which is often cited as an assessment weakness. The question of who is responsible for ensuring that data are actually used, especially for a general education curriculum, must be clearly established and faculty development opportunities must be directly tied to assessment results.

Related is the importance of administrators' support of these assessment efforts and the insurance that resources and rewards are available to faculty for making instructional and curricular changes based on assessment data. Gaining an administrative commitment may be difficult in some institutions, but is essential for promoting the message that assessment not only is essential for accreditation, but also for improving (maximizing) student learning.

Another finding relevant to other institutions' assessment processes is the importance of developing detail in rubrics. A rubric should provide clear operational descriptions associated with different levels of proficiency. For example, the criteria for paragraph construction that exceeds expectations is much clearer to faculty and students when a rubric uses language such as, "Each paragraph is unified around a topic that relates to the main idea. All paragraphs support the main idea and are ordered logically" rather than with simply "Exceeds Expectations." Furthermore, faculty tend to rate more consistently with each other when definitions are clearly articulated. Finally, we discovered that after these rubrics were fully developed that we were able to engage students in their own learning, improve students' writing and thinking, and demonstrate this improvement.

In sum, USF has learned a tremendous amount about its students' writing and has used this information to improve the quality of our instruction. To get to this point, however, required several years of careful thinking about how USF wants students to write, how to elicit this type of writing, and how to accurately assess it. That said, improving writing and its assessment at USF is still evolving. Implementation of the program could be more pervasive and support more robust. Persistence and sending clear messages to faculty and educating administrators that improving student learning is assessments' fundamental purpose may help diminish these challenges.

References

American Philosophical Association. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction, ("The Delphi Report"). (ERIC Document Reproduction No. ED 315 423).

Bloom, B. S. (ed.) (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook 1: Cognitive Domain. New York: McKay.

Madaus, G. F., Woods, N. E. & Nuttal, R. L. (1973). A causal model analysis of Blooms' Taxonomy. *American Educational Research Journal, 10*, 253-262.

Micceri, Ted (2006). Unpublished institutional document, http://usf.edu/assessment.

Olson, C. B. (1992). *Thinking/writing: Fostering critical thinking through writing*. New York, NY: Harper Collins Publishers.

Schwalm, D. E. (1985). Degree of difficulty in basic writing courses: Insights from the oral proficiency interview testing program. *College English, 47*(6), 629-640.

Spear, K. (1983). Building cognitive skills in basic writers. *Teaching English in the Two-Year College*, *9*(2), 91-98.

Steele, J.M. (1997). *Identifying the Essential Skills in Critical Thinking at the Post-secondary Level to Guide Instruction and Assessment.* [Draft]. Iowa City: American College Testing.

CLAQWA information and website may be found at http://usf.edu/assessment

Does it Matter How Data are Collected? A Comparison of Testing Conditions and the Implications for Validity

Carol L. Barry and Sara J. Finney
James Madison University

Abstract

The effects of gathering test scores under low-stakes conditions has been a prominent domain of research in the assessment and testing literature. One important area within this larger domain concerns the implications of a test being low-stakes on test evaluation and development. The current study examined one variable, the testing context, that could impact students' responses during low-stakes testing, and subsequently the decisions made when using the data for test refinement. Specifically, the factor-structure of college self-efficacy scores was examined across three low-stakes testing contexts, and results indicated differential model-data fit across conditions (the very controlled context yielded the best model-data fit), implying that testing conditions should be seriously considered when gathering low-stakes data used for instrument development.

Introduction

As the emphasis on accountability in education has increased, so has the need for a clear understanding of the validity of the inferences made from examinee scores. This need is more imperative when one considers that many times there are no consequences of poor performance or low effort for the examinee. In fact, oftentimes the measures given in order to make high-stakes decisions about program effectiveness have relatively little personal meaning or importance to the students completing them. Situations in which there are little to no consequences to the test-taker are termed "low-stakes." This paper focuses on the implications of low-stakes testing on the validity of inferences made from test scores when those scores are used for instrument development.

Low-Stakes Testing and Examinee Motivation

There is a well-documented link between low-stakes testing environments and examinee motivation. Because there are very few, if any, consequences associated with performance and because students may perceive no personal gain from the experience, low-stakes testing often leads to low effort and motivation on the part of the test-taker (Wise & DeMars, 2005). Students may feel that there is nothing in it for them and may not be motivated to perform their best. Thus, their scores may not serve as valid indicators of their true level of the construct of interest (Sundre, 1999; Sundre & Kitsantas, 2004; Wise & DeMars, 2005). Essentially, this decrease in student motivation results in an increase in construct-irrelevant variance, with further implications on the psychometric functioning of the test items.

