RESEARCH & PRACTICE IN ASSESSMENT

VOLUME SEVEN | WINTER 2012
www.RPAjournal.com
ISSN # 2161-4210

RPA

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

The goal of *Research & Practice in Assessment* is to serve the assessment community as an online journal focusing on higher education assessment. It is dedicated to the advancement of scholarly discussion amongst researchers and practitioners in this evolving field. The journal originated from the Board of the Virginia Assessment Group, one of the oldest continuing professional higher education assessment organizations in the United States. *Research & Practice in Assessment* is a peer-reviewed publication that uses a double-blind review process. Approximately fifty percent of submissions are accepted for issues that are published twice annually. *Research & Practice in Assessment* is listed in Cabell's Directory and will be indexed by ProQuest.

TABLE OF CONTENTS

CALL FOR PAPERS

Research & Practice in Assessment is currently soliciting articles and reviews for its Summer 2013 issue. Manuscripts submitted to *RPA* may be related to various higher education assessment themes, and should adopt either an assessment measurement or an assessment policy/foundations framework. Contributions to be evaluated by the *RPA* Review Board should be submitted no later than April 1. Manuscripts must comply with the *RPA* Submission Guidelines and be sent electronically to:

rpaeditor@virginiaassessment.org

Our journal is available for free download for iPad!
www.RPAjournal.com

Published by:
VIRGINIA ASSESSMENT GROUP
www.virginiaassessment.org

For questions about Virginia Assessment Group membership opportunities email **webmaster@virginiaassessment.org**. We welcome members and non-members to join our community. If you would like to be informed of Virginia Assessment Group events, please contact us and we will add you to our distribution list.

Publication Design by Patrice Brown
www.treestardesign.com

FROM THE EDITOR

Bringing in the Disciplines

Earlier this year, in a subtle yet bold move, *Research & Practice in Assessment* was restructured in order to bring diverse academic disciplines into the field of assessment. This intentional philosophical shift was meant to advance the assessment discourse in two respects. *First,* it is an attempt to reduce the uncoupled nature of assessment scholarship by linking seemingly independent perspectives. This shift emphasizes the shared areas of knowledge between fields by bringing relevant work into the profession that may have only appeared within a given discipline. *Second*, while the assessment literature has become exceptionally strong with regard to measurement, this shift attempts to further strengthen other aspects of the literature such as its theory and philosophy. The latter are essential features of educational scholarship in the social sciences. Consequently, as psychology has played a significant role in strengthening the measurement aspects of higher education assessment, disciplines such as sociology, history, political science, economics, anthropology, and others offer paradigms that will strengthen and broaden higher education assessment theory and philosophy.

In this vein, the Winter Issue of *RPA* opens by "bringing in" a special feature penned by two economists and a social demographer that engages the measurement of productivity in higher education. Written by William Massy, Teresa Sullivan, and Christopher Mackie, the piece highlights the collaborative efforts of the panel commissioned by the National Research Council. Two peer review articles are presented that advance conceptual issues of measurement for authentic assessments. John Hathcoat and Jeremy Penn provide a framework for conceptualizing measurement error when using authentic assessments and investigate the extent to which student writing performance may generalize across multiple tasks. Then, using a common assignment and combined rubric, Rana Khan, Datta Kaur Khalsa, Kathryn Klose, and Yan Zhang Cooksey present a model to assess graduate student learning in four competencies.

In the reviews, Linda Sax comments on *Assessment for Excellence*, the recently revised edition by Alexander Astin and anthony lising antonio. In light of the dominant discourse on innovation in higher education, Lisa Hatfield roots assessment professionals in our context with a review of *Reinventing Higher Education*. The conclusion of the issue is a deliberate contrast to its opening feature article. Here, in a piece entitled "Discovery", the fine art photography of Adam Barnes is combined with an excerpt from Maria Montessori's *The Discovery of the Child*.

As you engage the pieces herein, consider how your own disciplinary paradigm may be "brought in" to advance the higher education assessment discourse. There is value in variation and I hope you will consider penning a unique scholarly piece for submission to *Research & Practice in Assessment.*

Regards,

Joshua T. Brown

Liberty University

Abstract

William Massy is an emeritus professor and former Vice President for Business and Finance at Stanford University, Teresa Sullivan is President of the University of Virginia, and Christopher Mackie is a Study Director with the National Academies' Committee on National Statistics. This article summarizes the authors' recent National Research Council report, *Improving Measurement of Productivity in Higher Education,* which reviews the principles and pitfalls of measuring university productivity and proposes a practical method for doing so at the sector and institutional segment levels. The summary emphasizes the method's data requirements and describes needed changes in IPEDS and other databases.

AUTHORS

William F. Massy, Ph.D.
Stanford University

Teresa A. Sullivan, Ph.D.
University of Virginia

Christopher D. Mackie, Ph.D.
National Academies' Committee on National Statistics

CORRESPONDENCE

Email
wfmassy@comcast.net

Data Needed for Improving Productivity Measurement in Higher Education

Recognizing that higher education is a critical element of the American economy, The National Research Council of the National Academies, with support from the Lumina Foundation, convened a panel on measuring higher education productivity (NRC, 2012a). The panel members are listed in the sidebar (see next page). All are recognized experts in higher education and/or productivity analysis. This paper provides a brief summary of the panel's conclusions and, particularly, their implications for the Integrated Postsecondary Education Data System (IPEDS).

The panel's charge was to develop a practical approach for developing aggregate measures to track productivity for broad groups of institutions and for the sector as a whole. We were not asked to address the improvement of productivity itself and, likewise, research and public service productivity were outside our purview. We have proposed a conceptual structure for higher education productivity measurement. We also have documented the many difficulties and caveats associated with the use of available measurement tools.

Four touchstones guided the panel's thinking about the importance of productivity improvement and measurement:

- "Productivity should be a central part of the higher education conversation.
- Conversations about the sector's performance will lack coherence without a well-vetted and agreed-upon set of metrics.
- Quality should always be a core part of the productivity conversations, even if it cannot be fully captured by the metrics.
- The inevitable presence of difficult-to-quantify elements in a measure should not become an excuse to ignore these elements" (NRC, 2012a, p. 2).

An additional requirement was that our proposed productivity measures be derived as much as possible from the discipline of economics. Productivity is defined as the ratio of outputs to the inputs required for producing them, where both inputs and outputs are adjusted for quality differences. The panel felt strongly that ad hoc measures not related to economic science, such as graduation rates and time-to or cost-of degree statistics, are incomplete and likely to be misleading when used in isolation.

Key Issues and Their Resolution

The economic definition of productivity is, fundamentally, the relation between the physical quantities of outputs and the physical quantities of inputs. It is more an engineering concept than a financial one. Financial concepts, which involve prices as well as quantities, do not enter the picture except as weights in the aggregation of disparate quantity variables (an example will be given later). The distinction matters because policies aimed at productivity improvement must address what essentially are engineering issues, which often are lost in people's concerns about financial matters. For example, a higher cost per degree that is caused by escalating labor prices (wages and benefits) does not imply reduced productivity, whereas an increase in the *amount* of labor utilized would do so.

Measures that describe either inputs or outputs, but not both, give an incomplete picture of productivity. This means that familiar statistics such as numbers of enrollments or credit hours, degree production, and cost or "profit" per faculty member should not be used to assess productivity. A final consideration is whether to evaluate "single-factor productivity" (e.g., output per labor hour) or "multifactor productivity" (output related to total resource usage). Colleges and universities have historically been labor intensive, but this has changed in recent years because of information technology, increasingly complex facilities, and outsourcing of support services (for which the labor component, while potentially significant, does not show up on the university's books). Therefore, the panel chose to measure multifactor productivity as defined by the model described later.

Educational quality is the "elephant in the room" in most discussions of higher education productivity. The economic theory is clear: both input and output quantities should be adjusted for variations in quality. For example, because of changes in technology, today's cars and computers are not directly comparable to those produced a decade ago; thus, direct price comparisons (without quality adjustment) are not meaningful. Therefore, when measuring price and productivity changes in these sectors, economists use techniques that account for changes in input and output characteristics from one period to the next.

Productivity measurement outside higher education, in competitive markets, relies on one of two devices to police quality. First, other things equal, better quality usually commands higher prices (the more expensive car or computer is usually the better one). It turns out, however, that this approach does not work for higher education. The prevalence of government subsidies and regulation, coupled with a dearth of well defined and accurate market information about education quality (particularly as it pertains to learning), make it unwise to assume that either the tuition rates or financial aid awards of colleges and universities are determined by competitive market forces.

The other method is to measure quality through special studies – for example, to track differences in the speed and memory capacity of computers – and use the resulting measures to adjust the quantity variables. The panel looked carefully at the prospects for developing the kinds of comprehensive learning quality measures needed to make such adjustments. We would have liked nothing better than to propose such measures but, unfortunately, we were forced to conclude that this will not be possible anytime soon. Our report cites a great deal of good work in the area, which definitely should be continued; but, while current and prospective learning and engagement measures are useful in particular contexts, they cannot be brought together into comprehensive, robust, indices for quality adjustment.

All is not lost, however. In the United States, a variety of external quality assurance procedures are deployed, such as regional accreditation, subject-specific accreditation, and in some fields, stringent licensure requirements. To the extent these work as designed (and they should be made to work regardless of whether productivity is measured or not), they put a floor under the output quality of education. Institutions also employ a variety of internal quality assurance procedures based, for example, on faculty governance. When combined with the external procedures, these can be expected to deter any "race to the bottom" that might result from measuring the quantitative aspects of higher education production. While it is true that high-end quality differences will not be reflected in quantity-based productivity statistics, at least the downside dangers can be mitigated. We hope that, in due course, better and more comprehensive quality measures will shed light on how learning varies across segments and changes over time. In the meantime, subjective judgments can be used to interpret the quantitative productivity statistics in light of more fragmented evidence about output quality.

The panel focused on instruction rather than on research and public service—even though the latter are central to the mission of a large subset of institutions. Including the research mission would have carried us into territory already being considered by another National Academies panel (NRC, 2012b) and, in any case, it would have added huge complexities to the ones already confronting us. But, while the panel did not address research and public service productivity, we did carefully consider their impact on the measurement of educational productivity. Inputs must be parsed into their instruction and research/public service components. The parsing is mostly a straightforward application of cost accounting but, as mentioned later, we do propose a new approach for handling the vexing issue of departmental research.

The economic definition of productivity is, fundamentally, the relation between the physical quantities of outputs and the physical quantities of inputs. It is more an engineering concept than a financial one… The distinction matters because policies aimed at productivity improvement must address what essentially are engineering issues, which often are lost in people's concerns about financial matters.

Another issue is the importance of avoiding spurious comparisons among institutions with dramatically different characteristics and missions. Among other things, it is essential to take into account incoming student ability and preparation. For example, highly selective institutions typically have higher completion rates than open-access institutions. This may reflect more on the prior learning, preparation, and motivation of the entrants than on the productivity of the institutions they enter—which means institutional performance should be gauged in terms of value added, not the absolute quality of graduates. Therefore, for the purpose of generating performance statistics, institutions should be segmented into reasonably homogeneous categories – for example, as used in the Carnegie classification system and the Delta Cost Project (2009).

The list of measurement issues for the sector would not be complete without consideration of data availability. The panel adopted a two-pronged approach. Our "Base Model," described in the next section, relies almost entirely on current IPEDS and other government datasets. The "enhanced model" that follows requires additions to IPEDS. As explained later, we believe these additions will be worthwhile for their own sake as well as to improve productivity measurement. The paper ends with a brief description of the other data-related recommendations in the panel's report.

Base Productivity Measurement Model

The panel's conceptual framework employs a multifactor productivity index, the so-called "Törnqvist index." The Bureau of Labor Statistics (BLS) and the Organization for Economic Cooperation and Development (OECD) use this index to measure productivity in a variety of economic sectors (BLS 2007; OECD 2001, which includes references to the background literature). We first describe the index in general terms, then define the output and input variables, and, finally, illustrate the calculations with a numerical example.

The productivity index, as evaluated for time increment Δt, is:

$$\textit{Productivity index}\ [\Delta t] = \textit{Output index}\ [\Delta t] \div \textit{Input index}\ [\Delta t].$$

The input and output indices represent changes in the physical quantities over the time

increment Δt (e.g., from 2010 to 2011). In other words, the Törnqvist index defines *Productivity* as the change in outputs obtainable from the input changes observed over Δt. *Productivity change*, in turn is calculated from the ratio of successive productivity indices:

$$\textit{Productivity change}\ [\Delta t_1 \text{ to } \Delta t_2] = \textit{Productivity index}\ [\Delta t_2] \div \textit{Productivity index}\ [\Delta t_1] - 1.$$

Colleges and universities have historically been labor intensive but this has changed in recent years because of information technology, increasingly complex facilities, and outsourcing of support services.

These definitions are consistent with the conceptualization of productivity as an engineering concept. Productivity is the slope of the "production function"—the curve relating outputs to inputs. Looking at the slope of the function rather than the function itself amounts to a kind of "what if" analysis: what happens to outputs if the inputs change by a certain amount?

Outputs. We recommend a simple yet comprehensive measure for output quantity. It is based on two IPEDS variables that, in the panel's words, "are the standard unit measures of instruction in American higher education."

- *Credit hours*: 12-month instructional activity credit hours summed over student levels (e.g., undergraduates, first professional students, and graduate students);
- *Completions:* awards or degrees conferred, summed over programs, student levels, race or ethnicity, and gender (NRC, 2012a, p. 65).

The importance of completions is obvious, but a measure based only on completions would ignore the learning that takes place on a course-by-course basis. The panel's recommendation, therefore, is that the base definition of educational output be "Adjusted credit hours" (ACH), defined as follows:

$$\textit{Adjusted credit hours} = \textit{Credit hours} + (\textit{Sheepskin effect} \times \textit{Completions}).$$

Again to quote the panel, "The 'sheepskin effect' represents the additional value that credit hours have when they are accumulated and organized into a completed degree. Based on studies of the effect of earned credits and degrees on salaries, the panel believes that a value equal to a year's worth of credits is a reasonable figure to use as a placeholder for undergraduate degrees. Additional research will be needed to determine the appropriate weight for the sheepskin effect for graduate and 1st professional programs" (NRC, 2012a, p. 66). The same ideas apply to many community college programs.

