RESEARCH & PRACTICE IN ASSESSMENT

VOLUME ONE| WINTER 2006
www.RPAjournal.com
ISSN#2161-4210

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

Editors

Table of Contents

Comments from the Editor

Robin D. Anderson
Director of Institutional Research and Effectiveness
Blue Ridge Community College

This release of the first issue of *Research & Practice in Assessment* marks a new era in assessment related scholarship. The Board of the Virginia Assessment Group (VAG) and the Board of Editors of *Research & Practice in Assessment* have committed to a publication focused on student learning outcomes assessment. To some this focus may seem narrow; however, one needs only to review the inaugural issue of this journal to see the breadth of scholarship in the area of outcomes assessment. I am pleased to include four excellent articles addressing four different issues in assessment. Topics in this issue include assessing the impact of writing acros the curriculum, the handling of missing data, the use of effect size with the NSSE, and a wonderful piece on improving the assessment of student learning through peer review. I hope you will find this new era and this first issue as exciting, interesting, and informative as I do.

Finally, I would like to thank all of those who have worked over the last 18 months to develop and launch *Research & Practice in Assessment*. The Officers and Board Members of the Virginia Assessment Group worked tirelessly to make the needed changes to the organization's by-laws to support the establishment of the journal. The members of VAG voted overwhelmingly to support the development of the publications, transforming an organizational newsletter to an electronic journal. And a special thanks to Keston H. Fulcher, Ph.D. (Director of Assessment and Evaluation at Christopher Newport University), Dennis R. Ridley, Ph.D.(Director of Institutional Research and Planning at Virginia Wesleyan College) and Rufus Carter (Coordinator of Institutional Assessment at Marymount University) who make up the Board of Editors for their time and commitment to this new publication. I hope that upon reading *Research & Practice in Assessment* you will be inspired to contribute to the literature on learning outcomes assessment by submitting your own work for consideration in the next issue.

Solutions for Missing Data in Structural Equation Modeling

Rufus Lynn Carter
Marymount University

Abstract

Many times in both educational and social science research it is impossible to collect data that is complete. When administering a survey, for example, people may answer some questions and not others. This missing data causes a problem for researchers using structural equation modeling (SEM) techniques for data analyses. Because SEM and multivariate methods require complete data, several methods have been proposed for dealing with these missing data.What follows is a review of several methods currently used, a description of strengths and weaknesses of each method, and a proposal for future research.

Methods for Dealing with Missing Data

Listwise Deletion

Listwise deletion is an ad hoc method of dealing with missing data in that it deals with the missing data before any substantive analyses are done. It is considered the easiest and simplest method of dealing with missing data (Brown, 1983). It involves removing incomplete cases (record with missing data on any variable) from the dataset. This means the researcher removes all the records that have missing data on any variable. Depending on the sample size and number of variables this can result in a great reduction in the sample size available for data analysis. Listwise deletion assumes that the data are missing completely at random (MCAR). Data are missing completely at random when the probability of obtaining a particular pattern of missing data is not dependant on the values that are missing and when the probability of obtaining the missing data pattern in the sample is not dependant on the observed data (Rubin, 1976). An advantage in using listwise deletion is that all analyses are calculated with the same set of cases.

Pairwise Deletion

Another ad hoc method of dealing with missing data, pairwise deletion (PD), uses all available data. This means for each pair of variables PD calculates the covariance estimates from all cases with complete observations on both variables (Wothke, 1998). Pairwise deletion assumes that the data are missing completely at random (MCAR). Cases are removed when they have missing data on the variables involved in that particular computation (Kline, 1998). This can be problematic in that each element of the covariance matrix could be based on different groups of subjects. For example, if 300 subjects had complete scores for variables X_1 and X_2 then the effective sample size for the covariance between X_1 and X_2 is 300. Likewise, if 200 subjects had complete scores on X_1 and X_3 then the sample size for this covariance would be only 200. Kline (1998) points out that it would be impossible to derive some of these covariances if they were calculated using data from all subjects as in listwise deletion.

Imputation

The method of imputation involves placing estimated scores into the data set in the location of the missing data. Kline (1998) discusses three basic types of imputation. In each of these three types of imputations, the data are assumed to be MCAR. Mean imputation involves substituting missing cases with the overall sample average for each particular variable with missing data. While simple to execute this method does not take into consideration subjects patterns of scores across all the other variables. Regression imputation takes this into consideration by predicting a score for each subject by using multiple regression based on their non missing scores for other variables. For this method to work, Kline states that the variable with missing data must co-vary at least moderately with the other variables.

Pattern matching is the third from of imputation Kline (1998) describes. In this method the missing score is replaced with a score from another subject who has a similar profile of scores across the other variables. This method is not widely available on software packages but is available via PRELIS2 (Joreskog & Sorbom, 1996b), which performs pattern matching and can be used with LISREL. Kline notes these methods seem to work best with the proportion of missing data is low and scattered across different variables. Another imputation method is that of multiple imputations with the Expectation-

Maximization (EM) algorithm. Dempster, Laird, and Rubin (1977) presented an algorithm for computing maximum likelihood estimates from missing data sets. Each iteration of their algorithm consists of an expectation step followed by a maximization step. They assume a family of sampling densities $f(x|\phi)$ depending on parameters ϕ and they then derive their corresponding family of sampling densities $g(y|\phi)$. The EM algorithm attempts to find a value of ϕ which maximizes $g(y|\phi)$ given an observed y, but it does this by making use of the related family $f(x|\phi)$. Schafer and Olsen (1998) state that with the development of the EM algorithm, statisticians have stopped viewing missing data as a "nuisance" and have reevaluated it as a source of variability to be averaged over. Schafer and Olsen describe a technique developed by Rubin (1987) where each value is replaced with a set of $m > 1$ plausible values which allows the variances reported above to be averaged by simulation. After performing multiple imputations, each of these m data sets can be analyzed by SEM techniques intended for complete data. Then through a series of complex rules the estimates and standard errors are combined to provide overall estimates and standard errors that reflect missing data uncertainty. These rules properly applied are thought to provide unbiased estimates.

Schafer and Olsen (1998) describe their own iterative process, data augmentation (DA), which alternately fills in the missing data and makes inferences about the unknown parameters. The process is similar to the EM algorithm as DA fills in the missing data either randomly or else based on conjecture. DA performs a random imputation of missing data under assumed values of the parameters and then draws new parameters from a Bayesian posterior distribution based on the observed and imputed data. Schafer and Olsen explain the Bayesian distribution as requiring the researcher to specify a prior distribution for the parameters of the imputated model. Schafer (1997) developed a computer program NORM using the multivariate normal distribution to generate imputations for the missing values.