Uses of Low-Stakes Data and Threats to Validity

One of the main uses of data gathered in low-stakes environments is in evaluating program effectiveness for accountability purposes. Assessment practitioners may gather data to gauge whether or not a certain program delivered its intended effects. However, if low motivation results in test scores that are not truly representative of the construct of interest, the scores are then ambiguous at best and misleading at worst (Wise & DeMars, 2005). Thus, much of the research focused on low-stakes testing and motivation has emphasized either filtering out examinees with low motivation (e.g., Sundre & Moore, 2002; Wise & Kong, 2005; Wise, Wise, & Bhola, 2006) or attempting to increase examinee motivation (e.g., Wise, Bhola, & Yang, 2006) as ways to handle this construct-irrelevant variance.

Although it is often noted that one must exercise caution when making decisions about *program effectiveness* based on data from tests that are low- or no-stakes to the student, few studies note that caution must also be exercised when making decisions about the *test itself.* That is, the more fundamental and pervasive use of data collected in low-stakes environments is for instrument development purposes. Often, tests created to be used in high stakes conditions are evaluated and modified using data from low-stakes conditions (e.g., pilot testing; DeMars, 2000). Specifically, assessment and measurement professionals seem comfortable collecting data in low-stakes environments (e.g., through large-scale testing programs or university participant pools) and using these data to examine the psychometric functioning of the items in order to inform instrument development decisions. If students do not provide valid responses, test developers may make unnecessary changes (or *not* make necessary changes) to an instrument. The need for sound instrument development practices is made more imperative when one realizes that sound assessment practice begins with appropriate and well-functioning measures.

Although sparse, there is some research that has studied the impact of low-stakes conditions and, consequently, low motivation on the psychometric properties of test items. Most of this research has examined the psychometric functioning of dichotomously scored achievement test items. One study approached this by examining item-by-item differences in performance by two groups of students that differed in the stakes associated with the test (Wolf, Smith, & Birnbaum, 1995). These researchers found that mentally taxing items exhibited differential item functioning; when matched on ability, the group of students for whom the test was low-stakes performed worse than those for whom the test was high-stakes. Similarly, student performance has been shown to be lower in the low-stakes condition of pilot testing than in a high-stakes testing condition, which may lead to poor instrument refinement decisions if the item difficulties estimated under pilot conditions are thought to represent item difficulties under operational conditions (DeMars, 2000). An additional study focusing on the problem of low motivation found that the inclusion of examinees who demonstrated rapid-guessing (i.e., examinees with extremely low motivation) affected the estimation of item parameters (Wise & DeMars, 2006). Specifically, items that were known to have low item difficulties appeared more difficult and more discriminating when rapid-guessers were included in the sample.

Despite the research conducted on the effects of low-stakes and low motivation on the psychometric properties of achievement tests, one area of research that is lacking involves the effects on the psychometric properties of non-cognitive or developmental tests. The items on these instruments are typically polytomous or continuous in nature, and their psychometric properties are generally studied through the use of factor analysis. Interestingly, there appears to be very little, if any, research conducted on whether and how low-stakes environments impact the factor-structure of developmental measures. This is somewhat surprising given that student attitudes/affect are often of interest to student affairs personnel and that assessment specialists are often concerned with both learning and developmental outcomes. It seems reasonable to believe that, similar to achievement tests, the psychometric properties of developmental instruments would also be impacted by the decreased student motivation that accompanies low-stakes testing.

Purpose of the Current Research

Because low-stakes testing environments are unavoidable for many who study the properties of tests, and thus there may be inconsistency in the stakes associated with data gathered for test development versus data gathered for decision making (DeMars, 2000), it is important to determine the best way to collect useful and valid data that are of no- or low-stakes to the participants. Specifically, we were interested in examining if changes in the testing context would impact student responses to low-stakes tests. Would a more controlled testing context improve the quality of low-stakes data? To answer this question, we examined the factor structure of college self-efficacy scores from multiple samples gathered in several different testing contexts. Our main focus was the effect of testing context on model-data fit. That is, did the same factor structure emerge under the different contexts, or were the relationships between items different across context, resulting in different psychometric properties associated with the measure. The theoretical underpinnings of the models tested are discussed at length in another paper (Barry & Finney, 2007) and will not be the elaborated upon here; rather, the focus of this paper will simply be on comparing model-data fit across testing contexts.