Inputs. The model's inputs, defined in Table 1, are the quantity of labor, the amount of non-labor expenditure, and the rental value of capital (the depreciation of plant and equipment during use) utilized in the educational process. The data for each input consists of (a) physical quantities or surrogates for quantities, as required by the fundamental definition of productivity; and (b) nominal expenditures, which are used to combine the several inputs into a single index.

Labor FTEs is a direct physical quantity. No such physical quantity can be found for *Expenditures on intermediate inputs* or *Capital*, so for them it is necessary to use deflated dollars as surrogates. (The dollar figures sum up the myriad of individual items included in the definition for each variable, weighted by the items' unit costs.) To summarize, the model uses three input variables (labor, intermediate inputs, and capital), each of which is represented by a physical quantity (or surrogate) and nominal expenditures.

As stated earlier, the input variables must reflect instruction rather than the whole range of institutional activity. The Delta Cost Project (2009) and OMB Circular A21 provide the needed methodology. The steps, which must be performed for each variable in Table 1, are to: (1) separate total expenditures into their direct and indirect (administrative and support) components; (2) further separate the direct component into instruction, research, and public service; (3) allocate the indirect component according to the fractions of direct expenditures; and (4) add the indirect allocation for instruction to the direct instruction variable to produce an overall figure for instruction.

Table 1

Input Variables Used in the Proposed Model

a) *Expenditures on Labor* (LE): nominal value of salaries and wages plus fringe benefits, used as the weight of L when aggregating the input.
b) *Labor* (L): the quantity measure for labor input, approximated by full-time equivalent (FTE) employees. Both academic and non-academic employees are included in the calculation. FTE figures are calculated from total full- and part-time employees, with a part-time employee counting as one third of a full-time employee, as assigned in IPEDS (this, too, could be adjusted with empirical justification). Labor is the biggest input into higher education instruction.
c) *Expenditures on Intermediate Inputs* (IE): Nominal cost of materials and other inputs acquired through purchasing, outsourcing, etc. (the sum of the IPEDS 'operations & maintenance' (O&M) and 'all other' categories). These nominal values are used in calculating I weights.
d) *Intermediate Inputs* (I): Deflated nominal expenditures (IE) are used to represent the physical quantities.
e) *Expenditures on Capital* (KE): opportunity cost for the use of physical capital; also called rental value of capital. Expenditures equal the IPEDS book value of capital stock times an estimated national rate of return on assets, where book value capital stock equals the sum of land, buildings, and equipment. Overall, the book value reported in IPEDS is likely too low; however, it does include buildings that may not be specifically allocated to teaching, which offsets the total to an unknown degree. These nominal capital values are used in calculating the weighting for capital.
f) *Capital* (K): For the quantity of capital input, the book value is deflated by the Bureau of Economic Analysis's investment deflator for gross private domestic investment.

The deflators for intermediate expenditures and capital are, respectively, the Producer Price Index and the index for gross private domestic investment. Neither these figures nor the national rate of return on assets can be obtained from IPEDS, but they are available from other government sources.

Calculating the index. Table 2 illustrates the index's calculation. The table reflects three years of data for one institution, which allows representation of two Δt increments. The time periods need not reflect a single year and the intervals between periods need not be the same for all Δt increments. (Scaling adjustments are needed if the increments vary, however.) Given the period-to-period variability inherent in IPEDS data, for example, it may be desirable to consider each period as the average of, say, three to five years. We noted earlier that the productivity index should not be applied to single institutions, but rather to broad groupings of institutions (the "segments" discussed above). The availability of institution-specific IPEDS data makes this relatively easy, and the aggregation process further mitigates the data variability problem.

Other things equal, better quality usually commands higher prices (the more expensive car or computer is usually the better one). It turns out, however, that this approach does not work for higher education.

The output calculations appear at the top of the table. Credit hours and completions are extracted directly from IPEDS. Adjusted credit hours (ACH) are calculated using a sheepskin effect of 30 (one year's normal student load). Quantity change is obtained by dividing each ACH figure by the preceding one: e.g., 1.008 = 643,477 ÷ 638,435. This also equals the output index because enrollments and completions already have been combined into a single number by applying the sheepskin effect.

Table 2

Base Model Productivity Calculations

	Period 1	Period 2	Period 3
Output			
Credit hours	573,815	574,176	602,000
Completions	2,154	2,310	2,500
Adjusted credit hours (ACH)	638,435	643,476	677,000
Quantity change (= output index)		1.008	1.052
Inputs			
Quantities			
Labor FTEs (L)	3,926	4,296	4,705
Intermediate expenditures (I)	$324,680	$486,147	$643,599
Rental value of capital (K)	$261,834	$267,507	$348,033
Expenditures			
Wages and fringe benefits (LE)	$756,399	$867,311	$1,036,594
Intermediate expenditures (EI)	$324,680	$550,921	$777,193
Rental value of capital (KE)	$261,834	$301,954	$400,791
Total	$1,342,913	$1,720,186	$2,214,578
Quantity change			
Wages and fringe benefits		1.094	1.095
Intermediate expenditures		1.497	1.324
Rental value of capital		1.022	1.301
Expenditure weightings			
Labor FTEs		53.4%	48.6%
Intermediate expenditures		28.1%	33.6%
Rental value of capital		18.5%	17.8%
Total		100.0%	100.0%
Input index		1.180	1.204
Multifactor productivity			
Productivity index		0.854	0.874
Change in productivity		2.3%	

The "sheepskin effect" represents the additional value the credit hours have when they are accumulated and organized into a completed degree.

The situation is more complicated for inputs. First, as described earlier, the IPEDS quantity and expenditure variables must be adjusted to account for research and public service. The resulting quantity change figures are ratios of the successive quantity figures (e.g., 1.094 = 486,147 ÷ 324,680), but then they must be combined to produce a composite input index. This is accomplished using a weighted geometric average with the relative expenditure figures (also shown in the table) as weights. For example, 1.180 is the geometric average of 1.094, 1.497, and 1.022. (The weights for Intermediate expenditures and Capital in Period 1 are the same as the quantities, but they differ in later years because of the inflation adjustment.) The choice of a geometric average follows from the fact that ratios are being averaged, and also from the mathematical derivation of the Törnqvist index.

Calculation of the multifactor productivity index appears at the bottom of the table. The index itself is the ratio of the output index to the input index: e.g., 1.008 ÷1.180. The change in productivity is the ratio of the successive productivity indices minus one: in this case the shift from 0.847 to 0.854 is a 2.3% change.

The Törnqvist index has some very desirable properties. In particular, researchers have shown that, under fairly general conditions, this calculation makes the best possible use (in terms of productivity measurement) of the information embedded in the input and output variables. Among other things, it washes out extraneous financial factors like the effects of substituting one resource for another because of price changes. (Substituting computers for people in a production process simply because the former have become relatively cheaper does not represent a productivity increase, for example, whereas making the substitution because the computers are being used more effectively does.) Alternative methods of calculation—for example, weighting the input changes by something other than

nominal expenditures or using an arithmetic average instead of a geometric one—have been shown to produce inferior results. The panel recommended that a task force be set up to work on operationalizing our conceptual structure, and that it begin with consideration of the Törnqvist index.

The panel's charge called for addressing productivity measures at different levels of aggregation including the institution, system, and sector levels. Our proposed model is designed to operate at the sector or subsector (segment) level. While we feel obligated to raise the possibility of single-institution and state-specific indices (which are possible given that IPEDS provides institution-specific data), we do not want to invite use of the model for accountability purposes. To do so would produce malevolent incentives–and, possibly, a race to the bottom in terms of education quality. Moreover, state-level aggregations would necessarily obliterate mission-distinctions that are typically important within each state.

Enhancements to the Base Model

While the panel believes the base model to be a viable approach for tracking college and university productivity, we have identified two enhancements that, while requiring data collection beyond the current IPEDS structure, would add significantly to the model's power.

We do not want to invite use of the model for accountability purposes. To do so would produce malevolent incentives – and, possibly, a race to the bottom in terms of education quality.

Differentiating labor categories. The first enhancement is to track key labor categories, in our case academic versus non-academic and regular versus casual employees, separately from total FTEs. The panel's reasons for suggesting this differentiation are as follows (see NRC, 2012a, p. 74):

1. One of the critical assumptions of the conventional productivity model is not viable in higher education. The typical productivity study assumes that, because labor is procured in competitive markets, relative compensation approximates relative marginal products. There is, in such a situation, no need to differentiate labor categories. Unfortunately, tenure-track faculty labor may not be linked tightly to marginal product in education because such faculty often are valued for research and reputational reasons, or be protected by or locked into institutions by tenure. Similarly, many so-called "contingent" faculty (hired on a course-by-course basis, often without fringe benefits) are desired as an alternative to tenure rather than because of judgments about their marginal product.

2. Another assumption is that the market effectively polices output quality. This is manifestly not the case for higher education. Colleges pursue strategies—larger classes or less costly instructors, for example—that reduce cost per nominal output but which will dilute quality when taken to extremes. In the example above, it may be attractive to employ less expensive, but also less qualified, personnel who are not well integrated into institutional quality processes. The panel is concerned lest the measurement of productivity add to the already problematic incentives to emphasize quantity over quality in higher education.

3. Academic staff play a unique and critically important role in most institutions. They, and only they, can make the intellectual judgments needed to create new knowledge, decide curricular and pedagogical issues, and assure educational quality. It is true that the distinction between teaching and nonteaching staff blurs as information technology shifts the modalities of teaching and learning. In some institutions, for example, faculty time is leveraged by modern learning software, a change that may require entirely new kinds of labor inputs. Such technology-driven changes are not unique to higher education, but the pace of change seems unusually brisk at the present time. Yet the critical task of intellectual leadership remains with faculty. Singling them out as a separate labor category recognizes that, at root, this kind of "labor" is not truly substitutable.

4. Productivity statistics are more likely to weigh heavily in policy debates on higher education than in policy debates for other industries. The U. S. public policy environment includes a significant oversight and accountability component that

requires information about productivity. Therefore, it is important that the statistics be as complete as possible on the important issues, including those associated with labor substitution.

Differentiating labor categories will require IPEDS be modified to include the following three-way classification scheme.

1. *Regular faculty FTEs*: those on the tenure line or equivalent, whether full or part-time.

2. *Part-time teachers who are hired on a course-by-course basis:* may be measured in terms of FTEs, number of course assignments, or some other metric–perhaps related to the current "Part-time" and "Primarily-instruction" ("PT/PI") variable.

3. *All other FTEs*: total FTEs excluding the above.

Importantly, the new scheme will need to support allocations among education, research, and public service as described earlier.

The panel was able to approximate the three-way differentiation from current IPEDS data (see NRC, 2012a, p. 77), but we believe it will be worthwhile to produce the data directly. (Development of the methodology should involve consultation with the providers and users of IPEDS data.) In addition to their value for productivity measurement, these data will prove useful to developers of benchmarking statistics and to others in the higher education research community.

The U.S. public policy environment includes a significant oversight and accountability component that requires information about productivity.

Differentiating outputs. Controlling for output heterogeneity is the other important enhancement. The resources required to produce an undergraduate degree vary significantly across fields, and the cost of bachelors' degrees differs systematically from the costs of associate, graduate, and first professional degrees. Failure to control for these differences would risk mistaking output shifts from the more expensive disciplines of science, technology, engineering and mathematics (STEM) to non-STEM disciplines for a decline in productivity.

Enhancing the model in this way will require the data for both degrees and enrollments to be differentiated by field and award level. This poses no problem because IPEDS already provides the needed data. Differentiating enrollments is a different story, however, because IPEDS does not report credit hours by field. Institutions may track credit hours by the department or discipline in which the course is taught—in order to apply the Delaware cost benchmarks, for example—but these data cannot be mapped directly to degree production because students take many courses outside their matriculated areas. Researchers have made the necessary correspondences by creating course-taking profiles for particular degrees, but these matrices are difficult to manage and maintain on an institution-wide basis.

A better way is to collect data in a way that follows the students, not just the departments that teach them. The information needed to do this exists in most institutions' student registration files. Extraction of the needed data could proceed as follows:

- "Identify the students matriculated in a given degree program ('output category') as defined by the IPEDS fields and degree levels. Undeclared students and students not matriculated for a degree would be placed in separate ('non-attributable') output categories.
- For each output category, accumulate the credit hours earned by the students in that category, regardless of the department in which the course was offered or the year in which it was taken.
- Allocate credits earned by matriculated but undeclared students in proportion to the credit-hour fractions of declared students for the given degree. Retain non-matriculated students in their own separate category, one that has no sheepskin effect but in other respects is treated the same as the other categories" (NRC, 2012a, p. 78).

Once again, the value of such statistics will greatly transcend productivity measurement. One can envision new approaches to the analysis of graduation rates and times to degree, for example, and credit hour measures that "follow the student" will permit more accurate costing measures to be produced within institutions. As noted above, we believe the requisite data exist within existing institutional data files. Hence all that is necessary is to develop appropriate extraction algorithms.

Other Data-Related Recommendations

The NRC panel's report offers a number of additional recommendations pertaining to the definition and development of datasets. The importance of most of these is self-evident, and readers should refer to the original report for additional discussion.

Academic staff play a unique and critically important role in most institutions. They, and only they, can make the intellectual judgments needed to create new knowledge, decide curricular and pedagogical issues, and assure educational quality.

The first such recommendation states that, "Definitions should be established for outcomes and institutions other than traditional four-year colleges and universities with low transfer-out rates, and appropriate bonus figures estimated and assigned to those outcomes. This is especially important for community colleges where, in contrast to BA and BS degrees, outcomes may be successful transfers to 4-year colleges, completion of certificates, or acquisition of specific skills by students with no intention of pursuing a degree" (NRC, 2012a, p. 92).

Two recommendations address the handling of research in universities where it is a major mission element. First, "NCES or a designee should develop an algorithm for adjusting labor and other inputs to account for joint production of research and service. Faculty labor hours associated with instruction should exclude hours spent on sponsored research and public service, for example and the algorithm should provide an operational basis for adjusting other inputs on the basis of expenditures" (NRC, 2012a, p. 95). Our conceptual framework already provides an algorithm for separately budgeted research and service activities (based on the Delta Cost Project and A21), but it may be possible to refine the approach.