Schafer and Olsen (1998) note that multiple imputation methods resemble other methods of ad hoc case deletion because it addresses the missing-data issue at the beginning, before substantive analyses are run. They argue that unlike the other ad hoc methods, multiple imputations do not have to be MCAR but instead need only meet the less rigorous assumptionthat the missing data are missing at random (MAR). Data are missing at random when probability of obtaining a particular pattern of missing data is not dependant on the values that are missing (Rubin, 1987). Schafer and Olsen also state that while multiple imputation techniques are statistically defensible and incorporate missing-data into all summary statistics, they do suggest that the direct maximum likelihood methods may be more efficient than multiple imputations because they do not rely on simulation.

SEM Methods

One option available by SEM to deal with the problem of missing data is illustrated by Allison (1987). He proposes a maximum likelihood estimation for incomplete data. His model assumes multivariate normality, which as he states implies that the means, variances, and covariances are the sufficient statistics. However he also states that violations of multivariate normality will not seriously compromise the estimates. Allison discusses a confirmatory factor model where the goal is to estimate the correlation between father's occupational status (FAOC) and father's educational attainment (FAED) for black men in the U.S. He reports previous studies had estimated the correlation to be 0.433. He split a sample of 2,020 taken from Bielby et al. (1977b) into two groups, 348 with complete data and 1,672 with incomplete data. The small complete sample had two indicators of FAOC (y_1 and y_2) and two indicators of FAED (y_3 and y_4). The large sub-sample had only y_1 and y_3. Allison states that this design virtually guarantees that the missing data are missing completely at random. Sample variances and covariances for the complete-data sub-sample were obtained from the correlation matrix and standard deviations in the original study. By calculating sums of squares and crossproducts from the reported correlations and standard deviations of sample with the missing data, comparisons can then be made between the re-measurement sample and the full sample. These values are then used to recreate the covariance matrix for the sample with missing data (Allison, 1987). He goes on to state that while his method using LISREL produces non-biased estimates; it is exceedingly complex with the addition of more variables. The relationship of number of variables to number of possible missing data patterns is 2^k- 1. In these cases Allison (1987) suggests using the previously mentioned listwise and pairwise ad hoc practices to eliminate minor missing data patterns. His LISREL runs require the sample means and requires that each latent variable in each sub-sample have at least one indicator with a fixed, nonzero λ coefficient. The nonzero λ coefficients for y_1 and y_3 are fixed at 1.0, which define the metrics for the latent variables (Allison, 1987). For the sub-sample with no observations on y_2 and y_4 he set λ_{21}, λ_{23},λ_{42}, and λ_{43} equal to 0.0 and constrained variances ε_2 and ε_4 equal to 1.0. All the free parameterswere constrained to be equal across sub samples (Allison, 1987).

Another method of using maximum likelihood to estimate missing data is the Full-Information Maximum Likelihood (FIML) method. "The FIML method uses all of the information of the observed data, including mean and variance for the missing portions of a variable, given the observed portion(s) of other variables" (Wothke, 1998). Muthén, Kaplan, and Hollis (1987) present how the method applies to structural equation modeling. They state that their method using LISREL allows for the latent variable model to include missingness. Their paper examines maximum likelihood estimation of the θ parameters. Wothke (1998) states that FIML assumes multivariate normality, and maximizes the likelihood of the model with the observed data. He also states that two structural equation modeling programs, AMOS (Arbuckle, 1995) and Mx (Neale, 1994), implement this FIML method for dealing with missing data. He critiques other methods for estimation using FIML and states that those approaches are only practical when the data have just a few distinct patterns of missing data. In addition, he and states that using AMOS (Arbuckle, 1995) and Mx do not require the same level of technical expertise as do the methods of presented by Dempster et al. (1977) and Muthén et al. (1987) do. Wothke (1998) suggests that both AMOS and Mx maximize the case-wise likelihood of the observed data, computed by minimizing the function. He further states that both AMOS and Mx are not limited by the number of missing-data patterns, and do not require complex steps to accommodate missing data.

Comparisons of Methods in the Literature

Several of the techniques described earlier have been compared to determine which yields the least biased estimates in SEM. Wothke (1998) examined listwise, pairwise, mean imputation and maximum likelihood methods for growth curve modeling for examples where the data were MCAR and MAR. For the MCAR data estimates of the model parameters were unbiased for FIML, LD and MD, while mean imputation showed no bias in means but exhibited strongly biased variance and covariance estimates. For the MAR data FMIL produced unbiased estimates while PD estimates exhibited a small negative bias. Listwise deletion and mean imputation methods resulted in sampling distributions that did not include the parameter value. Similar results are reported in the literature by Muthén et al. (1987) and Arbuckle (1996). In these and other studies the comparison seems to be that of FIML methods with listwise and pairwise deletion. The results of the comparisons of these methods in the literature indicate that when the data are MCAR there is little difference in the estimation bias for listwise deletion, pairwise deletion and maximum likelihood. Some other comparisons were notably absent from the literature, and are the subject of the research proposal discussed below.

Future Research

In the literature, little attention has been paid to the use of pattern-based imputation. For MCAR data it would appear to be a viable alternative to listwise and pairwise deletion and perhaps to both multiple imputation methods and maximum likelihood methods. Further investigation into this area is needed.

One suggestion is to generate population values from a complete data set having no missing values. A random number generator like that found in SAS (version 9) software can provide random missing data points for an adequate number of data sets. A single model can be fit to each random sample taken from the original population sample as described above. Model fit can then be examined using FIML, listwise deletion, pairwise deletion, an application of the EM algorithm using NORM (Schafer, 1997) and finally the pattern-matching imputation method. This will enable researchers to make comparisons about estimate bias for missing data in SEM for the MCAR condition.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), *Sociological Methodology* (pp. 71-103). San Francisco: Jossey-Bass.

Arbuckle, J. L. (1995). *AMOS for Windows Analysis of Moment Structures.* Version 3.5. Chicago: Small Waters Corp.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.), *Advanced structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Publishers.

Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika*, *48(*2), 269-292.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B*,* *39*, 1-38.

Kline, R. B. (1998). Principles and practices of structural equation modeling. New York: Guilford.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431-462.

Neale, M. C. (1994) Mx: *Statistical modeling (2nd Edition)*. Department of Psychiatry: Medical College of Virginia.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 61*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. London: Chapman.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst'sperspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Wothke, W. (1998). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, and J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples.* Mahwah, NJ: Lawrence Erlbaum Publishers.

Peer Review and Organizational Learning: Improving the Assessment of Student Learning

Craig Herndon
State Council of Higher Education for Virginia

Abstract

Virginia's assessment of student learning outcomes has been lauded by national organizations for its respect of institutional autonomy while providing meaningful information on student learning outcomes. Virginia recently implemented a process by which each institution's plan to assess student learning outcomes are evaluated by peer institutions. The application of peer review to plans for assessment is described in greater detail and critiqued using the theoretical lens afforded by organizational learning. The article concludes with discussion and recommendations for the improvement of the peer review process as it applies to assessment.