We believe this study helps answer the call of Birenbaum (2007) to evaluate the validity of the *full* testing program. Specifically, Birenbaum emphasized the need for entrenching the comprehensive assessment process within an overarching validity framework. That is, one should not focus solely on the validity of inferences made based on scores, but rather should consider these inferences within the wider frame of how the assessment instruments map to the domain of study, the psychometric functioning and internal structure of the instruments (e.g., factor structure), and the *contexts* in which the data were collected (Birenbaum, 2007). In other words, the entire assessment process needs to be evaluated with respect to validity. Again, this study focuses on the impact of context on examinee test-taking behavior when tests are of no-stakes to students.

Methods

Participants and Procedures

Five samples of data were collected across a variety of testing conditions at a mid-sized, mid-Atlantic university. In all conditions, student responses were of no stakes to the individual student, and students were not provided with any information regarding their scores. Each sample and the method by which data were gathered are described below (see also Table 1 for a description of all samples).

Table 1
Description of Samples

Sample	Acronym	Age	Testing Condition	Item Order
Uncontrolled Freshman-1	UnFr-1	Freshmen	Students completed an online version of the instrument on their own time, unsupervised, prior to arriving on campus for the start of the Fall 2006 semester.	Non-Randomized
Uncontrolled Freshman-2	UnFr-2	Freshmen	same as above	Non-Randomized
Very Controlled Upperclassman	VCUp	Sophomores, Juniors, Seniors	Students completed the instrument in a small (~20 seats) classroom in the presence of a trained proctor. Care was taken to slow response time.	Non-Randomized
Controlled Upperclassman	CUp	Sophomores and Juniors	Students completed the instrument in large (i.e., number of seats ranged from 63-250), lecture-style classrooms, in the presence of trained proctors.	Non-Randomized
Controlled Upperclassman-Randomized	CUp-R	Sophomores and Juniors	same as above	Randomized

Uncontrolled freshman samples. Data were collected from 3,562 freshman students who completed the college self-efficacy measure as part of an on-line survey designed by the university to gather information about the incoming class. These surveys were approximately 60 items in length, with the college self-efficacy measure administered last. These students completed the instrument on their own time, unsupervised, prior to arriving on campus for the start of the Fall 2006 semester. Given this, we considered this a very uncontrolled testing context. The total sample was randomly split for replication purposes, and after screening the data and removing any outlying cases, sample sizes were 1,586 and 1,585 for Samples 1 (Uncontrolled Freshman 1: UnFr-1) and 2 (Uncontrolled Freshman 2: UnFr-2), respectively.

Very controlled upperclassman sample. Sample 3 consisted of 237 university upperclassmen (i.e., 66% sophomores, 22% juniors, and 11% seniors) recruited from the psychology participant pool during the Fall 2006 and Spring 2007 semesters. These students completed the instrument along with several other motivation-related measures in a small (~20 seats) classroom setting. The instruments were administered to the students by handing out a manila envelope containing all measures. It took approximately 40 minutes to complete the battery of instruments, and the college self-efficacy measure was administered first in all sessions. Participants completed the measures one at a time and were not allowed to begin responding to the next measure until everyone had completed the current measure. Each measure's instructions were read aloud by a trained proctor prior to student beginning the measure. This process was employed as an attempt to slow response rates, in the hopes that it would produce more thoughtful responses. Thus, Sample 3 (Very Controlled Upperclassman: VCUp) completed the college self-efficacy instrument in a highly controlled context.