The treatment of "departmental" (i.e., not separately budgeted, including sponsored) research is considerably more difficult. Such research is paid for by the university, often in the form of teaching-load reductions. The panel recommends the development of a statistical model to parse departmental research (DR) into two components: *Project-driven* and *Discretionary*. Such a study would use sample data on faculty activity to build statistical models for estimating: (a) the amount of activity that should be classified as departmental research, by institutional type, field, amount of sponsored research, and other descriptors; (b) the share of that activity directly associated with sponsored projects; and (c) the (residual) share of activity that should be classed as discretionary. We believe such a study will turn out to be practical. It would be immensely valuable for costing and other purposes as well as for productivity measurement.

As stated in the report, "The direct link between *Project-driven DR* and sponsored research provides a strong argument for excluding the former from instructional costs. Only the idiosyncrasies of university accounting and the market power of sponsoring agencies enable those agencies to enforce cost-sharing on academic-year effort in order to spread their funds further. Arguing on principle for inclusion of research cost and instructional cost is tantamount to arguing that the sponsored research itself should be included–which, in addition to being intrinsically illogical, would hugely distort the productivity measures" (NRC, 2012a, p. 97).

Discretionary DR, on the other hand, refers to work initiated by faculty members without regard to external support. In the panel's view, "Good arguments exist for including at least a part of such activity in the cost base for instruction. For one thing, it is difficult or impossible to separate the effects of educational research and development (R & D) from the other motivators of low teaching loads (other than those associated with sponsored research projects), and there is no doubt that educational R & D should be included in the instructional cost base. Meaningful education R & D expenses and work that sustains the life of disciplines (and is not sponsored research) should be defendable to stakeholders.

A better way is to collect data in a way that follows the students, not just the departments that teach them. The information needed to do this exists in most institutions' student registration files…Hence all that is necessary is to develop appropriate extraction algorithms.

Additionally, some allocation of faculty time entails service that is required to keep institutions running" (NRC, 2012a, p. 99).

Three more recommendations address more general data questions pertaining to higher education. The first says that, "Every effort should be made to include colleges and universities in the U. S. Economic Census, with due regard for the adequacy of alternative data sources and for the overall value and costs added, as well as difficulties in implementation" (NRC, 2012a, p. 110). Colleges and universities were included in the economic census only once, in 1977. Considering the importance of the sector, the amount of resources it consumes, and recent advances in university data systems, we could find no reason for continuing the exemption.

Another recommendation states that, "Standardization and coordination of states' student record databases should be a priority. Ideally, NCES should revive its proposal to organize a national unit record database" (NRC, 2012a, p. 117). We recognize that this is a politically difficult recommendation that may take years to realize, but emphasize that such a system, with appropriate privacy safeguards, would be extremely valuable for both policy and research purposes.

Finally, the panel supported the idea that, "The Bureau of Labor Statistics should continue its efforts to establish a national entity such as a clearinghouse to facilitate multistate links of unemployment insurance (UI) records and education data. This would allow for research on issues such as return on investment from postsecondary training or placement rates in various occupations" (NRC, 2012a, p. 118). The importance of state longitudinal databases, the national student clearinghouse, and various survey-based databases also are discussed in the report.

In closing, we hope that concerns about productivity measurement in higher education reflected in the charge to the panel continue to receive the attention they deserve. For example, a government entity such as the Department of Education, Census Bureau, or Bureau of Labor Statistics should be charged with overseeing and testing the implementation of our conceptual framework. We hope that such a group will be formed without delay, and also that individual institutions will begin to experiment with some of the data enhancements described in this paper.

References

Bureau of Labor Statistics (BLS). (2007). *Technical information about the BLS multifactor productivity measures*. Available at http://www.bls.gov/mfp/mprtech.pdf

Delta Cost Project. (2009). *Metrics for improving cost accountability* (Issue Brief No. 2). Washington, DC: Author. Available at http://www.deltacostproject.org/resources/pdf/issuebrief_02.pdf

Massy, W. F., Sullivan, T. A., & Mackie, C. (2013). Improving measurement of productivity in higher education. *Change*, January/February.

National Research Council (NRC). (2012a). *Improving measurement of productivity in higher education*. Panel on Measuring Higher Education Productivity: Conceptual Framework and Data Needs. T. A. Sullivan, C. Mackie, W. F. Massy, & E. Sinha (Eds.). Committee on National Statistics and Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (NRC). (2012b). *Research universities and the future of America: Ten breakthrough actions vital to our nation's prosperity and security*. Committee on Research Universities; Board on Higher Education and Workforce; Policy and Global Affairs; National Research Council. Washington, DC: The National Academies Press, June.

Organization for Economic and Cooperative Development (OECD). (2001). *Measuring productivity, measurement of aggregate and industry-level productivity growth*. Paris: Author.

Abstract

Critics of standardized testing have recommended replacing standardized tests with more authentic assessment measures, such as classroom assignments, projects, or portfolios rated by a panel of raters using common rubrics. Little research has examined the consistency of scores across multiple authentic assignments or the implications of this source of error on the generalizability of assessment results. This study provides a framework for conceptualizing measurement error when using authentic assessments and investigates the extent to which student writing performance may generalize across multiple tasks. Results from a generalizability study found that 77% of error variance may be attributable to differences within people across multiple writing assignments. Decision studies indicated that substantive improvements in reliability may be gained by increasing the number of assignments, as opposed to increasing the number of raters. Judgments about relative student performance may require closer scrutiny of task characteristics as a source of measurement error.

AUTHORS

John D. Hathcoat, Ph.D.
Oklahoma State University

Jeremy D. Penn, Ph.D.
Oklahoma State University

CORRESPONDENCE

Email
john.hathcoat@okstate.edu

Generalizability of Student Writing across Multiple Tasks: A Challenge for Authentic Assessment

For decades standardized testing in postsecondary education was limited to admissions testing. However, the influential report from the Secretary of Education's Commission on the Future of Higher Education recommended that all postsecondary education institutions should "measure student learning using quality-assessment data" with nationally standardized measures like the Collegiate Learning Assessment and institutions should make the results from those standardized tests "available to students and reported in the aggregate publicly" (United States Department of Education, 2006, p. 24). Critics have questioned the usefulness of standardized tests for both institutional accountability and institutional improvement. Common concerns with standardized tests include an overemphasis upon narrowly focused skills/abilities and content, the mismatch between the standardized tests and students' experiences at an institution, as well as students' motivation to complete such tests (Banta, 2006). Instead of standardized tests, researchers have suggested using what is called "authentic assessment," which includes approaches like assessment of e-portfolios, or assessment of writing and critical thinking (usually embedded in a course) using a common rubric (Banta, Griffin, Flateby, & Kahn, 2009).

Authentic assessment procedures may more directly reflect student experiences than standardized tests, though it remains unclear the extent to which it is appropriate to use authentic assessments in place of the many uses of standardized test scores. For instance, one desired use of standardized test scores is to compare students' performance across different schools (Benjamin, 2012). Standardized tests are standardized to control for specific sources of potential error – namely, differences in the characteristics of

If the assumption that student performance is consistent across multiple tasks is not tenable, and if authentic assessment is ever going to replace the numerous roles of standardized tests, then strategies must be developed to address task consistency.

tasks included within a test and the consistency of scores across alternative test forms. This does not negate many concerns with standardized tests, given that a standardized test score may reflect a single measure of a student's attribute, performance, or ability that fails to generalize to other settings. But authentic assessments, by their very nature, do not readily lend themselves to the same level of control across multiple sources of error. Just as score inconsistency across multiple items and/or alternate test forms restricts inferences from standardized tests, inferences derived from authentic assessments may be affected by multiple sources of error. Put differently, score consistency (i.e., reliability) is a necessary but insufficient condition to justify *any* use of scores deriving from an assessment regardless of whether it is standardized or authentic.

Examining the role of distinct sources of measurement error, along with interactions across these sources, remains of paramount importance in assessment practices. Such concerns, however, may lead to specific challenges for authentic assessment. It was therefore the purpose of the present study to address two concerns that that are potentially disconcerting for authentic assessment practices. First, common sampling strategies implicitly assume that some sources of measurement error are irrelevant. For example, many authentic assessment processes presume that *student performance is consistent across multiple tasks.* Users of assessment data may reasonably wonder if judgments about which students are doing best drastically changes across tasks or if measurement error is within acceptable limits. If the assumption that student performance is consistent across multiple tasks is not tenable, and if authentic assessment is ever going to replace the numerous roles of standardized tests, then strategies must be developed to address task consistency. Secondly, we investigate this assumption by applying generalizability theory to authentic assessment data (i.e., writing performance) collected at Oklahoma State University. Before proceeding to the study findings, a broad framework for vivisecting error variance through the lens of generalizability theory is provided.

Vivisecting Measurement Error across Multiple Tasks Using Generalizability Theory

This section provides an initial framework for conceptualizing the influence of assignment or task characteristics as a source of measurement error with respect to specific assessment goals and sampling strategies (Table 1). This framework is not meant to be inclusive, but is instead presented to illustrate a fundamental assumption with respect to sampling designs and measurement error: *If a single assessment or test is assumed to be representative of a student attribute, trait, or skill as a whole then evidence should be provided that such a use of that score is plausible.* This does not imply that assessment practitioners are explicitly aware of this principle when sampling specific assignments and/or tasks. In fact, we believe quite the contrary. In our own general education assessment practices at OSU, we have assumed that a single observation of a student's work is a reasonable estimate of performance when making comparisons at the institutional level. Although this is a low-stakes assessment for students, the kinds of inferences we hope to draw from this assessment process require that this assumption holds, and this may be specially concerning when employing specific sampling strategies (see Table 1). However, a failure to acknowledge or test this assumption does not render it unimportant. If this assumption is reasonable, judgments about student differences may be made irrespective of task characteristics. If this assumption is not tenable then judgments about students' performance may change if the researcher happened to sample a different task.

Evidence for person by task interaction effects may be particularly devastating given that this implies that judgments about which students are doing better depend upon the specific task that is assessed. With respect to writing assessment, this interaction would suggest that judgments about relative student differences drastically change across writing tasks or assignments, which may even occur within a single course. This particular source of measurement error can hinder assessment efforts targeting both within-group and between-group comparisons (Table 1). For example, two writing tasks, or assignments,

Table 1

Assessment Goals, Sampling Designs, and Measurement Error Related to Task Characteristics

Goals	Level of Generalization	Example Research Question	Sampling Design	Assignment as Source of Error
Between-Group / Cross-sectional	Institutional	On average are writing scores in 2012 higher than writing scores in 2011?	Random selection of student writing papers across level of interest.	Limited concerns; distribution of assignments should be checked across relevant comparisons.
	Program	Do students having experience 'X' tend to do better in writing than students who have not experienced 'X'?	Other than random	Potential bias due to distribution of assignments. Judgments about groups / individuals may change across assignments.
	Classroom	How well are students writing in this class?		
Within-Group	Institutional	On average, do freshman writing scores tend to improve by senior year?	Assignments have same prompt.	Limited concerns about assignment; check other issues stemming from design (e.g. practice effect)
	Program	How do writing scores change after participating in program 'X'?	Assignment has different prompts.	Potential bias resulting from a change in prompts.
	Classroom	Are writing scores improving across the semester?		Conduct G-study prior to large assessment, or design a study to control for prompt characteristics.

may be collected across the same students in order to assess changes in performance across time. Inferences about such changes are reasonable to the extent to which the two writing tasks are similar. A fundamental challenge, it would seem, is to provide evidence that tasks are sampled from the same theoretical domain, or the same universe of possible tasks. To once again place this argument within the context of writing assessment, claims about student writing performance must either be restricted to the specific task that is sampled, or evidence should be provided that performance generalizes across multiple tasks that are believed to be interchangeable.

If a single assessment or test is assumed to be representative of a student attribute, trait, or skill as a whole then evidence should be provided that such a use of that score is plausible.

Classical test theory (CTT), which is typically used to investigate score reliability via test-retest correlations, alternate forms, and/or internal consistency methods, is clearly limited for addressing these concerns. CTT, which assumes that an observed score may be decomposed into an expected true score and random error (Crocker & Algina, 1986), not only fails to consider multiple sources of error simultaneously but also fails to investigate interaction effects across sources of measurement error. Generalizability theory, or G-theory (Cronbach, Glesser, Nanda, & Rajaratman, 1972), has less restrictive assumptions than CTT and in many respects supplants this framework since it has been repeatedly demonstrated that investigations of reliability under CTT are special cases of G-theory designs (e.g., Brennan,

2011). Though both authentic assessment and G-theory have been utilized for some time now, for reasons that extend well beyond the scope of the present article, it appears that the utility of this approach for understanding sources of measurement error within the context of authentic assessment has yet to be fully realized. Others have addressed G-theory in detail (Brennan, 2001), and there are many good introductions to this topic (e.g., Shavelson & Webb, 1991). The following section will therefore close with a conceptual introduction to concepts employed within G-theory.

Conceptual Overview of Generalizability Theory

Claims about student writing performance must either be restricted to the specific task that is sampled, or evidence should be provided that performance generalizes across multiple tasks that are believed to be interchangeable.

G-theory utilizes analysis of variance techniques in order to further partition error into distinct sources of variation. These sources of variation are referred to as variance components, and estimating the relative magnitude of these components is of substantive interest in a *G-study*. A crucial task in designing a G-study is specifying the conditions of measurement, or *facets*, which presumably influence variation in observed scores. Facets may be either *crossed* or *nested*. A facet is considered crossed if every level of the first facet is observed at each level of the second facet (e.g., each student responds to every item), or alternatively a facet is considered nested within another if levels of one facet are observed at only one level of another facet (e.g., items may be nested within students if each student receives multiple items, but no student receives the same items). Facets may also be *random* or *fixed*. A facet is considered random if random sampling of each level has occurred or if the researcher is willing to treat observed levels as interchangeable (e.g., items may be replaceable with any other item of similar characteristics). A facet is considered fixed if the researcher has observed each level of facet or if the researcher does not wish to generalize beyond the observed levels of a facet.