Introduction

Virginia's assessment of student learning outcomes has been lauded by national organizations for its respect of institutional autonomy while providing meaningful information on student learning outcomes (Epstein, 2005). Virginia recently implemented a process by which each institution's plans to assess student learning outcomes are evaluated by peer institutions. The application of peer review to plans to assess student competency is described in greater detail and critiqued using the theoretical lens afforded by organizational learning. The article concludes with discussion and recommendations for the improvement of the peer review process as it applies to assessment.

In order to best understand the application of peer review to the process of competency assessment in Virginia, it is necessary to begin by describing the process of competency assessment in Virginia and the recent addition of peer review to this process. Following a description of assessment of student learning in Virginia, a brief discussion of organizational learning is conducted that includes an operational definition of the term and a tentative description of the organization in question. Next, the use of peer review in higher education is surveyed with particular attention paid to the benefits and shortcomings of the process. Finally, a discussion of improvements to peer reviewed student outcomes assessment is undertaken and recommendations are made using the theoretical constructs provided by organizational learning.

It should be noted from the onset that the topics of organizational learning, peer review, and student competency assessment are far broader than the limitations of this article. This article seeks to illuminate the areas of critical overlap between organizational learning, the use of a peer review process, and the assessment of student learning in an attempt to improve the assessment process.

Competency Assessment in Virginia

In 1998, the Governor of Virginia charged a Blue Ribbon Commission with evaluating the needs and goals of higher education in Virginia for the 21st century. Among the specific charges to the Commission, the Governor requested that the Commission, "advise the Governor on how the institutions, administrators, and faculty that comprise Virginia's system of higher education can be made more accountable to their stockholders (the taxpayers, the parents, and the private contributors who finance the system) for the quality of the academic content and the outcomes accomplished through the investment of public funds." (Executive Order 1, 1998). The Commission concluded that evidence of high quality outputs is essential in assuring stockholders that the substantial investment made by the Commonwealth in higher education is producing results (Governor's Blue Ribbon Commission on Higher Education, 2000). In order to provide this assurance, the Commission identified six areas of core competency—areas of knowledge and skill that supersede majors, disciplines, and institutional missions— recommending that these areas be assessed regularly and the results of such assessments be shared with the public. The core competencies identified by the Commission included written communication, mathematical analysis, scientific literacy, critical thinking, oral communication, and technology.

The Code of Virginia was subsequently amended such that the State Council of Higher Education for Virginia (SCHEV), Virginia's coordinating body for higher education, was charged with "develop[ing] in cooperation with institutions of higher education guidelines for the assessment of student achievement" (Code of Virginia, 2000). Biennially, and starting in 2001, each public four-year institution of higher education in the Commonwealth submitted plans to assess competency in two specified areas. The first of three rounds of assessment required institutions to submit plans to assess student competency in written communication and technology/information literacy to SCHEV for approval one year prior to submitting results. SCHEV staff reviewed the plans in light of each institution's mission and provided feedback with notification of approval. The same process was used in 2003 when plans to assess student competency in scientific reasoning and quantitative reasoning (adapted from the Blue Ribbon Commission's recommendation to assess scientific literacy and mathematical analysis, respectively) were submitted to SCHEV staff for review and approval.

In 2005, SCHEV staff instituted a process of peer review, by which each institution's plans to assess critical thinking and oral communication were shared with two other institutions in the Commonwealth for the purpose of (a) providing each institution with expert feedback; and (b) initiating inter-institutional communication on topic of competency assessment plans for the purpose of providing mutual benefit to reviewer and the reviewed while spurring creative approaches to assessment. Some six months before the peer review process began, SCHEV staff solicited the opinion of assessment professionals regarding the use of a peer review process in place of a review of institutional plans to assess student competency conducted exclusively by SCHEV staff. General support for such a process was expressed by assessment professionals.

In the spring of 2005, each of Virginia's 15 public four-year institutions submitted plans approved by its chief academic officer to assess student competency in critical thinking and oral communication. Plans included a definition of the competency used by the institution, criteria and standards for determining competency, and a methodology for scoring and deeming students competent. Each institution's designated assessment coordinator was then provided with plans from two other institutions in the Commonwealth: one institution that was of the same Carnegie classification and one that was from a different Carnegie classification (Carnegie Classification, 2000). Each institutional representative was encouraged to form a committee of knowledgeable staff from his or her institution to review the four assigned competency assessment plans (two competencies from two universities) and provide written feedback in accordance with a set of suggested components. Examples of the suggested components provided to referees for the purpose of evaluating an institution's plans to assess competency read, "adequacy of criteria and standards for determining competency" and "appropriateness of competency to the mission, goals, and objectives of the institution" (SCHEV memo to assessment officers, 2005).

Peer reviews were collected and compiled by SCHEV staff who acted as editors of the peer review comments much in the way that a journal editor does of peer reviewed publication. SCHEV staff read each plan and each review before sharing anonymous feedback with each institution. The feedback outlined the concerns and praise raised by referees in addition to concerns and praise generated by SCHEV staff. The process resulted in a single blind review, in that the reviewers were explicitly notified of the institutions they were reviewing, while recipients of review were not notified of the institutions that conducted the review. A double blind process was not possible given that referees were required to compare each institution's plans to assess competency with the institution's mission, goals, and objectives. Each institution received peer review comments within 45 days of submitting plans to assess competency and one full year before the results from the competency assessments were due.

Organizational Learning

Organizational theory is the study of how "groups and individuals behave in varying organizational structures and circumstances" (Shafritz & Ott, 2001, p.1). The study of organizational theory helps those that manage higher education to understand complex concepts (Berger, 2000; Birnbaum, 1988). Further, the application of organizational theory to complex concepts allows for a more complete understanding of the concept and the ability to take wellinformed action (Berger 2000, Birnbaum, 1988; Bolman & Deal, 2003; Morgan, 1997). The term organizational theory refers broadly to the theoretical frames and perspectives applied to the study of organizational behavior (Morgan, 1997; Shafritz & Ott, 2001). In an effort to better understand the processes of peer reviewed student learning assessment and make recommendations for its improvement, the theoretical frame of organizational learning will be applied to peer review and student competency assessment.

Organizational learning is, in itself, an umbrella term for a set of organizational theories that ascribe learning characteristics to organizations (Morgan, 1997). Taxonomists of organizational learning have classified its theories in a number of ways (Argyris & Schön, 1996; Dierkes, Berthoin Antal, Child,

& Nonaka, 2001; Morgan, 1997). Dierkes et al. (2001) distinguish between theories that speak to the creation of new knowledge and those that speak to the sharing, using, and storing of knowledge. Morgan (1997) distinguishes between theories associated with acquiring, processing, and using knowledge and those associated with storing and accessing knowledge. Argyris and Schön (1996) distinguish between a practically-oriented branch of organizational learning and a scholarly-oriented branch of organizational learning that is distant from practice. The immediate discussion of organizational learning, as it applies to the peer review process of plans to assess student learning outcomes, will begin by focusing on how knowledge is practically shared, processed, and used. The discussion will then migrate to the formation of new knowledge.