Controlled upperclassman and controlled upperclassman-randomized samples. Data for samples 4 and 5 were collected from a total of 854 upperclassman students. These participants completed the college self-efficacy measure during a mandatory university-wide assessment day during the Spring 2007 semester. The data collected during the assessment day were used for program effectiveness initiatives on campus. That is, the data were high-stakes for the administrators of programs on campus but of no stakes to the students completing the measures. Students completed a three-hour battery of tests in large (i.e., number of seats ranged from 63-250), lecture-style classrooms with proctors. The order of the tests differed across rooms, but the college self-efficacy measure tended to be administered during the last third of the testing session. We deemed this a slightly controlled testing context because, although there were proctors present, students were allowed to attend to the test as much or as little as they wanted. The combination of the larger room and the decreased proctor attention resulted in a higher degree of anonymity for the students and a potential for decreased motivation. After removing outliers and cases with missing data, Sample 4 (Controlled Upperclassman: CUp) consisted of 397 students and Sample 5 (Controlled Upperclassman-Randomized: CUp-R) consisted of 449 students. Sample 5 completed a version of the college self-efficacy instrument in which the order of the items was completely randomized.

Measures

College Self-Efficacy Inventory. The College Self-Efficacy Inventory (CSEI: Solberg, O'Brien, Villareal, Kennel, & Davis, 1993) was used to assess college self-efficacy and consists of 20 items written to represent participants' beliefs in their capabilities to successfully complete college-related tasks. Participants were asked to respond by indicating how confident they are in their ability to complete the task [1 (not at all confident) to 10 (extremely confident)]. The instrument, with its original and randomized item order, is presented in the Appendix.

Although the CSEI was administered for program evaluation purposes, a second, and equally important, purpose for its administration was to examine its psychometric properties. There had been little previous research on the properties of the instrument, and what existing research there was led us to believe that additional work on the measure may be necessary before trusting the inferences we made from its scores. It was important to collect data to evaluate its properties in the same context it would be gathered when used for program assessment: no-stakes. Moreover, it is difficult to imagine a situation in which students would complete this type of measure in a high-stakes environment. Therefore, we believe the contexts used in this study have high external validity.

Results

Confirmatory Factory Analyses

Confirmatory factor analysis (CFA) was used to test four models. All CFAs were conducted using LISREL 8.72 (Jöreskog & Sörbom, 2005). Because data screening indicated that the data for all samples were multivariate nonnormal, the Satorra-Bentler (S-B) correction was used in conjunction with maximum likelihood estimation to produce a corrected χ^2 and corrected standard errors. Global model-data fit was evaluated using the χ^2, along with the standardized root mean square residual (SRMR, with values of .08 or less indicating good model-data fit), the S-B adjusted root mean square error of approximation (RMSEA, with values of .07 or less indicating good model data fit), and the S-B adjusted comparative fit index (CFI, with values of .95 or above indicating good model-data fit). Areas of local misfit were identified by examining the standardized covariance residuals, which describe how well a model is able to reproduce each pair-wise relationship among items. These values can be positive or negative, indicating under- or over-representation of relationships, and absolute values of three or greater have been suggested as values to indicate a poorly reproduced relationship (Raykov & Marcoulides, 2000).

We were interested in examining whether model-data fit differed across the testing conditions. Although we expected there to be model-data misfit for all testing contexts given previous study of the measure, we expected greater overall misfit in the less controlled contexts compared to the controlled context. Given model-data misfit, we were then interested in examining how localized areas of misfit replicated across the testing conditions. Specifically, we questioned whether the more controlled condition would have fewer but *similar* areas of misfit than the other conditions or whether the more controlled condition would have fewer and *different* areas of misfit than the other conditions. Because specific areas of misfit often guide scale modification, ultimately we were interested in whether we would make different recommendations regarding scale modifications and refinement across the testing conditions.

Uncontrolled Freshman Samples 1 and 2

The theoretical model (Model 1) was fit to the data for the UnFr-1 sample and did not fit the data well (Table 2). Specific areas of misfit associated with this model were diagnosed by examining the standardized covariance residuals (Table 3). For Model 1, there were 41 standardized covariance residuals greater than three in absolute value, providing further evidence of model misfit. Theoretical and empirical considerations were used to derive and test a series of modified models through an iterative process until finding a model that fit the data adequately. Specifically, modifications were made to address areas of localized misfit, given that there was a theoretical or practical reason for doing so (e.g., redundancy in items, misalignment between item and subscale content). Three modified models were tested, with a 15-item three factor model providing adequate global model-data fit (Table 2).