Within a G-theory framework each observed level of a random facet may be viewed as a sample from a defined universe of acceptable observations. For example, within the context of writing assessment we are not necessarily interested in a student's performance on a specific assignment or writing task. *Instead, the specific task that is used may be viewed as one of many possible tasks that could have equally been utilized to assess writing performance*. In this case, we are interested in our ability to consistently estimate scores across tasks defining a universe of acceptable observations, irrespective of the specific writing task that was actually sampled in our assessment procedure. The generalizability coefficient ($E\rho^2$; Cronbach et al., 1972), which is the ratio of universe score variance to observed score variance (Webb, Shavelson, & Haertel, 2007), provides such an estimate by allowing us to examine the extent to which consistent estimates about relative student performance may be inferred across multiple tasks that are considered interchangeable. Generalizability coefficients range from 0 to 1.0, with acceptable coefficients ranging from .70 to .80 or higher (Brennan, 2001). Decision studies or D-studies may then be conducted to investigate how changes in specified facets may minimize error variance. We now summarize our own investigation of task variability as a source of measurement error within the context of general education assessment using G-theory.

Methods

Procedure

Each year Oklahoma State University (OSU) assesses the general education program and generates an annual report (http://tinyurl.com/osugened). This assessment effort typically involves sampling student papers (i.e., tasks) from courses across the campus. Each year tasks are sampled within the same semester, and faculty members act as paid raters who then score each paper in small independent groups of 2-3 members. Although the overall goal of the assessment process is to make general judgments about the extent to which students are achieving general education learning goals, as previously discussed, these judgments may still be affected by the task or assignment characteristics.

We began by examining the number of students for whom we had, by happenstance, scored more than one assignment or task in the entire set of data from 2001-2011. Of the scored areas, writing had been evaluated every year from 2001-2011, with the exception of 2007. In the 10 years in which writing was assessed there were a total of 1,831 scores, of which we identified 29 students who had more than one paper scored for writing. Of these, seven students had writing tasks sampled across different years of data collection. To avoid confounding results across years, these students were removed from subsequent analyses. The remaining 22 students were scored on two writing tasks sampled within the same semester, though each task was scored by an independent group of two faculty raters. This provided a total of 44 tasks, each of which was scored by two faculty raters, thus making 88 total observations. Since sample size may contribute to the stability of estimated variance components (Webb et al., 2007), the size of this design warrants some caution. However, the number of observations employed within this study is similar to many investigations utilizing G-theory.

Instead, the specific task that is used may be viewed as one of many possible tasks that could have equally been utilized to assess writing performance.

Instrumentation

Before faculty raters are assigned writing tasks to score they are first trained to use a rubric developed at OSU (see Appendix A). Scoring procedures have slightly varied throughout the years, though typically each faculty member rates tasks independently and then meets with their group in order to reach consensus with respect to each task's assigned score. Each task is scored on a 1-5 scale on content, organization, mechanics, and documentation so that higher scores reflect greater writing performance. In addition to dimensional scores faculty raters also provide an overall score reflective of the general writing performance exemplified by a student paper. The overall score provided by each faculty rater prior to consensus was utilized in the present analysis. Inter-rater reliability estimates tend to vary across groups of raters when approached under a CTT framework. A benefit of setting up a G-study is that distinct sources of error may be simultaneously examined in terms of their relative contribution to error. Reliability analyses are detailed in the results section.

Analytic Design

There were a total of 22 students who were sampled on two different writing tasks. Each task was scored by an independent group of two raters. Raters were therefore nested within tasks. However, given that each task was also different across persons tasks are considered to be nested within persons. Though there are statistical disadvantages to a fully nested design (i.e., confounded sources of error) this design resulted from restrictions deriving from decisions that were made about previous sampling strategies. Persons were treated as the object of measurement and both raters and tasks were conceptualized as a random sample from a potentially infinite number of observations. This entailed a fully nested, random effects design wherein the following variance components were estimated: persons (σ^2_P), tasks within persons $(\sigma^2_{T:P})$, and raters within tasks within persons $(\sigma^2_{R:T:P})$. The main effect of persons (σ^2_P) indicates the estimated variance component for between-person differences in average writing performance. Within the current study this variance component reflects the 'universe' from which we wish to make consistent inferences about student writing. The variance component for tasks within persons $(\sigma^2_{T:P})$ reflects mean differences in assignment scores for each person across the pairs of raters. Given that each task was assigned to a different group of raters this variance component cannot be disentangled from a person by task interaction. The variance component for raters nested within tasks nested within persons $(\sigma^2_{R:T:P})$ indicates differences in assigned scores within a single group of raters for a particular task. This source of variation is also confounded with unexamined sources of error.

Given that persons were the object of measurement we focused on the ability of scores to provide relative comparisons about inter-individual differences in writing

performance. In estimating the generalizability coefficient relative error is a function of both $(\sigma^2_{T:P})$ and $(\sigma^2_{R:T:P})$:

$$\sigma^2_{\delta} = \frac{\sigma^2_{T:P}}{n_T} + \frac{\sigma^2_{R:T:P}}{n_T n_R} \quad (1)$$

These values suggest that raters within each task tended to display little disagreement about the overall writing performance indicated by a particular student paper…A failure to find such differences however, indicates little about the consistency of rank ordering student writing ability.

From this equation it can be seen that both increases in $(\sigma^2_{T:P})$ and $(\sigma^2_{R:T:P})$ will inflate the amount of error in making normative judgments about student writing ability. Variation in average ratings assigned to tasks within a person and variation of raters within each task contribute to an inability to make consistent judgments about relative student writing performance. Increases in the number of observed tasks (n_T) and the number of assigned raters within each task (n_R) will decrease relative error given that the estimated variance components remain constant. Estimates of relative error are utilized in order to calculate the generalizability coefficient:

$$Gen\_Coefficient = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_{\delta}} \quad (2)$$

From equation 2 it can be seen that the generalizability coefficient is expressed as a ratio of total between person variation (i.e., universe score variance) to estimated observed score variation. Increases in the magnitude of relative (σ^2_{δ}) error will reduce the generalizability coefficient whereas increases in universe score variance (σ^2_P) will tend to increase the generalizability coefficient. As previously indicated this coefficient may be interpreted as the extent to which one may make consistent normative inferences about student writing performance across all possible raters and tasks.

Results

Descriptive statistics were first examined on the 22 students (i.e., 44 writing tasks) who had each task assessed by an independent group of two raters (see Table 2). It is of particular interest to note that the variation of assigned scores within raters for each task was relatively low. Within-task rater variance ranged from 0.00 to 1.00 with an average variance across each task of 0.13. These values suggest that raters within each task tended to display little disagreement about the overall writing performance indicated by a particular student paper. With such data many researchers may choose to utilize inferential statistics in order to investigate either mean difference across each writing task or the linear relationship between assigned scores across each writing task. For this analysis the mean rating provided by both judges for a single task was the outcome variable. A dependent sample t-test indicated no statistically significant differences across mean ratings assigned across writing tasks. A failure to find such differences however, indicates little about the consistency of rank ordering student writing ability. The observed correlation across each writing task was .178 (p = .429), which implies that the pattern of student writing scores across each task was highly inconsistent. Estimated variance components from the G-study were examined in order to investigate these inconsistencies using EduG 6.0 (Cardinet, Johnson, & Pini, 2010).

Results from this analysis are presented in Table 3. The object of measurement, between-person differences in student writing, consists of approximately 12% of the total variation. Though not large, this represents the signal that the assessment procedure is attempting to detect. Rater variation within each task that is also nested within each person consists of approximately 23% of the total error variation. Though the magnitude of this variation is substantive it is of particular interest that 77% of the error variance derives from differences within a single person across each task. As previously indicated, the design of this study confounds a task effect with a person by task interaction. The vast majority of error variance may be attributed to either a task effect or the possibility that the rank ordering of individuals in writing changes across each sample of tasks. The estimated generalizability coefficient was .28 (*SEM* = .55), which is far below acceptable standards. If we assume that error is normally distributed we may utilize the standard

Table 2

Task Variation and Mean across Raters

Person	Task	Task Mean	Task Variation
1	1	3.00	0.00
1	2	4.00	0.00
2	1	3.00	0.00
2	2	2.50	0.00
3	1	3.50	0.25
3	2	3.00	0.25
4	1	4.00	0.00
4	2	3.00	0.00
5	1	4.00	0.00
5	2	3.50	0.00
6	1	3.00	0.25
6	2	2.00	0.00
7	1	4.50	0.00
7	2	1.50	0.25
8	1	3.00	0.25
8	2	3.00	0.00
9	1	3.00	0.00
9	2	2.00	0.00
10	1	3.50	0.00
10	2	4.00	0.25
11	1	4.00	0.00
11	2	5.00	0.00
12	1	3.50	0.00
12	2	2.50	0.25
13	1	3.00	0.25
13	2	4.00	0.00
14	1	3.00	0.00
14	2	4.50	0.00
15	1	3.50	0.25
15	2	1.50	0.25
16	1	3.50	0.25
16	2	3.50	0.25
17	1	2.00	0.25
17	2	2.00	0.00
18	1	2.00	0.00
18	2	3.00	0.00
19	1	2.00	0.00
19	2	2.00	0.00
20	1	3.50	0.00
20	2	2.50	0.25
21	1	4.00	0.25
21	2	3.00	1.00
22	1	2.00	0.00
22	2	2.50	1.00

Note : Two raters are nested within each task. Tasks are also nested within students.

Table 3

Variance Component Estimates for Raters Nested in Tasks Nested in Persons Design

Source of Variance	*SS*	*df*	*MS*	Variance Estimate	Percent of Error Variance
Person	35.09	21	1.67	0.12	N/A
Task within Person	1.63	1	1.63	0.47	77.4%
Raters within Tasks within Persons	0.59	2	0.29	0.27	22.6%

Note: *SS* = sum of squares; *df* = degrees of freedom; *MS* = mean square. N/A = not applicable. N=22 persons rated by two groups of independent raters on two different tasks.

Examination of these confidence intervals suggest that individuals receiving a mean score of one…are indistinguishable from students assigned a mean score of two…though they may be distinguished from individuals assigned a score of three…or higher.

error of measurement in order to construct confidence intervals around mean scores on the writing rubric. Examination of these confidence intervals suggest that individuals receiving a mean score of one (95% CI = -0.078 to 2.078) are indistinguishable from students assigned a mean score of two (95% CI = 0.922 – 3.078), which in turn are indistinguishable from those with a mean score of three (95% CI = 2.922 – 4.078). Individuals with a mean score

It appears that, in order to use authentic assessments to make direct comparisons of students' scores, understanding the impact of task characteristics may very well be the biggest challenge.

of four (95% CI = 3.922 to 5.078) are, for all practical purposes, indistinguishable from students receiving a mean score of five (95% CI = 4.922 – 6.078). Stated differently, current assessment practices may distinguish those with relatively low scores (i.e., mean score of 1 and 2) from those with relatively high scores across both assignments (mean score of 4 and 5). However, more subtle distinctions in student performance across these tasks may not be consistently inferred.

Several D studies were conducted in order to evaluate the expected impact of increasing both the number of sampled writing tasks and the number of raters assigned to score each task. As indicated by Figure 1, little increase in the generalizability coefficient is predicted by increasing the number of raters assigned to each task. While holding the number of tasks constant at five the predicted generalizability coefficients range from .51 to .53 across three to seven raters. However, greater gains in the generalizability coefficient may be made from increasing the number of tasks collected on each person. While holding the number of raters constant at three the predicted generalizability coefficient ranges from .51 to .75 when increasing the number of tasks from 5 to 15. This pattern substantiates inferences from the G-study that suggested an increase in the number of raters assigned to a particular task may be of limited utility given the relative magnitude of error associated with differences assigned to tasks within a person. Instead, greater precision in making judgments about relative student writing performance may derive from increasing the number of observed writing tasks. Unfortunately, the number of tasks needed to substantially improve these inferences may be unobtainable in most assessment contexts due to the cost of collecting and scoring a substantially larger number of assignments.

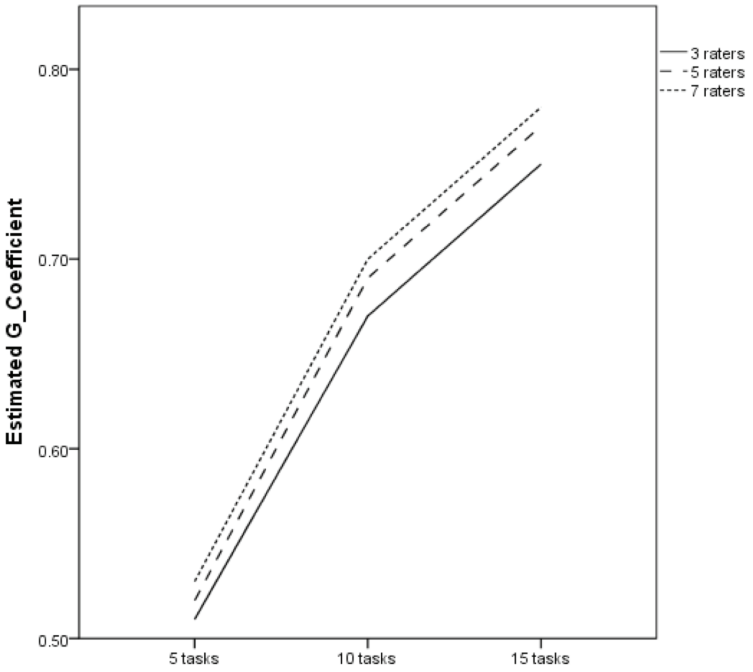

Figure 1. D studies for raters within each task by number of sampled tasks.

Discussion

Inter- and intra-rater reliability were once believed to be the biggest problems with authentic assessments. Instead, it appears that, in order to use authentic assessments to make direct comparisons of students' scores, understanding the impact of task characteristics may very well be the biggest challenge. Without an a priori equating of writing tasks, judgments derived from authentic assessment may largely depend upon the kinds of tasks students

are asked to perform. This is not to say that authentic assessment should be abandoned, nor does this evidence imply that standardized tests should replace authentic assessment. Instead, further investigation is needed to explicate the conditions under which generalized inferences are justified. The success of authentic assessment may therefore depend upon systematic efforts to articulate why judgments about relative student performance seems to change across separate tasks.