In order to best understand how organizational learning will be applied to the concept of peer review and ultimately the assessment of student learning, it is necessary and appropriate to provide a functional definition of organizational learning that narrowly describes its use in this process by which organizations share, process, and use knowledge by scanning the environment, comparing what is observed to operating norms, and correcting accordingly. The act of scanning, comparing, and correcting reflects a single loop learning orientation. A double loop learning orientation to organizational learning, in which existing norms are questioned, will be introduced in the discussion section.

Community of Practice

Given the existence of learning organizations, it must be established that an organization among assessment professionals in the Commonwealth of Virginia exists if the theory is to be applied to the discussion at hand. Wenger and Snyder (2001) coined the phrase community of practice to describe people informally bound by shared expertise who, in turn, share knowledge beyond the traditional boundaries of their formal organization for the purpose of creatively approaching shared problems. Looking at this definition in parts, the argument can be made that assessment professionals at Virginia's public institutions of higher education (a) contain shared expertise; (b) are formally bound to their own institution; and (c) are encouraged to share knowledge with colleagues for the purpose of approaching shared problems through a process of peer review.

The argument that assessment professionals in Virginia constitute an informal organization is not without its flaws. First and foremost, the association of assessment professionals formed through peer review is not a free association, as the members have been, to some degree, compelled to participate. But the very nature of organization results in variation in participation levels and willingness to participate among members (Bolman & Deal, 2003; Morgan, 1997). Second, members of the loosely coupled organization of peer reviewers are not bound to creatively approach problems. Yet the most recent assessment of the critical thinking and oral communication core competencies were thought to be the most difficult competencies to assess of the six that Virginia has identified. Peer review, as a mechanism for sharing information, was intended to heighten creative problem solving with regard to the development of plans to assess critical thinking and oral communication.

Peer Review

Peer review is a widely practiced form of certifying quality in higher education. Peer review has been described as a formative evaluation process in which participants work collaboratively to strengthen a product (Keig & Waggoner, 1994). Common uses of peer review in higher education include the awarding of research funds, evaluating academic publications, reviewing faculty performance for tenure and promotion, and granting regional and disciplinary accreditation. Peer review is generally said to encourage critical examination, promote the exchange of ideas, reduce non-academic interference, guide academic discourse, and reinforce academic values (Berkencotter, 1995). In addition to its benefits, peer review has been criticized for suppressing innovation, promoting cliques, and providing irreproducible results (Harnard, 1982; Peters & Ceci, 1982; Rothwell & Martyn, 2000). Focusing on the benefits and shortcomings of peer review, it is important to draw connections between peer review and organizational learning where such connections exist.

An alignment may be identified between the constructs of organizational learning defined previously and the benefits and shortcomings of peer review. It is this alignment that permits for a better understanding of peer review, an improvement of its practice, and an improvement in the assessment of student learning. Peer review, like the single loop learning process of organizational learning, assumes the existence of norms by which a peer's work may be judged. Through critical examination, norms are used to compare a peer's work to accepted practices. If a peer's work deviates significantly from accepted norms, then an attempt to correct will likely occur. Harnard (1982) refers to this action of peer review as "selfcorrective", in the sense that experts in the disciple are maintaining the discipline's accepted norms. The same terminology is used by organizational learning scholar Morgan (1997) to describe single loop organizational learning. Peer review, as a form of organizational learning, uses norms to guide a self-corrective process.

Self-correction exposes peer review and single loop learning to a major criticism. Single loop learning is, by its very nature, a perpetuation of the norms of the organization. Like peer review, ideas that deviate from the norms of the organization are corrected. The perpetuation of norms can lead to the suppression of innovation in peer review and the inability to adapt and change in an organization. Up until this point, the discussion of organizational learning has been confined to the single loop learning process.

Discussion

In order to overcome the impediment to organizational learning that is created by the suppression of innovation, a new view of organizational learning must be adopted. Double loop learning encourages participants in organizational learning to challenge the norms that guide corrective action (Morgan, 1997). The cliché "thinking outside the box" is often used to describe the process of challenging existing norms. When faced with an idea or practice that deviates from existing norms, double loop learning encourages the learner to challenge the norms rather than immediately discard the innovation, as single loop learning would dictate. For example, a method of assessment that does not resemble the status quo may be discouraged by peer review that is guided by single loop learning. The introduction of double loop learning permits the reviewer to challenge the status quo and further explore the innovative technique through a dialogue with its creator.

Perhaps the strongest bond that exists between peer review and organizational learning is that which is exposed only when double loop learning is introduced. Organizational learning has the potential to promote the mutual sharing of knowledge (Argyris & Schön, 1996). Double loop learning furthers this sharing of knowledge by permitting for the creation of new knowledge. Peer review also seeks mutual benefit through the sharing of information. Peer reviewers and those receiving review can benefit from the exchange afforded by the peer review process. If the mutual benefits of peer review are to be realized, an iterative process must be instituted (Rubin, 1982). The peer review of plans to assess competency in critical thinking an oral communication fell short of a process that was truly iterative, in that reviewers were only afforded a one-time, one-way opportunity to address an institution's plans to assess competency. An iterative process would provide the opportunity to review, rebut, and revise in a cyclical method until a suitable finished product is reached.

The following suggestions are made in an effort to improve the peer review process as it applies to the assessment of student learning. First, double loop learning requires that the norms used to guide decision making and corrective action (i.e. the norms that guide an organization) must be continuously challenged (Morgan, 1997). In efforts to advance new ideas and promote innovation in student outcomes assessment, operating norms and assumptions must be confronted, ultimately resulting in their affirmation or their dismissal and replacement. Morgan suggests that organizations interested in fully developing double loop learning strategies (a) anticipate change; (b) develop capabilities for questioning operating norms; and (c) foster emergent organization.

Second, a balance must be struck between innovation and regulation. Steps must be taken to foster innovation and the free exchange of ideas, as is required in a community of practice, while still engaging in a process that is ultimately regulated by state code. Bureaucracies are widely criticized for stifling innovation. At the same time, a level of consistency and order must be achieved to comply with the intent of the policy that aims to provide substantive information on the quality of student learning. Goodsell (1994) notes that the ability to stabilize and provide predictability are among bureaucracy's greatest virtues. In all, forces for innovation must confront forces for stabilization and predictability.

Finally, in order to achieve an environment in which innovation is fostered and properly balanced, communication must be optimized (Berkencotter, 1995). Managers of information must act diligently and deliberately to establish networks that support collegial interaction. Organizers of peer review should consider how a truly iterative process may be implemented if the full benefits of peer review are to be realized. Further, members of the organization must be willing to engage in ongoing and substantive discussions on what constitutes good assessment. The original intent of using peer review in evaluating student competency assessment plans was to (a) provide expert feedback, and (b) encourage inter-institutional communication for the purpose of providing mutual benefit to reviewer and the reviewed while spurring creative approaches to assessment. Given that the plans reviewed have yet to be fully implemented, it is too soon to determine if both aims have been achieved. Evaluation of the first aim will require the completion of the assessment cycle to determine if the expertise of peer comments provided greater value than those of SCHEV staff. The second aim will also require the sort of reflection that is best accrued with time. Ultimately, the literature suggests that a double loop learning process will contribute to mutual learning and the use of innovative assessment techniques, if the proper conditions are established. Participants and organizers (i.e. the reviewed, reviewers, and editors) must engage in an ongoing discussion that seeks to clarify good assessment without suppressing innovation.