Table 2
Fit Statistics for Hypothesized and Modified Models

	Model	χ^2_{S-B}	df	SRMR	RMSEA$_{S-B}$	CFI$_{S-B}$
UnFr-1						
	1) 17-item, three-factor	2135.30	116	0.075	0.10	0.92
	2) 16-item three-factor a	1444.85	101	0.075	0.09	0.94
	3) 16-item three-factor b	1286.40	101	0.054	0.09	0.95
	4) 15-item three-factor	976.92	87	0.051	0.08	0.95
UnFr-2						
	1	1886.78	116	0.069	0.10	0.94
	2	1282.44	101	0.069	0.09	0.96
	3	1224.22	101	0.055	0.08	0.96
	4	820.45	87	0.048	0.07	0.97
VCUp						
	1	349.39	116	0.083	0.090	0.90
	2	223.05	101	0.079	0.070	0.94
	3	225.05	101	0.076	0.070	0.94
	4	182.78	87	0.070	0.070	0.95
CUp						
	1	575.68	116	0.079	0.09	0.95
	2	396.97	101	0.079	0.09	0.95
	3	368.79	101	0.068	0.08	0.96
	4	305.09	87	0.062	0.08	0.96
CUp-R						
	1	674.41	116	0.076	0.10	0.92
	2	522.85	101	0.078	0.10	0.93
	3	487.85	101	0.073	0.09	0.94
	4	451.72	87	0.069	0.10	0.93

Note. The 17-item, three-factor model is a model in which three of the original items were removed prior to analyses due to poor functioning found in prior studies and item content issues; the 16-item three-factor a model is the model in which item 5 was removed; the 16-item three-factor model b is the model in which item 5 was removed and item 1 was moved to the Roommate subscale; the 15-item three-factor model is the model in which item 1 was removed from the scale.

Table 3

Areas of Localized Misfit for Models

	Model 1		Model 2		Model 3		Model 4	
	# resids > | 3 |	item pairs w/resid > | 5 |	# resids > | 3 |	item pairs w/resid > | 5 |	# resids > | 3 |	item pairs w/resid > | 5 |	# resids > | 3 |	item pairs w/resid > | 5 |

UnFr-1	41	1 with 2, 15, 16, 20;
		2 with 3, 4;
		3 with 5, 6;
		4 with 18, 19;
		5 with 6, 11;
		6 with 11, 18;
		9 with 17
			33	**1 with 2**, 15, 16, 20;
				2 with 3, 4;
				4 with 18, 19;
				6 with 17;
				9 with 17;
				11 with 18
					29	1 with 3;
						2 with 3, 4;
						4 with 18, 19;
						6 with 17;
						9 with 17;
						11 with 18
							22	**2 with 3**, 4;
								4 with 18, 19;
								6 with 17;
								9 with 17;
								11 with 18

UnFr-2	43	1 with 2, 15, 20;
		2 with 3, 4;
		3 with 4, 5, 6, 20;
		4 with 18;
		5 with 6, **11**;
		6 with 11, 18;
		8 with 18
			33	**1 with** 2, 3, 15, **20**;
				2 with 3, 4;
				3 with 4, 13, 20;
				4 with 18;
				8 with 18;
				11 with 15
					32	1 with 3, 6, 16, 20;
						2 with 3, 4;
						3 with 4;
						4 with 18;
						8 with 18;
						11 with 15
							25	**2 with 3**, 4;
								3 with 4, 20;
								4 with 18;
								8 with 18

VCUp	8	4 with 15; **5 with 6**	9	**4 with 15**	9	**4 with 15**	5	4 with 15

CUp	22	1 with 20;
		2 with 3;
		5 with 6;
		9 with 14
			22	1 with 20;
				2 with 3;
				6 with 13;
				9 with 14
					18	1 with 3;
						2 with 3;
						6 with 13;
						9 with 14
							16	2 with 3;
								6 with 13;
								9 with 14;
								11 with 16

CUp-R	19	3 with 5;
		5 with 6;
		6 with 18;
		8 with 19
			16	6 with 19;
				8 with 19
					15	6 with 17, 19;
						8 with 19;
						9 with 15
							13	6 with 17, 19;
								8 with 19

Note. Largest residual indicated by bolded item pair.

Although modifications to the tested models resulted in improved fit for the UnFr-1 sample, there are several problems associated with re-specifying and testing modified models on the same sample (MacCallum, Roznowski, & Necowitz, 1992). Because the fit of the modified models may capitalize on chance (i.e., fitting the idiosyncrasies of the sample), the fit of modified models may not generalize to other samples. Given this, all models were tested again using the UnFr-2 sample to (a) determine whether the pattern of misfit associated with the four theoretical models was reproduced in an independent sample, (b) provide the first a priori testing of the modified models. As expected, results for UnFr-2 were extremely similar to UnFr-1, both in regard to global fit and areas of local misfit (Tables 2 and 3). This was not a surprise given that the two samples were derived by randomly splitting the overall sample and both were fairly large in size, which results in more stable estimates.