Within our sample, it was clear that the ordering of students by writing performance depended upon which task was selected. This study suggests that if researchers want to make comparisons about students' performance from authentic assessments between institutions or within an institution, they should greatly increase the number of tasks that are sampled for each student, establish statistical controls based on variables that are shown to impact students' performance (such as motivation), or take steps to standardize some task characteristics (which may not be palatable for users of authentic assessment). More than twenty years ago Elliot Eisner wrote, "Our nets define what we shall catch" (1992). Our study supports this statement by suggesting that what our students are able to show they can do is in part influenced by what we ask them to do.

If researchers want to make comparisons about students' performance from authentic assessments between institutions or within an institution, they should greatly increase the number of tasks that are sampled for each student, establish statistical controls based on variables that are shown to impact students' performance (such as motivation), or take steps to standardize some task characteristics.

While standardized tests may make a stronger case for controlling specific sources of measurement error, other aspects of standardized tests may not compare favorably with authentic assessments. First, if the content of a standardized test is selected at the national level and does not represent the goals, mission, and learning outcomes desired by an institution, it may be just as dubious to claim such a test is a reasonable comparison between institutions. Second, the extent to which scores on standardized tests extends to the kinds of tasks students perform throughout their education remains controversial. Just as our evidence implies that a single observation of writing performance fails to generalize to performance across other tasks, a similar issue may exist with standardized tests since these scores may also fail to generalize to scores observed on similar tasks outside a controlled testing environment. Third, there is some research to suggest that task characteristics are important to standardized tests as well. Russell and Plati's (2000) study illustrates this point. They found that student performance on a standardized test depended upon the mode of administration (computer or paper), and student performance was also a function of prior keyboarding skills. Even though standardized tests use a similar task across all examinees, the characteristics of the task still matter when making inferences using the scores from standardized tests.

Finally, regardless of whether authentic assessment or standardized tests are used to draw inferences, this study highlights the importance of explicitly addressing assumptions about the contribution of particular sources of measurement error. Specifically, when observing a single student assignment, or task, there are dangers in interpreting the scores as though they were independent of the task being sampled. Findings from the present study suggest interpretations that fail to account for task variation may be problematic since they presume that judgments about relative student differences are consistent across distinct tasks. Numerous authors have raised similar concerns (e.g., Kroll & Reid, 1994; Schoonen, 2005; Shavelson, Baxter, & Gao, 1993), and this study provides additional support that may serve to caution drawing unwarranted inferences from assessment results. Before proceeding to the implications of the present study, a central limitation will first be addressed.

Limitations

Numerous limitations exist with the present study; however, one limitation is particularly salient. Historical data were used in an effort to investigate the extent to which assumptions about the consistency of student performance across multiple tasks may be problematic. Methodological choices of previous assessment efforts restricted the analytic design employed within the current study. Within the current study raters were nested within writing tasks, which in turn were nested within persons. This design confounds important sources of error that may be important when deciding which strategies to adopt in subsequent assessment procedures. For example, this nested design makes it impossible

to disentangle a task effect and a person by task interaction. A fully crossed design would allow the separation of interaction effects between persons and tasks and persons and raters. Though the analytic design was not ideal, it provides tentative evidence in support of a growing concern about task variability as a source of measurement error within assessment practices.

Future Research

If a random number appeared with each observation of a pocket-watch, it would be challenging, but more importantly extremely misleading, to argue for the validity of a particular interpretation of these observed "times." We would not be able to use the pocket-watch to complete even a simple task accurately, such as putting students in order based on their time of arrival to class. No matter how carefully we analyze the scores from the random-number generating watch, they remain of little value. Without score consistency (e.g., we observe a similar time when each observation is conducted with the sun being at a particular point in the sky) nothing is being measured (Thompson, 2003). Reliability is thus a prerequisite, and principle justification, for the assignment of meaning to a set of scores. It is the role of validation to investigate evidential support for proposed interpretations given to a set of scores. Stated differently, reliability justifies the assignment of meaning to a set of scores, and validation is a constructive act whereby evidence is accumulated to articulate the limits, boundaries, and extension of a particular interpretation. Both reliability and validity are hence central considerations that inform decisions derived from educational assessment procedures.

When observing a single student assignment, or task, there are dangers in interpreting the scores as though they were independent of the task being sampled.

Fluctuations in student performance across multiple tasks, particularly if performance is a function of task characteristics, restrict the kinds of inferences that are justified when interpreting assessment results. Unfortunately, a simple panacea does not, at least as far as current research suggests, exist. As a first step, it is necessary to replicate the present findings in a study explicitly designed to control for confounded sources of error. Instead of using a nested design, a fully crossed design wherein every rater scored the same students on the same multiple tasks would be ideal from a G-theory perspective. Despite such constraints, the present results have led to concerns with our own assessment procedures, and additional data is currently being collected in an effort to further investigate the role of task characteristics as a source of measurement error when attempting to assess students' writing performance. Note that the present study is delimited to student writing performance, though we suspect that similar issues may arise when investigating other valued learning outcomes (e.g., critical thinking, intercultural competence, etc.). Examining this source of error in other authentic assessment processes (i.e., portfolios, critical thinking rubrics, etc.) is warranted. Though simple heuristic devices fail to account for the complexity of assessment practices, three general considerations are addressed that may be used to inform subsequent assessment efforts.

First, authentic assessment and G-theory have been discussed in the literature for some time now. Reliability estimation under classical test theory cannot address the complexity of task characteristics as a source of measurement error within authentic assessment practices. Consequently, we propose a "marriage" between G-theory and many assessment practices. G-theory provides greater flexibility to assessment practitioners, has less restrictive assumptions than classical test theory, and may be utilized to check data quality prior to implementing large scale investigations. Additionally, once specified, decision studies may be utilized to investigate ideal assessment procedures. The flexibility provided by G-theory does come at a cost. G-theory can be computationally complex and implementing this approach not only requires foresight into methodological design, but also careful consideration of how facets of measurement are specified as sources of measurement error. G-theory may not be appropriate for all assessment purposes, but continual advancement of this field appears to require practitioners to confront the challenges introduced by distinct sources of measurement error.

Second, person by task interaction effects may demand increased precision in how the universe of generalization is conceptualized. Stated differently, the presence of person by task interaction effects is suggestive of at least two possibilities that are in

need of subsequent investigation. First, it is conceivable that sampled writing tasks are interchangeable in that they are derived from the same theoretical domain. Under this view, one task should be equivalent to others in that the specific tasks that are sampled are indifferent with respect to judgments about relative student performance. The present analysis, which sampled two tasks, suggested that at least ten tasks may be necessary to derive reasonably consistent estimates about relative student differences. This could imply that the two sampled tasks, just by happenstance, failed to be representative of the universe of all possible tasks. Increasing the number of observations should therefore provide a better representation of the theoretical universe of all possible tasks.

An alternative interpretation is also possible. It is conceivable that the sampled tasks are not interchangeable, suggesting that it is mistaken to treat these tasks as a reflection of the same theoretical domain, or universe of generalization. In this case, either inferences about student writing must be restricted to the specific tasks that are sampled or greater care should be taken when conceptualizing the kinds of tasks that are judged to reflect the same theoretical domain. It is possible that writing tasks with characteristic "X" compose a separate universe of generalization than writing tasks with characteristic "Y." If so, then tasks may be sampled while controlling for characteristic "X," and consequently generalized inferences about relative student performance would be restricted to tasks denoted by such a characteristic. At this juncture there are many more questions than answers, and clearly more work is needed to investigate which of these alternatives may be more viable.

Reliability justifies the assignment of meaning to a set of scores, and validation is a constructive act whereby evidence is accumulated to articulate the limits, boundaries, and extension of a particular interpretation.

Finally, we wish to draw this discussion back to the controversies surrounding the issue of standardized tests and authentic assessment practices. As previously indicated, reliability estimates within authentic assessment practices, particularly with the use of rubrics, have generally focused on score consistency across or within raters (Finley, 2011/2012). Though controlling for this source of error remains important, this is only part of the story. Consistency across tasks is also an important source of error that stands in need of clarification. Elucidation of this source of measurement error, we contend, is intricately connected to criticisms of standardized tests, specifically criticisms residing in the question of whether *general* skills can be assessed (Banta & Pike, 2012; Benjamin, 2012). Person by task interaction effects, at least in principle, may be utilized as evidence to address such debates. For example, students may be given writing tasks across two disciplines that are then scored by trained raters using a common rubric. A person by task interaction would indicate that judgments about relative student differences changes across discipline, or in other words this evidence may suggest that performance is domain specific, which could then be used to argue for further refinement of the universe of generalization from which writing tasks are sampled. Alternatively, we could sample writing tasks within a single discipline utilizing the same procedures. A failure to find a person by task interaction may then imply that generalized inferences within a specific discipline are justified.

In conclusion, the current study suggests that caution is warranted when interpreting many assessment results. This caution stems from a generally unrecognized source of measurement error, namely the introduction of task variability. An accumulating body of evidence suggests that students' performance may be highly varied across tasks, and that judgments about which students are doing better may change across seemingly similar tasks. These problems can restrict warranted inferences from assessment results by limiting desired comparisons both within and between institutions. However, we do not universally reject authentic assessment as an important component of educational practice. To the contrary, we believe authentic assessment plays a critical role in evaluating educational programs and for making decisions about program improvement so long as such inferences carefully address distinct sources of measurement error investigated within this and other studies. This study underscores our concern with task variability as a source of measurement error, while acting as an invitation to other users of authentic and standardized assessment to join us in this investigation.

References

Banta, T. W. (2006). *A warning on measuring learning outcomes.* InsideHigherEd.com. Retrieved July 9, 2012 from http://www.insidehighered.com/views/2007/01/26/banta

Banta, T. W., Griffin, M., Flateby, T. L., & Kahn, S. (2009). *Three promising alternatives for assessing college students' knowledge and skills*. Occasional paper #2. National Institute for Learning Outcomes Assessment. Retrieved July 9, 2012 from http://learningoutcomesassessment.org/documents/AlternativesforAssessment.pdf

Banta, T.W., & Pike, G.R. (2012). *Making the case against—One more time. Occasional paper #15*. National Institute for Learning Outcomes Assessment. Retrieved July 9, 2012 from http://learningoutcomesassessment.org/documents/AlternativesforAssessment.pdf

Benjamin, R. (2012). The seven red herrings about standardized assessments in higher education. *National Institute for Learning Outcomes Assessment. Occasional paper #15*. National Institute for Learning Outcomes Assessment. Retrieved September 18, 2012 from http://learningoutcomesassessment.org/documents/HerringPaperFINAL.pdf

Brennan, R.L. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R.L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1-21.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Taylor and Francis Group.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L.J., Glesser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: John Wiley.

Eisner, E. (1992). In Blekin, G.M., & Kelly, A.V. *Assessment in early childhood education*. United Kingdom: Sage.

Finley, A. P. (2011/2012). How reliable are the VALUE rubrics? *Peer Review*, *13*(4), 31-33.

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3*, 231-255.

Russell, M., & Plati, T. (2000). *Mode of administration effects on MCAS composition performance for grades four, eight, and ten*. Malden, MA: Massachusetts Department of Education.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1-30. doi: 0.1191/0265532205lt295oa

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

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Thompson, B.T. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-30). Thousand Oaks, CA: Sage.

United States Department of Education (2006). *A test of leadership: Charting the future of U.S. higher education* [Secretary of Education's Commission on the Future of Higher Education]. Washington, D.C.

Webb, N.M., Shavelson, R.J., & Haertel, E.H. (2007). Reliability coefficients and generalizability theory. In C.R. Rao and S. Sinharay (Eds.), *Handbook of Statistics, Vol. 26* (pp. 81-121). Oxford, UK: Elsevier B.V.

Appendix A

Writing Rubric Developed at Oklahoma State University

	Skill	Level of Achievement 1	2*	3	4**	5
A	Content	Topic is poorly developed; support is only vague or general; ideas are trite; wording is unclear, simplistic; reflects lack of understanding of topic and audience; minimally accomplishes goals of the assignment.		Topic is evident; some supporting detail; wording is generally clear; reflects understanding of topic and audience; generally accomplishes goals of the assignment.		Topic/thesis is clearly stated and well developed; details/wording is accurate, specific, appropriate for the topic & audience, with no digressions; evidence of effective, clear thinking; completely accomplishes the goals of the assignment.
B	Organization	Most paragraphs are rambling and unfocused; no clear beginning or ending paragraphs; inappropriate or missing sequence markers. No clear over-all organization		Most paragraphs are focused; discernible beginning and ending paragraphs; some appropriate sequence markers. Overall organization can be inferred and is appropriate for the assignment		Paragraphs are clearly focused and organized around a central theme; clear beginnings and ending paragraphs; appropriate, coherent sequences and sequence markers. Overall organization is clearly marked and is appropriate for the assignment
C	Style and Mechanics	Inappropriate or inaccurate word choice; repetitive words and sentence types; inappropriate or inconsistent point of view and tone. Frequent non-standard grammar, spelling, punctuation interferes with comprehension and writer's credibility.		Generally appropriate word choice; variety in vocabulary and sentence types; appropriate point of view and tone. Some non-standard grammar, spelling, and punctuation; errors do not generally interfere with comprehension or writer's credibility.		Word choice appropriate for the task; precise, vivid vocabulary; variety of sentence types; consistent and appropriate point of view and tone. Standard grammar, spelling, punctuation; no interference with comprehension or writer's credibility.
D	Documentation	In text and ending documentation are generally inconsistent and incomplete; cited information is not incorporated into the document.		In text and ending documentation are generally clear, consistent, and complete; cited information is somewhat incorporated into the document.		In text and ending documentation are clear, consistent, and complete; cited information is incorporated effectively into the document.

Abstract

Since 2001, the University of Maryland University College (UMUC) Graduate School has been conducting outcomes assessment of student learning. The current 3-3-3 Model of assessment has been used at the program and school levels providing results that assist refinement of programs and courses. Though effective, this model employs multiple rubrics to assess a wide variety of assignments and is complex to administer. This paper discusses a new outcomes assessment model called C2, currently being piloted in UMUC's Graduate School. The model employs a single common activity (CoA) to be used by all Graduate School programs. It is designed to assess four of the five student learning expectations (SLEs) using one combined rubric (ComR). The assessment activity, scored by trained raters, displays pilot results supporting inter-rater agreement. Pilot implementation of the C2 model has advanced its reliability and its potential to streamline current assessment processes in the Graduate School.