Author note: Craig Herndon serves as Associate for Academic Affairs and Research Policy Analyst for the State Council of Higher Education for Virginia. In addition, he is a doctoral candidate in Virginia Tech's Educational Policy and Leadership Studies program.

References

Argyris, C., & Schön, D. A. (1996). *Organizational learning II: Theory, method, and practice.* Reading, MA: Addison.

Berger, J. B. (2000). Organizational behavior at colleges and student outcomes: A new perspective on college impact. *The Review of Higher Education 23*(2), 177-198.

Berkencotter, K. (1995). The power and perils of peer review. *Rhetoric Review*, *13*(2), 245-248

Birnbaum, R. (1988). *How colleges work: The cybernetics of academic organization and leadership.* San Francisco: Jossey-Bass.

Bolman, L. G., & Deal, T. E. (2003). *Reframing organizations: Artistry, choice and leadership.* San Francisco: Jossey-Bass.

Carnegie Classification (2000). *Carnegie Foundation for the Advancement of Teaching.* Stanford, CA: Author

Code of Virginia § 23-9.6C1 (2005). *Duties of the State Council of Higher Education for Virginia.* Retrieved November 8, 2005, from http://leg1.state.va.us/cgibin/legp504.exe?000+cod+23-9.6C1.

Dierkes, M., Berthoin Antal, A., Child, J., & Nonaka, I. (2001). *Handbook of organizational learning and knowledge.* New York: Oxford University Press.

Epstein, D. (2005). Measuring the pulse of liberal education. *Inside Higher Ed.* Retrieved November 7, 2005, from http://www.insidehighered.com/news/2005/11/07/aacu.

Executive Order 1. *Creating a Blue Ribbon Commission to evaluate the needs and goals of higher education in Virginia in the 21st century.* Retrieved November 1, 2005, from http://www.epi.elps.vt.edu/Perspec tives/order.htm.

Goodsell, C. T. (1994). *The case for bureaucracy: A public administration polemic.* Chatham, NJ: Chatham House.

Governor's Blue Ribbon Commission on Higher Education. *Final Report of the Governor's Blue Ribbon Commission on Higher Education.* Richmond, VA: Author.

Harnard, S. (1982). *Peer commentary on peer review.* New York: Cambridge University Press.

Keig, L., & Waggoner, M. D. (1994). *Collaborative peer review: Role of faculty in improving college teaching.* Washington, DC: ERIC-ASHE.

Morgan, G. (1997). *Images of organizations.* Thousand Oaks, CA: Sage Publications.

Peters, D. P. & Ceci, S. J. (1982). Peer review practices or psychological journals: the fate of published articles, submitted again. In S. Harnard (Ed.), *Peer commentary on peer review* (pp. 12-24). New York: Cambridge University Press.

Rothwell, P. M., & Martyn, C. N. (2000). Reproducibility of peer review in clinical neuroscience. *Brain*, *123*, 1964-1969.

Rubin, D. R. (1982). Rejection, rebuttal, revision: some flexible features of peer review. In S. Harnard (Ed.), *Peer commentary on peer review* (pp. ??-??). New York: Cambridge University Press.

Shafritz, J. M. & Ott, J. S. (2001). Classics in organizational theory. Belmont, CA: Wadsworth.

Wenger, E. C. & Snyder, W. M. (2001). Communities of practice: the organizational frontier. In Harvard Business Review's *Organizational learning*, (pp 1-20). Cambridge, MA: Harvard Business School.

Writing Across the Curriculum Works: The Impact of Writing Emphasis upon Senior Exit Writing Samples

Dennis R. Ridley
Virginia Wesleyan College

Edward D. Smith
Longwood University

Abstract

Seniors' writing skills were assessed in 1998 at a medium-sized public university. Blind scoring, a standard scoring guide, and trained graders were used. Curricular writing emphasis was assessed through a syllabus study, yielding a Curricular Emphasis Score. Controlling for entry-level skill in writing, Writing Score and Curricular Emphasis were highly correlated.

Introduction

Writing across the curriculum is an emphasis that, like apple pie, enjoys widespread appeal. According to the MLA Commission on Writing and Literature (1985), 47% of 4-year colleges had writing across the curriculum programs, and that percentage continues to climb. The belief in its effectiveness remains strong in colleges and universities across the land. However, among writing professionals skeptics point out that after programs have been set up, assessment is often lax or non-existent. Convincing evidence sufficient to satisfy the wary researcher, policy-maker or administrator, controlling for relevant variables, is difficult to find. A partial catalogue of these relevant variables would include, at the top of the list, pre-existing writing ability as it can be estimated upon college entry. In addition, given the fact that increasing numbers of students attend more than one institution on their journey toward graduation, there is the factor of the writing emphasis found at multiple institutions. Since the experience gained earlier at another institution is beyond the control of a college or university, how does the latter institution determine what part of the writing proficiency of its graduates was contributed by its courses and what part by the prior experience?

Another major problem facing research comes from the nature of the curriculum. Accreditation standards require that virtually all institutions have curricula that foster writing proficiency along with other general education skills. Many institutions, including the one that is the object of this presentation, address this requirement through both required freshman writing courses and other required courses in the curriculum, sometimes designated as having a "writing emphasis." The problem that remains is that learning does not necessarily follow prescribed patterns laid down in the curriculum. Students may improve their writing in many ways through a myriad of course-taking patterns, including those without any official designation as "writing-intensive." What may be needed is a method that brings to bear an independent outside assessment of the writing-intensiveness of the curriculum. Further, such a method must allow for individualized measurement of the writing intensiveness for students' particular courses of study.

Method

Such a method was developed by the one of the authors, the late Dr. Edward D. Smith (see Note 1), and presented at a regional assessment conference. The method requires the collection and review of syllabi for all, or almost all, courses in the undergraduate curriculum. Each syllabus is rated by an independent rater on a 3-point scale to measure the degree of emphasis on a number of process variables. These variables included the following: written communication, oral communication, problem solving, computer applications, mathematical applications, international perspectives, and diverse perspectives. The 3-point scale defines three degrees: (a) no emphasis that the process variable was being address in the course (score = 0); (b) some emphasis that the process variable received some attention at some point (score = 1); and (c) strong emphasis that the process variable received emphasis throughout the semester (score = 2). For example, strong emphasis in written communication was defined as two or more assigned papers. The method has been used at several public institutions in Virginia including the object of the current presentation. In the latter, a sample of syllabi from six departments yielded satisfactory reliability estimated by correlations between two independent raters in the approximate range of r =.7 to .9.