Despite the adequate global fit for the 15-item, three-factor model, several areas of local misfit remained for both samples, as evidenced by a number of large residuals. Especially puzzling were the large residuals associated with the relationships between items 2, 3, and 4. These three items represent different subscales and appear to represent completely different areas of confidence. One possible explanation lies in the fact that these items were presented in succession, and the strong relationships may have been caused by an item-ordering effect; especially when expressing attitudes, preceding questions can influence the responses given to subsequent ones (e.g., Schwarz, 1999; Tourangeau & Rasinksi, 1988). It is possible that these items were correlated with one another simply because they were located next to one another on the instrument.

Very Controlled Upperclassman Sample

The results from the UnFr-1 and -2 samples indicated that the three-factor model (Model 1) did not fit the data well and that, even after removing two items that consistently performed poorly across samples, a great deal of localized misfit remained. Again, the important point is that areas of misfit replicated across the two random samples from the uncontrolled condition, and if these were our only samples, we may claim there was no clear structure to the data and most likely recommend not using the measure for assessment purposes. We were now interested in evaluating if these same results would emerge for data collected in a controlled condition. Thus, data for the VCUp sample were gathered to address these concerns.

Similar to the UnFr-1 and -2 samples, the theoretical model did not fit the data (Table 2). Additionally, the *patterns* of local misfit for Model 1 were similar, although not identical, to those found using the Freshman samples. In order to fully compare the results across samples, the three modified models tested using the UnFr-1 and -2 samples were fit to data from the VCUp sample, and the reduced 15-item, three-factor model provided fairly good model-data fit. Moreover, it is quite interesting to note that the local misfit associated with items 2, 3, and 4 no longer was present, and overall, standardized covariance residuals were fewer in number and smaller in magnitude, with values between 0 and 1 for most items (Table 3).

Obviously, one possible explanation for the substantially better local fit concerns the method of administration. Unlike UnFr-1 and -2, students in the VCUp sample completed the instrument in a much more controlled testing context. It is very likely that the high number of large residuals were not present for this sample because these students provided more thoughtful answers to the questions and were not able to simply rush through the questionnaires. However, it is important to note that the age of the student in the controlled condition was different from that in the uncontrolled condition; students in the controlled condition were older and had more experience in college. Because there were two variables that changed between these samples (i.e., freshman vs. upperclassman and uncontrolled vs. controlled condition), it is not possible to disentangle which was the cause of the better model-data fit.

Controlled Upperclassman Samples

The CUp and CUp-R samples were used to collect data to address questions raised by the results from the previous three samples. Specifically, one question concerned why there were fewer areas of local misfit when using the VCUp sample compared to the UnFr- and -2 samples. As noted above, one possibility could be the method of administration (an uncontrolled condition vs. a very controlled setting with explicit instructions to answer slowly and carefully); however, it is possible that the year in school of the participants was the underlying factor contributing to these differences. The CUp sample (i.e., upperclassmen in a slightly controlled condition) was gathered to help disentangle these variables. We believed the reduction of misfit for the very controlled condition was due to the testing environment and not the age of the student. Therefore, we expected to find more misfit associated with the CUp sample (upperclassmen in slightly controlled condition) compared to VCUp (upperclassmen in a very controlled setting).

A second question that remained was why items 2, 3, and 4 in particular exhibited large residuals. We believed we were seeing an item-order effect (e.g., Schwarz, 1999; Tourangeau & Rasinksi, 1988) due to low motivation. Specifically, if students don't respond in a thoughtful manner, they may choose similar response options for items placed next to each other on the measure. The CUp-R sample was used to test this hypothesis. That is, if items 2, 3, and 4 were no longer positioned next to each other on the scale and the testing condition was slightly controlled, would the items still have large standardized residuals? We hypothesized that they would not; instead, items positioned next to each other in this new randomized order would have large residuals.