AUTHORS

Rana Khan, Ph.D.
University of Maryland, University College

Datta Kaur Khalsa, Ph.D.
University of Maryland, University College

Kathryn Klose, Ph.D.
University of Maryland, University College

Yan Zhang Cooksey, Ph.D.
University of Maryland, University College

CORRESPONDENCE

Email
rana.khan@umuc.edu

Assessing Graduate Student Learning in Four Competencies: Use of a Common Assignment and a Combined Rubric

University of Maryland University College (UMUC) has been involved in institutional assessment of student learning in both its undergraduate and graduate schools since 2001. According to Palomba and Banta (1999), assessment is "the systematic collection, review, and use of information about educational programs undertaken for the purpose of improving student learning and development" (p. 4). UMUC's institutional assessment plan, consistent with Walvoord's (2004) recommendations, aligns with its mission, core values, and strategic plans. The plan also provides an overarching conceptual framework that defines student learning outcomes, provides a roadmap for assessing student learning, and ensures the use of findings for the improvement of UMUC programs. In the Graduate School, the current model of assessment is based on a framework introduced in 2010. This framework measures five student learning expectations (SLEs) and consists of *three* rounds of assessment at *three* stages carried out over a *three* year period each spring semester and has been named the 3-3-3 Model. Though the current process is effective in systematically collecting data across the Graduate School, it is a complex process to administer. This paper describes two phases of a pilot study, the intent of which was twofold: (a) to simplify the current Graduate School assessment process and (b) to examine and refine a new model that employs a recently developed assessment instrument. This article contributes to educational literature that focuses on graduate school assessment methods and will assist assessment practitioners by sharing the authors' experiences with piloting a new

assessment model. Details and results of the pilot study, including information on the current model, design of the new assessment model, online rater training, and interpretation of the pilot results follow.

Graduate School Assessment Process-Current Assessment Model

The Graduate School views assessment as an ongoing and collaborative process driven by continuous reflection and improvement.

In line with university priorities and strategies, UMUC's Graduate School has established a commitment to systematic assessment and the use of assessment results to improve student learning. The Graduate School views assessment as an ongoing and collaborative process driven by continuous reflection and improvement as described by Maki (2004). The current 3-3-3 assessment model employed by the Graduate School obtains evidence of student learning by assessing five student learning expectations (SLEs; Appendix A). The five SLEs include Communication (COMM), Critical Thinking (THIN), Information Literacy (INFO), Technology Fluency (TECH), and Content Knowledge (KNOW) and are expected of all UMUC graduate students.

The 3-3-3 model consists of three rounds of assessment carried out over a three-year period each spring semester, with each round assessing all five SLEs (See Figure 1). This model takes a "snapshot" of student learning at three points in a program lifecycle. Assessments are run within the first 9 credits, between 10 and 18 credits and at 19-36 credits, marking beginning, intermediate and advanced levels of study.

For each round, program directors, who manage courses in the Graduate School, identify the courses/sections that will conduct assessment activities. Within each course/section, class activities are chosen that will allow students to demonstrate their abilities in specific SLEs.

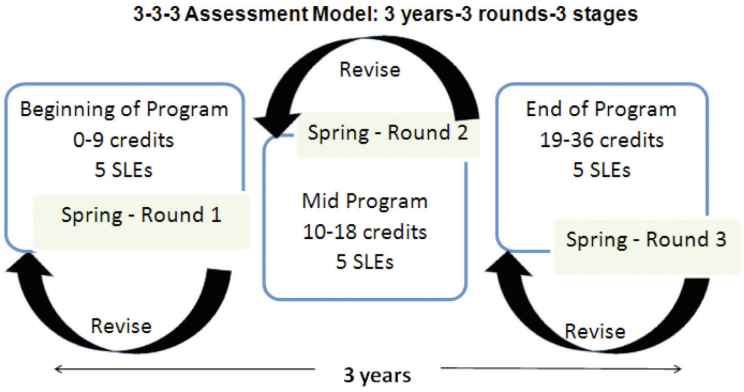

Figure 1. UMUC's 3-3-3 assessment model.

While the current 3-3-3 model has served the Graduate School well and proven reliable in delivering useful data for our goals, it has limitations and challenges.

There are a variety of tools that may be used for assessing student learning, including standardized tests, interviews, surveys, external examiners, oral exams, rubrics, and e-portfolios (Prus & Johnson, 1994). UMUC's Graduate School has chosen to use rubrics to assess student learning for each SLE for reasons aligned with the thinking of Petkov and Petkova (2006), who cite ease of implementation, low costs, student familiarity, and applicability to a variety of performance criteria. Rubrics can also be used in both formative and summative evaluation. For use with its current 3-3-3 model, the Graduate School designed a set of analytic rubrics where rubric criteria align with each of the school's five SLEs. Each rubric describes student performance over four progressively increasing levels of attainment (unsatisfactory, marginal, competent & exemplary).

Consistent with the design recommendations offered by Moskal (2000) and Nitko (2001), each Graduate School rubric contains criteria that serve to identify and describe the separate dimensions of performance that constitute a specific SLE. Instructors are required to score each rubric criterion and sum the scores. For example, the Graduate School

has identified the criteria of Conceptualization, Analysis, Synthesis, Conclusions and Implications as dimensions of the Critical Thinking SLE. When assessing assignments associated with Critical Thinking, faculty assign a score to each criterion, which is then summed up. By assigning a score to each criterion, faculty and course/program administrators receive multidimensional information on student performance. In addition to providing insights on specific levels of student learning, the inherent design of analytic rubrics employed in the 3-3-3 model provides students with specific feedback via the criteria definitions. The feedback enables students to focus on areas where they need improvement. The analytic rubric lends itself to formative use of rubric information, as opposed to the more summative approach inherent in holistic rubrics (Mertler, 2001). In this way, UMUC faculty and administrators use the results derived from the rubric scores to inform improvements to their courses and programs. In line with the iterative approach to rubric design described by Wiggins (1998), the Graduate School has over the past three rounds of assessment refined its rubrics based on assessment findings and user feedback. An example of a rubric currently employed in the 3-3-3 model is contained in Appendix B.

The primary resource needed for the development of the C2 model was time. The collaborative process took over a year from the time the idea was first proposed by the researchers to the Graduate School Assessment Committee to the time the pilot was conducted in Spring 2012.

When Graduate School faculty carry out assessment activities in their classes, they are responsible not only for assigning a class grade to select assessment assignments, but must also score the assignments using the appropriate Graduate School rubrics. The faculty must record the students' rubric scores for each specific SLE criteria on a summary sheet and submit the sheet to the Graduate School. Faculty and administrators are later provided with a summary of the assessment findings and asked to develop action plans to address the most significant areas of weakness in their programs. This completes the assessment cycle by looping actionable improvements into the course/program.

An example of this loop-back into courses and programs is the implementation of an Accounting and Finance Research Module designed by UMUC's Library Services. Round 1 assessment findings indicated that, related to the SLE of Information Literacy, students in Accounting and Finance scored low on the criterion of Identification and were not able to competently differentiate between scholarly and trade journals when conducting research. Upon analyzing the findings, the program director asked UMUC Library Services to develop a resource exclusively for helping students understand how to evaluate the quality of publications used in their research. Subsequent findings in Rounds 2 and 3 showed improvement in the criteria of Identification among Accounting and Finance students.

The Graduate School completed its first 3-year assessment cycle under the 3-3-3 model in Spring 2012. While the current 3-3-3 model has served the Graduate School well and proven reliable in delivering useful data for our goals, it has limitations and challenges that include:

- extra grading workload for faculty who teach courses identified for assessment,
- no training or norming for faculty on rubric use,
- disparities in the types of assignments used for assessment across the Graduate School,
- misalignments between the assignments and rubrics, and
- inconsistencies in grading practices among faculty.

As described by Buzzetto-More and Alade (2006), the reflection that occurs in relation to the assessment cycle often stimulates discussion and suggestions for improvements, and plans for implementing change. With the completion of the cycle came the opportunity to review the current model, which led to the design of the C2 model and current pilot study discussed in this paper.

Graduate School Assessment Process-Proposed Assessment Model

The limitations and challenges of the 3-3-3 model are not unusual in nature and relate to those described by those writing in the area of outcomes based assessment such as Banta (2002), Bresciani (2011), and Maki (2010). These challenges relate to understanding

the goals of assessment and having the resources and time necessary to carry out assessment activities. To address some of the aforementioned challenges, the authors proposed a new model called C2 to simplify the current annual process.

Development of Common Activity (CoA)

In the C2 model, a single common activity (CoA) is used by all UMUC's Graduate School programs to assess four SLEs (COMM, THIN, INFO, and TECH). The CoA requires that students respond to a question in a short essay format to demonstrate their levels of performance in the four learning areas. Collaboratively developed with representatives of all the Graduate School departments, the question relates to commonly addressed program themes (i.e., technology globalization and leadership) and does not require prior knowledge of the topic. The CoA instructions present the essay question, clearly describe for students the steps for completing the task, and explain how the submission will be evaluated. Of note, the SLE, KNOW, was excluded from the model design. While it is a learning outcome expected of all students in the Graduate School, it is viewed as very program/discipline-specific and therefore, more appropriately assessed by other means, which may include standardized exams or special projects.

Moskal and Leydens (2000) suggest that discussing differences in raters' scores helps improve reliability, as does making performance criteria more precise, though narrow criteria definitions may preclude applicability to other activities. Bresciani, Zelna and Anderson (2004) contend that norming ensures that raters understand the rubric in a similar manner, which promotes consistency in scoring, and thereby enhances reliability.

Design of Combined Rubric (ComR)

A new rubric (ComR) was designed for use in the C2 model by initially combining relevant and established criteria from the current rubrics used in the 3-3-3 model, excluding those related to knowledge (KNOW). The researchers remained committed to the use of analytic rubrics in the C2 model for the benefits cited previously, including their ability to present a continuum of performance levels, provide qualitative information on observed student performance, and the potential for tracking student progress (Simon & Forgette-Giroux, 2001). The ComR rubric removed overlaps between the four existing rubrics. The steps in the design of the ComR involved:

- Consolidation of individual rubrics from four SLEs (COMM, THIN, TECH, INFO) into a single rubric (ComR) with fourteen criteria
- Review and revision based on feedback from the Graduate School Assessment Committee
- Use of ComR in Phase I to test content validity and alignment between ComR and the CoA
- Review and revision based on feedback from raters in Phase I to further consolidate ComR into nine criteria
- Application of the refined ComR in Phase II

The ComR rubric employed in Phase I is presented in Appendix C and Appendix D shows the refined ComR rubric used in Phase II.

The C2 model was designed to provide the means to evaluate multiple SLEs simultaneously and to score the common activity (CoA) by trained raters. Table 1 contrasts the new C2 model with the current 3-3-3 model.

Allocation of Resources

The primary resource needed for the development of the C2 model was time. The collaborative process took over a year from the time the idea was first proposed by the researchers to the Graduate School Assessment Committee to the time the pilot was conducted in Spring 2012. Fortunately, all members of the committee were in agreement that the existing 3-3-3 assessment model needed to be simplified and improved, therefore it did not take much convincing for them to agree to participate in the pilot. The most time expended was in the development of the common activity (CoA) and the combined rubric (ComR). The essay question for the CoA was developed over a period of several months until a consensus was reached across the Graduate School. The ComR was created through

Table 1

Comparison of the Current 3-3-3 Model and the Combined Activity/Rubric (C2) Model

Current 3-3-3 Model	Combined Activity/Rubric (C2) Model
Multiple Rubrics: one for each of 4 SLEs	Single rubric for all 4 SLEs
Multiple assignments across graduate school	Single assignment across graduate school
One to multiple courses/4 SLEs	Single course/4 SLEs
Multiple raters for the same assignment/course	Same raters/assignment/course
Untrained raters	Trained raters

an iterative process, which included sharing each draft edition and making adjustments until the committee was in agreement. Additional resources included a stipend paid to the seven hired raters trained for grading. The funds for the stipends were provided from a federal grant. These stipends resulted in a total cost of $7,000.

Implementation of C2 Model

The pilot study was conducted sequentially through two phases: Phase I and II. In Phase I, the ComR was used in three graduate programs to determine its reliability for grading the CoA. The three Masters' programs that were part of Phase I included Biotechnology, Master of Arts in Teaching, and Master of Education in Instructional Technology. The three programs were selected based on the interest and willingness of the degree program directors to participate in the pilot and their ability to easily incorporate the pilot activity into their courses. The CoA was explained in the syllabi of the courses selected for the pilot study and was scheduled to be completed during the first quarter of the semester.

Raters' Training and Norming

The C2 model appears to have simplified the assessment process. The new C2 assessment model implemented a common activity (CoA) and used a combined rubric (ComR) for the outcomes assessment process.

Adding trained raters to the C2 model was done for the purposes of simplifying faculty workloads and improving scoring consistency. Program directors were asked to suggest faculty who could act as raters for the pilot papers. The faculty raters needed to fit the following guidelines: they were *not* teaching any of the pilot courses in Spring 2012, had experience teaching and grading in the participating programs, and therefore could easily become 'raters' for the pilot study. All seven recommended faculty members were contacted and 100% agreed to participate in the study. Contracts for the faculty raters were discussed, signed and processed with an agreed-upon timeline for training, scoring procedures and follow-up.

A total of 91 students completed the activity. The papers were collected, redacted of any identifiable information, and assigned a code number prior to being distributed to the raters. Raters were given a set of anchor papers, selected from the submissions, which provided the raters with samples of varying levels of student performance (Tierney & Simon, 2004). To strengthen reliability and yield a consistency in grading with the rubric, raters were required to participate in norming sessions (Trochim & Donnelly, 2006) prior to and after the scoring of the anchor papers. Since raters were geographically dispersed, the norming sessions were conducted online, both asynchronously and synchronously, to allow for flexibility and scalability. All raters actively engaged in the training and norming sessions, which provided them with the opportunity to practice scoring anchor papers and discuss in detail the interpretation and application of the combined rubric for grading. Moskal and Leydens (2000) suggest that discussing differences in raters' scores helps improve reliability, as does making performance criteria more precise, though narrow criteria definitions may preclude applicability to other activities. Bresciani, Zelna and Anderson (2004) contend that norming ensures that raters understand the rubric in a similar manner, which promotes consistency in scoring, and thereby enhances reliability.