Validity was also addressed. At a second institution, validity in the area of written communication was indicated in a special study connecting curricular writing intensiveness with seniors' writing proficiency. Writing intensiveness of the curricula in the institution's divisions (percent of courses with "strong emphasis") correlated significantly with pass-rates on the institution-wide senior writing test; i.e., the measure of writing intensiveness in those divisions tracked the proportion of students passing the test, having majors in those divisions. This result suggests that, for measuring course process variables, the syllabus method can connect such variables with important outcomes. Further, it supports an extension of the study to another institution in the particular area of written communication. Another study based in the first institution found evidence of validity of syllabus ratings as a measure of writing intensiveness. That study found a significant correlation ($r = .80, p < .01$) across departmental programs between syllabus ratings and referrals to the university's Writing Center.

The question of the stability of the syllabus method also was addressed. The syllabus study was repeated for the institution under study during two different years, 1996 and 1999. The same process variables named above were measured. The primary rater and investigator was the same person, an independent outside consultant. For each of the variables, chi squared analysis showed that the combined percent of "some emphasis" and "strong emphasis" did not differ significantly for any of the variables between the two years. For example, Written Communication showed a mean of 69% in 1996 and 66% in 1999.

In the current study, the institution is much smaller than the institution that was the focus of the previous study reported above. Therefore, students ($N = 71$) rather than divisions of the institution were the focus of study. It thus became necessary to devise an individual measure of the writing intensiveness of the individual's prior course of study. Since this phase of study was quite labor-intensive, a random sample of 25 students was drawn. The academic records of these students revealed the number of courses taken in each discipline. A weighted sum for each student, with weighting by the percent of syllabi that showed "strong emphasis" on writing in each discipline from the syllabus study, yielded a Curricular Emphasis Score.

For the writing outcome measure, since the university did not require a senior writing test it was necessary to recruit a sample ($N = 71$) of the graduating seniors to participate in a senior exit writing test. The testing was done in the spring of 1998. Similar writing sample data were collected from samples of seniors from 1995 ($N = 60$) and 1997 ($N = 93$). Students were selected from the list of prospective graduates, using a sampling plan designed to guarantee a good representation of subjects who transferred in freshman writing course credits and those who took those courses at the institution. With this one restriction, the sample was a random sample. Students were recruited by letter from the provost with a follow-up letter and telephone calls from the Director of Assessment. The letter invoked the catalogue graduation requirement to participate in various forms of assessment. It also offered a small stipend ($15) for all participants. Testing was conducted during the last two weeks of the term. Participation was nearly 90 percent of those students eligible and invited to participate.

The writing sample followed the same procedures and used the same test as was used by freshman writing students for the final of the spring term. For security purposes, it was necessary to schedule the testing immediately after the writing prompt was selected and just before finals week when freshmen were scheduled to take the examination. Grading also followed the same structured procedures as used for freshmen. A standard, structured Scoring Guide was used and all graders, teachers of freshman English, were trained for grading consistency. Six dimensions were rated on a 5-point scale. The scale points were as follows: 0 = failing, 1 = below average, 2 = average, 3 = above average, and 4 = superior. The dimensions of the scoring guide were the following: Summary of Reading, Critique of Reading, Personal Response to the Reading, Structure, Correctness, and Style. For this study, a Total Score based on all dimensions was also constructed to summarize performance. In addition, instructors assigned a holistic grade on the standard 4-point scale. Batches of papers were assigned randomly to readers, who were professors of English. These assignments were blind with respect to graders' knowledge of the special status as seniors of those being graded. Readers did not read their own students' papers, and all writers' identities were disguised. While some graders may have suspected, they were not intentionally informed that such a special status existed. Only the Director of the Freshman Writing Program, who assisted with the project, was aware of those assignments. Superficially, freshmen and senior papers looked identical.

For statistical purposes, other data were added to the dataset: the mean grades for the two freshman English courses, the student's cumulative GPA at the university, transfer GPA, overall GPA, and the SAT-

Verbal score. Due to the importance of examining pre-existing writing ability, the SAT-Verbal score was used as a surrogate although it is a fallible measure for that purpose. Support for this choice can be found in the results (below) where it will be seen that, of the variables in the study, the SAT-Verbal was most highly correlated with the Total Score or composite exit writing score.

Results

The syllabus study revealed that approximately two-thirds of courses at the university incorporate written communication, defined as the combined ratings of "some" and "strong" emphasis. About one-fourth showed a "strong" emphasis or incorporationof two or more writing assignments. For comparison, this degree of curricular emphasis was less than the problem-solving emphasis (> 80%), about equal to the oral communication emphasis, and greater than all the other dimensions of curricular emphasis. As already reported, results were highly consistent over a 3-year period. Senior exit writing results can also be compared on the same six dimensions plus the Total Score. These comparisons will not reveal the same degree of consistency over time as shown in the syllabus study.

Comparisons between freshman scores and senior scores were studied in depth in the 1997 study. While these were interesting and can be briefly reported, the focus of the current study was on the 1998 writing study (combined with the 1999 syllabus study), i.e., the relationships between the writing outcomes variables and the curricular process variables.

Correlations among all variables were examined with a view toward predicting skill in writing. For summary purposes, the Total Score will be used in this proposal in place of the other six writing outcome variables, of which it is the composite. Total Score correlated significantly with freshman English grades ($r = .43$), cumulative GPA ($r = .38$), and SAT-Verbal ($r = .57$). In addition, freshman English grades correlated ($r = .58$) with the cumulative GPA. All correlations were significant at the $p<.01$ level of confidence.

The role of transfer was also examined. One comparison was made between participants who transferred their freshman English credits from another institution and those who took their freshman English course at the university. The dependent variables in the comparison were the holistic grade received on the writing test and the Total Score. These differences were not statistically significant. A second related comparison looked at native students (students with no transfer credits) versus transfer students on the same two dependent variables. Again, these differences were not statistically significant. These results were consistent with those found in the 1997 writing study. For 25 students randomly selected to derive a Curricular Emphasis Score, the correlatiobetween that score and the Total Score on the writing test was $r = .40, p < .05$. To control for entry-level skill in writing a partial correlation was conducted on Curricular Emphasis and Total Score on the Writing Test, using SAT-Verbal scores as the controlled variable. The partial r was $.78, p < .001$.

Conclusions

Mean grades in freshman English courses, cumulative GPA, and SAT-Verbal scores all correlate significantly with an independent assessment of writing skills of seniors. There are no significant differences on this senior assessment of writing skills between native and transfer students or between students who had taken freshman English at the university or elsewhere. Curricular emphasis on writing correlated significantly with this senior assessment of writing skills. This correlation was even stronger after controlling for differences in entry-level writing skills as measured by SAT-Verbal scores. Thus, pre-existing writing ability (estimated by the SAT-Verbal) continues to be a strong influence on writing skills later in college, continuing through until the time of graduation. However, again there is a highly significant contribution of the curricular emphasis on writing that comes through strongly when pre-existing writing skill level is controlled. Certainly, there are flaws in the study; one could wish for a larger sample in the crucial test reported here. Nonetheless, the substantial correlationand level of significance are worthy of note.