As found previously, the theoretical model did not fit the data for the CUp sample (Table 2). Examination of the standardized covariance residuals for the models (Table 3) indicated that, overall, patterns local misfit was similar to that found for the UnFr-1 and -2 samples. As expected, the number of standardized covariance residuals was higher than that found using the VCUp sample, and the specific misfit associated with items 2, 3, and 4 was again found, suggesting that its presence was a function of the testing condition (i.e., degree of control) rather than age.

The CUp-R sample was used to determine whether the misfit associated with items 2, 3, and 4 found using the UnFr-1 and -2 samples and the CUp sample was an item ordering effect caused by low motivation. Again, the theoretical model did not fit the data adequately (Table 2). Misfit associated with Model 1 and the three modified models was again examined (Table 3). Consistent with all samples completing the non-randomized version of the instrument, the largest residual for the CUp-R sample was that for the relationship between items 5 and 6. However, the pattern of residuals overall was *not* very similar to those found in any of the previous samples that administered the non-randomized version. Moreover, Model 4 (the 15-item, three-factor model) did *not* fit the data from this sample, which is understandable since the model was created based on misfit from the previous samples and since the CUp-R sample did not share the same areas of misfit. Of particular importance, there were no longer large residuals associated with items 2, 3, and 4, but there *was* misfit associated with items 6 and 19, which were located next to one another in the randomized version. This suggests that the misfit associated with items 2, 3, and 4 was indeed an item order effect caused by testing context. Essentially, these results highlight the fact that in low-motivation contexts there can be dependencies among items simply because they are located adjacent to one another on an instrument and that randomizing the order of the items will result in *different* sets of items displaying these dependencies. This of course will affect the psychometric properties of the measure (i.e., the factor structure) and subsequent decisions regarding test refinement.

Discussion

Given the risks associated with using low-stakes data and the widespread use of this type of data for instrument development purposes, this research was conducted to examine the dimensionality of college self-efficacy scores from multiple samples gathered in several different testing contexts in order to determine whether the amount of proctor control impacted the fit of the data to the tested models. Although some similarities were found across all samples and testing conditions (e.g., the theoretical model did not fit, there was an extremely large standardized covariance residual for item 5 and 6), there were differences in model-data fit across the three testing conditions. As the testing conditions increased in level of control, the amount of localized misfit decreased. That is, the largest numbers of standardized covariance residuals were found when using data collected in an uncontrolled testing condition (i.e., UnFr-1 and -2 samples), smaller numbers of residuals were found when using data collected in a controlled condition (i.e., CUp and CUp-R samples), and the smallest numbers of residuals were found when using data collected in a very controlled condition (i.e., VCUp). Thus, the measure could have been considered inadequate when employing the two Uncontrolled samples and the two Controlled samples, whereas it may have been considered acceptable when employing the Very Controlled sample. If item deletion was conducted in order to create a "better" measure, more items would be removed from the test using these Uncontrolled or Controlled samples than if conducting the same process using the Very Controlled sample. As items are labor-intensive to construct and, in turn expensive to write, throwing out quality items is something instrument developers and evaluators would like to avoid. Collecting data in a controlled setting appears to minimize the chance of removing quality items.

Thus, one possible way to alleviate these problems is to increase the level of control in the testing condition, as was done with the Very Controlled sample. Specifically, the participants in this sample completed the instrument in a small campus classroom with the experimenter present, were given explicit instructions to carefully answer the questions, and were not allowed to rush through the questionnaires. This was done to slow responding in the hopes that participants would provide more thoughtful responses to the questions. As mentioned previously, the residuals for the tested models were fewer in number and much smaller in magnitude for this sample than they were for the samples who participated in the large-scale testing.[2] Thus, it appears that the testing condition played a very important a role in how much effort participants put into their responses, how thoughtfully they responded, how well the models fit the data, and ultimately the proposed modifications to the measure. Slowing responding eliminated what appeared to be a sort of response style/acquiescence and eliminated some of the dependency of the items on one another.