Papers were assigned to raters in a discipline-specific manner in Phase I such that the raters from the Education department received and scored papers from students

in Education, while raters from Biotechnology graded papers from the Biotechnology program course.

Inter-rater Reliability

The pilot norming results emphasized the importance of providing a range of anchor papers that represented different levels of student performance in order to determine and discuss baseline scoring.

In this study, each paper was randomly assigned to two independent raters and graded by them using the same scoring rubric. This process is called coding because the raters are creating the data when they assign scores (ratings) to each student paper. Stemler (2004) states that in any situation that involves judges (raters), the degree of inter-rater reliability is worthwhile to investigate, as the value of inter-rater reliability has significant implication for the validity of the subsequent study results. There are numerous statistical methods for computing a measurement estimate of inter-rater reliability (e.g., simple percent-agreement, Cohen's Kappa, generalizability theory, Pearson *r*, etc.) and each of them has advantages and disadvantages (Stemler, 2004). For example, Pearson *r* can be a useful estimator of inter-rater reliability only when one has meaningful pairings between two and only two raters (linear relationship between the two set of ratings). Cohen's Kappa is commonly used for calculating inter-rater reliability for qualitative (categorical) data (i.e., gender, age, education level, etc.). Its greatest advantage is taking into account chance agreement between two or more raters. However, Kappa assumes that all raters have similar training and experience. When raters have dissimilar training and experience, the Kappa statistic is likely to be underestimated (Crewson, 2005).

Intraclass Correlation Coefficients (ICC) were used in this study for the estimation of inter-rater reliability. An ICC is a measure of the proportion of a variance that is explained by the objects (i.e., students) of measurement (i.e., raters' ratings). ICC has advantages over bivariate correlation statistics, such as Pearson *r*, as it accounts for variability between multiple raters and among the multiple dimensions of the rubric. Reliability assessed by ICC is a scaled agreement index under ANOVA assumptions. As discussed in the works of McGraw and Wong (1996) and Shrout and Fleiss (1979), to select an appropriate form of the ICC, one has to make several decisions related to (a) which ANOVA model should be used to analyze the data (one-way or a two-way); (b) whether differences in raters' mean ratings relevant to the reliability of interest (ICC for consistency vs. absolute agreement) and (c) whether the unit of analysis is a mean of several ratings or single rating (ICC for average vs. single measurements).

In this study, each student paper was rated by a randomly selected group of two raters from a larger pool. In other words, the same two raters did not grade all the papers. No effort was made to disentangle the effects of the rater and student paper, but only the objects (students) were treated as a random factor. Therefore, a one-way random effects ANOVA model was used to calculate the ICC (measures of absolute agreement were selected, as consistency measures were not defined in this model). The "average measures" ICC was provided in the results, which indicates the inter-rater reliability when taking the mean of all ratings from multiple raters and multiple dimensions of the rubric. The ICC will approach 1.0 if there is less variance within item ratings. According to Nunnally (1978), an ICC of 0.7 is generally considered an acceptable level for the type of study employed in this pilot.

Multiphase Approach

The researchers anticipated that the development of the C2 model would be a process of continuous improvement. For this reason, Phase II was performed and lessons learned from Phase I were applied that included further refining the ComR based on feedback provided by the raters and modifying the pilot process. Refining the rubric involved eliminating what the raters determined were redundant or overlapping criteria and clarifying criteria descriptions. In terms of modifying the pilot process, the same set of papers and raters from Phase I were used in Phase II, but the raters were given different subsets of papers and the papers were *not* assigned in a discipline-specific manner. This modification was made to allow us to gain insight into how well raters would handle rating papers from different disciplines, which is an ultimate goal in the Graduate School's full implementation.

Results

In both Phases I and II, each paper was rated by two raters and the ICC was computed. Table 2 displays a value of 0.44 in ICC from the Phase I data, which means that approximately 44% of the time two independent raters assessed an item and then scored it with the same value. The ICC is lower than the generally acceptable level of 0.70. In an attempt to increase the relative low reliability (0.44) generated in Phase I, the authors refined and consolidated the ComR to remove redundancy, and thereby reduced the number of criteria from fourteen to nine. The authors carefully selected a different set of anchor papers than those used in Phase I that clearly represented different levels of student performance. In addition, in Phase II, a third rater was used for papers when the scores between two raters had discrepancies greater than 1 point in at least 3 criteria. These extreme scores were discarded before calculating the Phase II ICC.

Table 2

Average Measures of ICC – Phase I & II

	Intraclass Correlation Coefficients	
	Phase I	Phase II
Average Measures	0.44	0.75

By implementing the refinements and consolidations to the rubric and common activity, Phase II ICC provided a value of 0.75, meaning approximately 75% of the time two independent raters assessed an item and then gave it the same score (Table 2). Since the ICC for Phase II reached the generally acceptable level (0.70) of agreement among these raters, it provided confidence in the reliability of the C2 model.

Discussion

As mentioned earlier, the present 3-3-3 Graduate School assessment model has some limitations. One of those is the increased faculty workload of grading a wide variety of assignments that are used for assessment across the Graduate School programs. With the 3-3-3 model, there can also be grading inconsistency and weak alignment between the assignment and the rubrics.

The C2 model appears to have simplified the assessment process. The new C2 assessment model implemented a common activity (CoA) and used a combined rubric (ComR) for the outcomes assessment process. It also addressed the concerns with the current 3-3-3 model in that it:

- shifted the faculty grading workload to external, trained raters,
- incorporated training and norming sessions to improve rubric consistency and use,
- eliminated assignment disparities by employing one common activity across the Graduate School, and
- provided tighter alignment between the assignment and rubric.

Instructors often feel a pressure to align assessment scores with assignment grades, whereas raters can focus solely on the criteria under assessment. External raters may also possess more knowledge and understanding of the specific criteria under assessment.

Rezaee and Kermani (2011) write that "raters' inconsistencies in scoring can be attributed to different factors among which are raters' gender, age, teaching experience, scoring experience, first language and scoring environment" (p. 109). Furthermore, Bresciani et al. (2004) report that low reliability among raters may be influenced by the (a) objectivity of the task or scoring, (b) complexity of the task, (c) group homogeneity of the raters, (d) work pace of the raters, and (e) number of assignments scored. A lower agreement among raters may result from various reasons such as ambiguity of the rubric criteria and activity instructions, misunderstanding of rubric criteria, preconceived notions held by raters, and using a small pool of raters. In Phase I the ICC of .44 was lower than the generally acceptable .70 level, indicating the potential presence of such issues for the participating raters. In Phase II, the authors addressed some of these issues in an attempt to improve the inter-rater reliability, the results of which, was an improved ICC of .75.

Although the effect of norming on inter-rater reliability may be disputed, the researchers recognized the importance of the norming process for refining the rubric and the activity. The pilot norming results emphasized the importance of providing a range of anchor papers that represented different levels of student performance in order to determine and discuss baseline scoring. Rater feedback during the norming process also informed further rubric consolidation. The iterative process of refining the CoA and ComR worked toward ensuring that the criteria for each SLE were discrete, not dependent on each other and directly assessable. As a result, the original combined rubric (ComR) with fourteen criteria was consolidated further in Phase II to nine criteria, again simplifying the use of the rubric and potentially contributing to better application and agreement among raters.

In addition, there appears from the pilot to be benefits in using external raters to score assessment activities as opposed to the teaching faculty. Instructors often feel a pressure to align assessment scores with assignment grades, whereas raters can focus solely on the criteria under assessment. External raters may also possess more knowledge and understanding of the specific criteria under assessment. In addition, providing a potential point of comparison between rater and teacher evaluations may serve in evaluating assessment findings.

Limitations of this Study

Even though the main goals of this pilot study were met and simplification of the current Graduate School assessment process seems promising, there are limitations to this study and future research is needed to address them.

The use of a single assignment and rubric to evaluate multiple competencies may be construed as a limitation. As Maki (2004) points out, "Relying on one method to assess the learning described in outcome statements restricts interpretations of student achievement within the universe of that method" (p.156); using multiple measures to assess different learning outcomes, on the other hand, has its advantages. However, others have explored the possibility of combining various rubrics to evaluate multiple learning outcomes based on a single student assignment (Stanny & Duer, 2012). In addition, just as the trained raters provided feedback for the rubric in Phase I of this pilot study, the researchers expect to continue to receive feedback for further refinements in future phases of our studies.

This model is an attempt to improve the comparability of the data across programs, increase clarity of the process, decrease faculty workload, and therefore greatly simplify the outcomes assessment process.

Another limitation may result from the design of the study. In this pilot study no two raters graded all the same papers. This was intentional as eventually a pool of raters will be expected to grade all the papers that come out of the Graduate School. Having the same two or more raters grade all the papers will not be practical for implementation purposes. Consequently, one-way (or one-factor) random effect ANOVA model using objects (students) as the only effect was used to calculate ICCs. This approach limited the ability to evaluate the rater effect as a variable because specific raters and the interactions of raters with students were not disentangled. Intra-rater reliability, a measure of the rater's self-consistency, also could not be defined in this study.

Conclusions and Further Studies

This study describes the implementation of a unique assessment model, C2. Our findings indicate that this model may have a higher rate of reliability than the Graduate School's current 3-3-3 model. Using the C2 model, several core learning competencies may be assessed simultaneously through a common assignment, a combined rubric, and trained raters across different graduate programs. This model is an attempt to improve the comparability of the data across programs, increase clarity of the process, decrease faculty workload, and therefore greatly simplify the outcomes assessment process. To evaluate both object (student) and rater effects, either the two-way random or mixed effects model, in which each student paper is rated by the same group of raters, may be used in future studies.

In order to further improve on the reliability of scores from the common activity and the combined grading rubric, Phase III of the C2 model will be applied to several

programs across the Graduate School in Fall 2012 in preparation for a potential graduate school-wide implementation. Post graduate school-wide implementation, the authors will focus on methods to establish the validity of the C2 model.

References

Banta, T. W. (2002). *Building a scholarship of assessment.* San Francisco, CA: Jossey-Bass.

Bresciani, M.J. (2011). Identifying barriers in implementing outcomes-based assessments program review: A grounded theory analysis. *Research and Practice in Assessment, 6*, 5-16. Retrieved from http://www.rpajournal.com/archive/

Bresciani, M.J., Zelna, C.L., & Anderson, J.A. (2004). *Assessing student learning and development: A handbook for practitioners.* Washington, DC: National Association of Student Personnel Administrators (NASPA).

Buzzetto-More, N.A., & Alade, A. J. (2006). Best practices in e-assessment. *Journal of Information Technology Education, 5*, 251-269. Retrieved from http://informingscience.org

Crewson, P. (2005). Fundamentals of clinical research for radiologists: Reader agreement studies. *American Journal of Roentgenology, 184*(5), 1391-1397. Retrieved from http://www.ajronline.org

Maki, P. L. (2004). *Assessing for learning: Building a sustainable commitment across the institution.* Sterling, VA: Stylus.

Maki, P. L. (2010). *Coming to terms with student outcomes assessment: Faculty and administrators' journeys to integrating assessment in their work and institutional culture*. Sterling, VA: Stylus.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. Retrieved from http://psycnet.apa.org

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25). Retrieved from http://pareonline.net/

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment Research & Evaluation, 7*(3). Retrieved from http://pareonline.net/

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubrics development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(10). Retrieved from http://pareonline.net/

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

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

Petkov, D., & Petkova, O. (2006). Development of scoring rubrics for IS projects as an assessment tool. *Issues in Informing Science and Information Technology Education, 3*, 499-510. Retrieved from http://proceedings.informingscience.org

Prus, J., & Johnson, R. (1994). A critical review of student assessment options. *New Directions for Community Colleges*, *88*, 69-83.

Rezaee, A. A., & Kermani, E. (2011). Essay rater's personality types and rater reliability. *International Journal of Language Studies, 5*(4), 109-122. Retrieved from www.ijls.net

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2*, 420-428. Retrieved from http://psycnet.apa.org

Simon, M., & Forgette-Giroux, R. (2001). A rubric for scoring postsecondary academic skills. *Practical Assessment, Research & Evaluation, 7*(18). Retrieved from http://pareonline.net/

Stanny, C. J., & Duer, J. D. (2012). Authentic assessment in psychology: Using rubrics and embedded assessments to improve student learning. In D. S. Dunn, S. C. Baker, C. M. Mehrotra, R. E. Landrum, & M. A. McCarthy (Eds.), *Assessing teaching and learning in psychology* (pp. 19-34). Belmont, CA: Wadsworth, Cengage.

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://pareonline.net/

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research and Evaluation, 9(2*). Retrieved from http://pareonline.net/

Trochim, M.K., & Donnelly, J. (2006). *Research methods knowledge base*. New York, NY: Cornell University.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Walvoord, B. (2004). *Assessment clear and simple: A practical guide for institutions, departments, and general education*. San Francisco, CA: Jossey-Bass.

Appendix A

UMUC Graduate School Student Learning Expectations (SLEs)

STUDENT LEARNING EXPECTATIONS (SLEs)	
Written Communication (COMM)	Produce writing that meets expectations for format, organization, content, purpose, and audience.
Information Literacy (INFO)	Demonstrate the ability to use libraries and other information resources to effectively locate, select, and evaluate needed information.
Critical Thinking (THIN)	Demonstrate the use of analytical skills and reflective processing of information.
Technology Fluency (TECH)	Demonstrate an understanding of information technology broad enough to apply technology productively to academic studies, work, and everyday life.
Content/Discipline-Specific Knowledge (KNOW)	Demonstrate knowledge and competencies specific to program or major area of study.