Beyond the findings themselves, this study illustrates the use of two methods that have shown considerable promise. The first is the syllabus study method, which has been used with good results in at least four different institutions. While syllabi present only one window on the important process variables related to valued general education outcomes, these early studies are promising. They suggest that instructor's educational intents as stated in syllabi are stable and often may be valid indicators of the general knowledge and skills that students gain.

The second innovative method introduced here is a disciplined process of assessing seniors' writing skills. We used trained graders and a structured scoring guide with which graders have become comfortable and can use efficiently. We also made the process as blind as we could. The problem of assuring student motivation to perform well under these institutional circumstances remains difficult. However, compliance with the task was good and students' performance was reasonably in accord with their academic records as regards writing. In other circumstances, where real consequences are attached to performance, this approach might be even more successful.

The real message is that of the title of our presentation: "Writing Across the Curriculum Works." We believe our study has overcome at least some of the obstacles that stand in the way of making such a claim with confidence, if not certainty. Professors of English as well as many others, who believe in and care deeply about fostering writing ability in college, and who have labored toward this end for many years, can be encouraged by these results.

Notes

[1] An earlier version of this paper was presented at the meetings of the Association for the Study of Higher Education, Sacramento, CA, November 17, 2000. It was also presented at the 14th annual Virginia Assessment Group Conference, November 3, 2000. The first author particularly acknowledges a debt to Edward D. Smith, Ph.D., who was a well-known and respected member of the assessment community in Virginia, and a professor of psychology at Longwood University, for many years until his untimely death in 2003. Dr. Smith was the inspiration for the method used and its successful application in many contexts including the present case, in which he collaborated fully. The first author thanks Mrs. Sherry L. Smith for her gracious permission to include her late husband posthumously as honored co-author.

[2] Requests for additional information may be sent to Dennis R. Ridley, Institutional Research and Planning, Virginia Wesleyan College, Norfolk, Virginia 23502-5599.

Using Effect Size in NSSE Survey Reporting

Robert Springer
Elon University

Abstract

The National Survey of Student Engagement (NSSE) provides participating schools an Institutional Report that includes (among many documents) mean comparisons, frequency distributions, and student respondent data as part of its standard reporting package. Sifting through all this data can leave even experienced researchers wondering where to start and what to report. For example, how meaningful is it to report frequency percentages or statistically significant differences between your school and other NSSE schools? Fortunately, NSSE also provides an effect size (Cohen's *d*) or practical significance indicator that can help bring context to the results. In addition to its value in conveying NSSE results, using effect sizes in survey research helps to easily identify areas/items of praise as well as areas/items for improvements.

Introduction

Perhaps one of the most overlooked and more useful statistics is effect size. While statistical tests of significance indicate the likelihood that results would differ by chance (and are depend upon sample size), effect size measurements tell us the relative importance or magnitude of the treatment. As a result of the inability of statistical significance to indicate importance or practical significance (Kirk, 1996; Thompson, 1999; Valentine & Cooper, 2003), there is an ongoing debate as to the practical usefulness of statistical significance tests (Hunter 1997; Kirk, 1996;Thompson, 1999), particularly statistical significance tests used as a sole indicator of the meaningfulness of results.

In essence, effect sizes are practical significance/importance indicators (Kirk, 1996; Vacha-Haase & Nilsson, 1998; Valentine & Cooper, 2003). In a time where collecting large sample sizes has become relatively easy and affordable (i.e. web-based surveys), it is important to distinguish between statistical significance and practical significance. Consider the following example. Elon University had 950 first-year and senior students participate in 2005 NSSE. Of the 170 items on the NSSE (85 questions for both freshmen and seniors), 114 are statistically significant at the $p< 0.001$ level. So, which of these 114 items should be presented to stakeholders? While mapping survey items to institutional or department purpose is always good practice and would help to identify specific questions for reporting, what about other important items (perhaps equally important to another department/program) that might be slipping through the analysis? Short of presenting the entire NSSE results, which in all likelihood will never be read, what items can be identified that indicate meaningful or practical differences? As previously stated, the effect size indicates the relative importance or magnitude of the difference in scores between a treatment and a non-treatment (control) group. In multi-school surveys, it is reasonable to view your institution as the treatment group and other institutions as the non-treatment group. One reason an institution might participate in a national survey is to compare its results to that of other colleges – to see how they are doing by comparison. One way to evaluate the comparison is to use an effect size to indicate meaningful differences between colleges on particular items/areas?

While there are various types of effect size statistics (e.g., ω^2, adjusted R^2, Hedge's g, Fisher's Z, Glass's Δ, η^2), Cohen's *d* will be the focus of this paper, since it is supplied by NSSE . Cohen's *d* as a reported effect size is becoming very popular (Thalheimer & Cook, 2002). As a result, more and more research is including Cohen's *d* which allows for easier comparisons of the magnitude of treatments across experiments (Thalheimer & Cook, 2002). Besides the advantage of its popularity, this effect size also has theadvantage of allowing comparisons to known benchmarks established by Cohen. He describes a d-value of 0.20 as small, 0.50 as medium (moderate), and 0.80 as large. Adding some perspective to these effect sizes, he states that a moderate effect size is "visible to the naked eye" (Cohen, 1988, p.26).

Purpose

This paper is intended as a best practices presentation by demonstrating the use of effect sizes to assist in reporting. This statistic can quickly identify items of practical significance, which adds to interpretation of results.

NSSE and Cohen's d in Reporting

The National Survey of Student Engagement provides Cohen's *d* in their standard Institutional Report under the Means Comparison results. Thus, any institution that participates in the NSSE will have this statistic provided as part of the standard reporting documentation (in paper and electronic form). Of particular interest is the electronic spreadsheet of the NSSE results. This allows for easy sorting of items by Cohen's *d*.

Effect sizes can be negative. For this to happen, the treatment group is performing at a lesser level than the control group. However, the negative sign could be function of scale direction rather than a perceived lack of performance. For example, in the NSSE question about coming to class unprepared (2005 NSSE item 1f), a negative sign is preferable - meaning that fewer students are coming to class unprepared.

While effect sizes may change from one survey administration to the next, they tend to remain fairly stable. In other words, if the institution has not changed what it is doing with respect to certain items, effect sizes, in all likelihood, will not change from one level to another (small, moderate, or large).

Effect sizes are not new to statistics. Effect sizes can be traced back to at least 1901 with the work of Karl Pearson (Kirk, 1996). Yet as such, reporting effect sizes to stakeholders may not be desirable. Doing so might be confusing and could easily lead to dismissal of the report. Reporting figures that are more widely understood such as percent positive frequencies is more advisable. Stakeholders will understand percent positive frequencies. Elon University uses a percent positive frequency (for example, the number of students that select Very Often or Often for a series of questions). The percent positive scale used in Elon's reporting is believed to be fair and it appears to make sense to various stakeholders.