One related and particularly concerning result of this study involves the dimensionality and pattern of residuals obtained for the CUp-R sample, which received the randomized form of the CSEI. Although randomizing the item order eliminated the residuals between items 2, 3, and 4, the overall patterns of misfit for the models were alarmingly dissimilar when fit to these data than when fit to data from samples who received the non-randomized form. Moreover, the modified models fit the data worse in this sample than any other. It is important to note that this was true when comparing the CUp sample to the CUp-R sample, which involved the same age students (i.e., upperclassmen) in the same testing condition (less controlled testing situation); the only aspect that differed was item order. It is very possible that all modifications made to the instrument in the original item order might not have been made using this randomized order in an uncontrolled setting and other modifications *would* have been made. However, the results of this study do suggest that students attend to items to a higher degree when in a more controlled testing context, resulting in a clearer understanding of item functioning. It follows that the effects of randomizing the item order on model-data fit may not be so problematic if data are collected in a more controlled testing condition. Presumably, a more controlled testing condition and the subsequent decrease in error variance would allow areas that are truly problematic to be identified. We unfortunately did not have a sixth sample to test this hypothesis, and we call for additional work in this area.

Conclusion

The results from the current study have serious implications for the manner in which data for instrument development should be gathered. In an instrument development context, data are typically gathered through a large-scale testing program or a university participant pool (i.e., an environment that is extremely low-stakes to the test takers providing the data), several models are fit to that data, and changes to the instrument are made based on areas of misfit associated with the tested models. However, this study has shown that the amount of misfit present is dependent upon how controlled the testing condition is. Because of this, data collected from students in an uncontrolled testing condition might lead assessment specialists or test developers to make unnecessary changes (or fail to make necessary changes) to an instrument. On the other hand, a testing condition in which there is a high degree of control, although more costly in terms of time and resources, appears to increase student motivation despite the fact that the test is low-stakes to these students. Consequently, the test developer is more able to trust the validity of inferences made regarding these scores and will therefore make more appropriate decisions about changes to an instrument. This is important given that sound assessment practice begins with appropriate and well-functioning instruments, and before one can trust the inferences made regarding student performance or development and, ultimately, program effectiveness, one must be able to trust the instrument with which these are measured.

References

Barry, C. L., & Finney, S. J (2007, October). *A Psychometric Investigation of the College Self-Efficacy Inventory.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*, 29-49.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13,* 55-77.

Gore, P. A., Leuwerke, W. C., & Turley, S. E. (2006). A psychometric study of the College Self-Efficacy Inventory. *Journal of College Student Retention: Research Theory & Practice, 7,* 227-244.

Jöreskog, K. G., & Sörbom, D. (2005). LISREL (Version 8.72) [Computer software]. Lincolnwood, IL: Scientific Software.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490 – 504.

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.

Solberg, V. S., O'Brien, K., Villareal, P., Kennel, R., & Davis, B. (1993). Self-efficacy and Hispanic college students: Validation of the College Self-Efficacy Instrument. *Hispanic Journal of Behavioral Sciences, 15,* 80-95.

Sundre, D. L (1999). *Does examinee motivation moderation the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Sundre, D., L. & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29,* 6–26.

Sundre, D. L., & Moore, D. L., (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14,* 8-9.

Tourangeau, R., & Rasinksi, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Wise, S.L., Bhola, D.S., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25,* 21-30.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

Wise, S. L., & Kong, J. (2005). Response Time Effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11,* 65-83.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.

Footnotes

[1] Although the instrument consists of 20 items, three items were removed prior to testing these models. These three items had functioned poorly in past studies of the instrument (e.g., Barry & Finney, 2007; Gore, Leuwerke, & Turley, 2006; Solberg et al., 1993) and were written such that they may not be relevant to all students. Thus, all models tested in this paper were based on the remaining 17 items.

[2] One might question whether the differences in the number and magnitude of the standardized residuals were due to differences in sample sizes rather than differences in the level of control. This is because the standardized covariance residuals used to examine misfit are computed by dividing the covariance residuals by the standard error. Given that standard errors can be affected by the sample size (i.e., smaller samples tend to yield larger standard errors and, in turn, may lead to smaller standardized covariance residuals), it was possible that the large residuals in the UnFr-1 and -2 samples were due to their large N, that the moderate residuals in the CUp and CUp-R samples were due to their smaller N, and that the small residuals in the VCUp sample were due its small N. To ensure that this was not a plausible explanation for the pattern of results, all analyses were conducted a second time, using the correlation matrix (i.e., the standardized covariances) as input; when conducted in this manner, correlation residuals are computed, which are not impacted by the standard error and consequently the sample size. The results indicated that the correlation residuals followed a similar pattern and had similar relative magnitudes, providing evidence that the differences in the number and magnitude of standardized covariance residuals across the five