If survey results are to be used to help make improvements at an institution, it is good practice to identify items where the school does well (areas for praise and celebration) and items where it does not do well (areas for improvement) as compared to other schools. For Elon, items of a practical significance are those that approach or exceed a moderate effect size level (d > 0.40). A Cohen's *d* of at least 0.40 is approximately two-thirds the distance between small and moderate levels. As a result, *d* values of at least 0.40 are considered approaching a moderate level.

How do effect size and percent positive frequencies relate to each other? Consider the following example. The criterion of an effect size of 0.40 or higher to report items is applied. One item where first-year student responses met that criterion was item 1h – *worked with classmates outside of class to prepare class assignments*. Its effect size is 0.41, it is statistically significant at the $p<.001$ level, and the percent positive frequency is 63% versus 43% for all NSSE schools. To say that Elon is noticeably different compared to other colleges with respect to this item would probably make sense to the lay person, because they can see the large gap (a 20-point difference) in the percent positive frequencies. In addition, the effect size supports that statement. As a further example, another item where first-year student responses resulted in an area targeted for improvement is item 5, *the extent your examinations during the current school year challenged you to do your best work.* Its effect size was 0.03 (very small), it is not statistically significant, and the percent positive frequency is 54% versus 52% for all NSSE schools. The differences in percent positive frequencies, as well as the small effect size, indicate little if any practical difference exists between the two comparisons. Elon wants academic challenge and rigor to be a hallmark for distinction. Given these results, its first-year students appear to be no more challenged than other schools first year students.

How did we actually use the effect size for reporting purposes? A supplemental two-page report interpreting the NSSE results is sent to senior staff and then to the faculty. The first page of that report provides basic information about NSSE and describes the three tables on the second page. In addition, a short paragraph explains that effect sizes are used to select the items for inclusion into the tables. Since Elon participates in the NSSE each year and in order to address stakeholders possible concerns about effect sizes shifting from year-to-year, we selected items that were extremely consistent with respect to reported effect sizes for a five to six year period depending upon when the item was introduced (that being an effect size equal to or greater than 0.40, or near zero). Table 1 indicates items for first year students that have effect sizes of at least a 0.40 (i.e., high performing items). Table 2 indicates items for senior students that have effect sizes of at least 0.40 (i.e., high performing items). Table 3 indicates items whose effect sizes are at or near zero (low performing items). Each table provides the survey questions and the percent positive frequencies for Elon and all other NSSE schools – effect sizes are not presented (effect sizes are included in the appendices as a point of reference for the reader). The items presented for improvement represent areas that are important for Elon to achieve in its strategic plan. In general, this two-page report is very effective in convening what Elon does very well (based upon effect size) and what it needs to do better (based upon effect size and areas deemed important by the institution).

Discussion

We hope that readers will see the adaptability in using effect sizes to assist in reporting. For example, while a Cohen's *d* of 0.50 is considered moderate and 0.20 is considered small, we selected effect sizes of 0.40 or higher and those that were near zero. For items where Elon performed well, we simply selected items with effect sizes of 0.40 or higher. Items selected for improvement were also items that are identified as important to Elon. Interesting was the fact that the high performing items tended to confirm institutional belief. This also added support for acceptance of the low performing items.

While not all reporting of survey data should or can be reported using effect sizes, it should be obvious that having NSSE supply effect sizes as part of its standard reporting packages enables a researcher to quickly sort and analyze practical differences between itself and comparison groups. Creating a two-page report is acceptable and desirable at Elon - this may not be the case at other institutions.

National surveys that do not provide an effect size or the statistics necessary to calculate one, would be aiding institutions by adopting such standards in their reports. This would allow true peer/aspirant comparisons on a number of dimensions from students and faculty.

Finally, effect size has much broader implications that just survey data. Many publishers are requesting effect size(s) with interpretation(s) from researchers. As the popularity of reporting effect sizes continues to grow, researchers should take caution intheir interpretations of effects sizes being reported as small, moderate, or large. In other words, let the context of the research help establish typical effect sizes and, therefore, what is worth reporting.

References

Cohen, J. (1988). *Statistical power and analysis for the behavioral sciences* (2nd ed.) Hilldale, NJ: Lawrence Erlbaum Associates.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8,* 3-7.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Thalheimer, W. & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology.* Retrieved November 1, 2005 from http://work-learning.com/effect_sizes.htm.

Thompson, B. (1999). Why 'encouraging' effect size reporting is not working: The etiology of researcher resistance to changing practices. *The Journal of Psychology*, *133*(2), 133-140.

Vacha-Haase, T. & Nilsson, J. (1998). Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling & Development*, *31*(1), 46-57.

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes.* Washington, DC: What Works Clearinghouse.

Appendix

Table 1

Items which Distinguish Elon University Freshmen from Those at All NSSE Schools (2005)

Item	Elon % Positive	All NSSE Schools % Positive	*d*
Make significantly more class presentations	52	33	0.42
Work more frequently with classmates outside of class	63	43	0.41
Worked w/faculty members on activities other than course work (committees, orientation, student life activities, …)	26	14	0.45
Write significantly more short papers (5 pages or less)	73	39	0.64
Write significantly papers/reports between 5 -19 pages	26	11	0.78
More frequently attend exhibits, galleries, plays, or dances	50	29	0.47
Are more likely to have participated in a learning community than their peers	48	36	0.41
More frequently attend campus events (athletics, special speakers, cultural)	90	56	0.54
Feel support to thrive socially	61	43	0.43
Cognitive/behavioral development: Students contributing to the welfare of their community	67	46	0.45
Are more satisfied about their educational experience	97	87	0.44

Table 2

Items which Distinguish Elon University Seniors from Those at All NSSE Schools (2005)

Item	Elon % Positive	All NSSE Schools % Positive	*d*
Make significantly more class presentations	86	64	0.49
Work more frequently with classmates outside of class	83	59	0.52
Use e-mail more frequently to communicate with their instructors	92	83	0.42
Worked w/faculty members on activities other than course work (committees, orientation, student life activities, …)	42	26	0.50
Write significantly papers/reports between 5 -19 pages	62	38	0.48
Are more likely to have participated in practicum, internship, co-op, field experience, …	91	77	0.58
Are more likely to perform community/volunteer service	93	76	0.52
Are more likely to have studied abroad	73	25	1.35
Are more likely to have a culminating senior experience (capstone, thesis, project, comprehensive exam)	89	65	0.68
Attend campus events (special speakers, cultural performances, athletic events)	82	56	0.62
Are more satisfied about their educational experience	97	88	0.41

Table 3

Items that Do Not Distinguish Elon University Students from Other NSSE Schools (2005)

Item	Class	Elon % Positive	All NSSE Schools % Positive	*d*
To what extent have your exams challenged you?	Freshmen	54%	52%	0.03
	Seniors	50%	53%	-0.04
Course work emphasizes making judgments about the value of information, arguments, or methods	Seniors	78%	72%	0.11