Appendix B

Rubric for 3-3-3 Assessment Model

University of Maryland University College
Graduate School
Writing Rubric for Outcomes Assessment Spring 2012

CRITERIA	EXEMPLARY 4	COMPETENT 3	MARGINAL 2	UNSATISFACTORY 0-1	SCORE
Context/Purpose Considers the audience, purpose, and the circumstances surrounding the writing assignment(s).	Shows **superior understanding** of context, audience, and purpose that is **extremely appropriate** for the assignment(s).	Shows **good understanding** of context, audience, and purpose that is **mostly appropriate** for the assignment(s).	Shows **fair understanding** of context, audience, and purpose that is **somewhat appropriate** for the assignment(s).	Shows **insufficient** or **poor understanding** of context, audience, or purpose of the assignment(s).	
Content/Ideas/Support Articulates and supports a main idea(s) that is consistent with context and purpose.	**Highly original** main idea(s) is **clearly articulated** and **strongly supported** by **predominantly current and relevant** evidence that may be researched based. Main idea(s) is **exceedingly consistent** with context and purpose.	**Mostly original** main idea(s) is **generally well articulated** and **sufficiently supported** by **mainly current and relevant** evidence that may be researched based. Main idea(s) is **generally consistent** with context and purpose.	Main idea(s) is **vague**, and/or **inadequately supported**, and/or **inconsistent** with context and purpose.	Main idea(s) is **hardly** or **not evident** and/or **lacks support** and/or **scarcely relates** to context and purpose.	
Organization Uses logical sequencing including introduction, transitions between paragraphs, and summary/ conclusion to develop main idea(s) and content.	Uses **highly logical sequencing** including introduction, transitions between paragraphs, and summary/ conclusion to **fully develop** main idea(s) and content.	Uses **mostly logical sequencing** including introduction, transitions between paragraphs, and summary/ conclusion to **generally develop** main idea(s) and content.	Uses **partially logical sequencing. Makes inadequate use** of introduction, and/or transitions between paragraphs, and/or summary/ conclusion. Main idea(s) and content are **incompletely developed**.	Uses **little** or **no logical sequencing**. Lacks introduction, and/or transitions between paragraphs and/or summary/ conclusion. Main idea(s) and content **remain undeveloped**.	
Sources Incorporates use of and identifies sources and/or research, according to APA and/or instructor guidelines.	Demonstrates **superior judgment** in selection, incorporation, and identification of **entirely appropriate quality and quantity** of sources and/or research that **fully meet or exceed** established guidelines.	Demonstrates **good judgment** in selection, incorporation, and identification of **mainly appropriate quality and quantity** of sources and/or research that **mostly meet or exceed** established guidelines.	Demonstrates **limited judgment** in selection and/or incorporation and/or identification of sources and/or research. Quality and/or quantity and/or appropriateness **partially meet** established guidelines.	Demonstrates **little** or **no judgment** in selection and/or incorporation and/or identification of sources and/or research. Quality and/or quantity and/or appropriateness **do not meet** established guidelines.	
Word Usage/ Grammar/Spelling/ Punctuation Uses wording, grammar, spelling and punctuation accurately and correctly.	Demonstrates **virtually error-free** grammar, spelling and punctuation.	Demonstrates **very few errors** in grammar, spelling and punctuation.	Demonstrates **numerous errors** in grammar, spelling and punctuation.	Demonstrates **unacceptable amount and/or type of errors** in grammar, spelling and punctuation.	

Appendix C

ComR Rubric for Phase I

University of Maryland University College
Graduate School of Management and Technology
COMBINED Rubric for Outcomes Assessment for Spring 2012

CRITERIA	EXEMPLARY 3.1-4.0	COMPETENT 2.1-3.0	MARGINAL 1.1-2.0	UNSATISFACTORY 0-1.0	Score
Conceptualization Identifies and describes nature of idea(s) or issue(s) in relation to research and assignment context.	Shows a superior ability to identify and describe basic and complex issues with exceptional depth and clarity within context for full understanding.	Shows a good ability to identify and describe basic and complex issues with sufficient depth and clarity within context. Omissions do not seriously impede understanding.	Shows fair ability to identify and describe basic and complex issues within context with some depth and clarity. Ambiguities and omissions impede understanding.	Shows insufficient or no ability to identify basic and complex issues. Lack of clarity or depth impedes understanding.	0.00
Analysis Considers pros/cons; compares/contrasts in logical examination of issue(s) and source data.	Analyzes information in a highly organized and logical manner. Is exceptionally consistent and accurate in identifying embedded hypotheses, biases, causalities, and conclusions.	Analyzes information in a mostly organized and logical manner. Is generally consistent and accurate in identifying embedded hypotheses, biases, causalities, and conclusions.	Analyzes information in a somewhat organized and logical manner. Is slightly inconsistent and/or inaccurate in identifying embedded hypotheses, biases, causalities, and conclusions.	Analyzes information in a disorganized and illogical manner. Is inconsistent and/or inaccurate in identifying embedded hypotheses, biases, causalities, and conclusions.	0.00
Synthesis Integrates key concepts from research and analyses in coherent manner to form a cohesive response.	Consistently incorporates analyses with other information to connect key concepts in a highly coherent way. Provides strong base for further application and perspective.	Usually incorporates analyses with other information to connect key concepts in a mostly coherent way. Provides adequate base for further application and perspective.	Occasionally incorporates analyses with other information to connect key concepts in a partially coherent way. Provides minimal base for further application and perspective.	Rarely or never incorporates analyses with other information to connect key concepts. Work is incoherent. Provides no base for further application and perspective.	0.00
Conclusion Integrates analysis and synthesis to formulate a new perspective or position that is appropriate to the conceptualization of the question or assignment.	Integrates prior criteria in a highly effective manner demonstrating an original, well-reasoned, and justifiable perspective(s).	Integrates prior criteria in a mostly effective manner demonstrating a generally original, well-reasoned, and justifiable perspective(s).	Integrates prior criteria in a partially effective manner demonstrating weakness in originality, reasoning, and justifiable perspective(s).	Integrates prior criteria in an ineffective manner. Lacks an original, well-reasoned, or justifiable perspective(s).	0.00
Implications Based upon the positions, perspectives or conclusions, determines practices or processes and/or the need for further study.	Suggests highly appropriate considerations or actions for practice, policy and future research.	Suggests mostly appropriate considerations or actions for practice, policy and future research.	Suggests somewhat appropriate considerations or actions for practice, policy and future research.	Suggests inappropriate or fails to make considerations or actions for practice, policy and future research.	0.00
Evaluation Identifies appropriate resources by critically assessing reputation and quality of information.	Thoroughly analyzes information sources for currency, relevance, accuracy, authority and objectivity.	Sufficiently analyzes information sources for currency, relevance, accuracy, authority and objectivity.	Partially analyzes information sources for currency, relevance, accuracy, authority and objectivity.	Insufficiently analyzes information sources for currency, relevance, accuracy, authority and objectivity.	0.00
Incorporation Use information to accomplish specific purpose.	Expertly synthesizes and presents information to fully achieve a specific purpose with clarity and depth.	Sufficiently synthesizes and presents information to fully achieve a specific purpose with some clarity and depth.	Partially synthesizes and presents information with little clarity or depth.	Inadequately synthesizes and presents information with little or no clarity or depth.	0.00
Ethical Use Understands and complies with institutional policies related to access and use of information, demonstrating academic integrity.	Fully demonstrates understanding of ethical and legal guidelines for published, confidential and proprietary information.	Mostly demonstrates understanding of ethical and legal guidelines for published, confidential and proprietary information.	Partially demonstrates understanding of ethical and legal guidelines for published, confidential and proprietary information.	Fails to demonstrate understanding of ethical and legal guidelines for published, confidential and proprietary information.	0.00
Context/Purpose Considers the audience and purpose of the assignment.	Shows superior understanding of context, audience, and purpose that is extremely appropriate for the assignment(s).	Shows good understanding of context, audience, and purpose that is mostly appropriate for the assignment(s).	Shows fair understanding of context, audience, and purpose that is somewhat appropriate for the assignment(s).	Shows insufficient or poor understanding of context, audience, or purpose of the assignment(s).	0.00
Content/Ideas/Support Articulates and supports a main idea(s) that is consistent with context and purpose.	Highly original main idea(s) is clearly articulated and strongly supported by predominantly current and relevant evidence that may be researched based. Main idea(s) is exceedingly consistent with context and purpose.	Mostly original main idea(s) is generally well articulated and sufficiently supported by mainly current and relevant evidence that may be researched based. Main idea(s) is generally consistent with context and purpose.	Main idea(s) is vague, and/or inadequately supported, and/or inconsistent with context and purpose.	Main idea(s) is hardly or not evident and/or lacks support and/or scarcely relates to context and purpose.	0.00
Organization Uses logical sequencing as required of the assignment to develop main ideas and content.	Uses highly logical sequencing including introduction, transitions between paragraphs, and summary/ conclusion to fully develop main idea(s) and content.	Uses mostly logical sequencing including introduction, transitions between paragraphs, and summary/ conclusion to generally develop main idea(s) and content.	Uses partially logical sequencing. Makes inadequate use of introduction, and/or transitions between paragraphs, and/or summary/ conclusion. Main idea(s) and content are incompletely developed.	Uses little or no logical sequencing. Lacks introduction, and/or transitions between paragraphs and/or summary/ conclusion. Main idea(s) and content remain undeveloped.	0.00
Grammar/Spelling/ Punctuation Uses wording, grammar, spelling and punctuation accurately and correctly.	Demonstrates virtually error-free grammar, spelling and punctuation.	Demonstrates very few errors in grammar, spelling and punctuation.	Demonstrates numerous errors in grammar, spelling and punctuation.	Demonstrates unacceptable amount and/or type of errors in grammar, spelling and punctuation.	0.00
Technology Management Creates accurate electronic document with appropriate layout, formatting, and accuracy.	Shows exceptional skills in creating accurate electronic documents appropriate for the assignment or context.	Shows good skills in creating accurate electronic documents appropriate for the assignment or context.	Shows fair skills in creating accurate electronic documents appropriate for the assignment or context.	Shows minimal or no skills in creating accurate electronic documents appropriate for the assignment or context.	0.00
Information Retrieval Utilizes technology to research, evaluate, inform, and communicate information retrieved from appropriate resources.	Uses technology extremely effectively to research, evaluate, inform, and communicate information from very appropriate resources.	Uses technology very effectively to research, evaluate, inform, and communicate information from mostly appropriate resources.	Uses technology somewhat effectively to research, evaluate, inform, and communicate information from appropriate resources.	Uses technology ineffectively or not at all to research, evaluate, inform, and communicate information from often inappropriate resources.	0.00

Appendix D

ComR Rubric Phase II

University of Maryland University College

COMBINED Rubric for Outcomes Assessment for Spring 2012, The Graduate School

CRITERIA	EXEMPLARY 3.1-4.0	COMPETENT 2.1-3.0	MARGINAL 1.1-2.0	UNSATISFACTORY 0-1.0	Score
Conceptualization/Content/Ideas Identifies and articulates the main idea(s) or issue(s) in a way that is appropriate for the audience, research, context, and purpose of the assignment.	Identifies and articulates the main ideas/issues as appropriate with **exceptional** depth and clarity for full understanding with no ambiguities.	Identifies and articulates the main ideas/issues as appropriate with **sufficient** depth and clarity. Ambiguities and omissions do not seriously impede understanding.	Identifies and articulates the main ideas/issues within context with **some** depth and clarity. Ambiguities and omissions impede understanding.	**Insufficiently** identifies and articulates the main ideas/issues. Lack of clarity or depth impedes understanding.	0.00
Analysis/Evaluation Determines essential components and characteristics of the idea(s) or issue(s) while considering connections and significance.	Examines information in a **highly** logical and accurate manner and **extensively** exposes relationships, causalities, and importance of the ideas/issues.	Examines information in a **mostly** logical and accurate manner and **sufficiently** exposes relationships, causalities, and importance of the ideas/issues.	Examines information in a **somewhat** logical and accurate manner and **insufficiently** exposes relationships, causalities, and importance of the ideas/issues.	Examines information in an **illogical** and **inaccurate** manner and **fails** to expose relationships, causalities, and importance of the ideas/issues.	0.00
Synthesis /Support Integrates key concepts from research and analyses in a coherent manner to form a cohesive response.	**Consistently** incorporates analyses with other information/research to connect key concepts in a **highly** coherent way.	**Usually** incorporates analyses with other information/research to connect key concepts in a **mostly** coherent way.	**Occasionally** incorporates analyses with other information/research to connect key concepts in a **partially** coherent way.	**Rarely or never** incorporates analyses with other information/research to connect key concepts. Work is **incoherent**.	0.00
Conclusion/Implications Formulates a new perspective or position based upon consequences for practice, policy and/or the need for future study.	Forms a conclusion in a **highly** effective manner demonstrating an original, well-reasoned, and justifiable perspective(s) that **extensively** considers potential implications.	Forms a conclusion in a **mostly** effective manner demonstrating a generally original, well-reasoned, and justifiable perspective(s) that **sufficiently** considers potential implications.	Forms a conclusion in a **partially** effective manner demonstrating weakness in originality, reasoning, and justifiable perspective(s) that **insufficiently** considers potential implications.	Forms a conclusion in an **ineffective** manner. Lacks an original, well-reasoned, or justifiable perspective(s) with **no** consideration of potential implications.	0.00
Selection/Retrieval Chooses appropriate resources identified through online searches and critically assesses the quality of the information according to the criteria in the assignment.	Displays **thorough** evidence that information sources were chosen and assessed according to assignment expectations.	Displays **mostly complete** evidence that information sources were chosen and assessed according to assignment expectations.	Displays **incomplete** evidence that information sources were chosen and assessed according to assignment expectations.	Displays **very little** or **no** evidence that information sources were chosen and assessed according to assignment expectations.	0.00
Organization Uses logical sequencing as required of the assignment to develop the main ideas and content.	Uses **highly** logical sequencing including introduction, transitions between paragraphs, and summary/conclusion to **fully** develop the main idea(s) and content.	Uses **mostly** logical sequencing including introduction, transitions between paragraphs, and summary/conclusion to **generally** develop the main idea(s) and content.	Uses **partially** logical sequencing. Makes **inadequate** use of introduction, and/or transitions between paragraphs, and/or summary/conclusion. Main idea(s) and content are **incompletely** developed.	Uses **little** or **no** logical sequencing. Lacks introduction, and/or transitions between paragraphs and/or summary/ conclusion. Main idea(s) and content remain **undeveloped**.	0.00
Writing Mechanics Uses wording, grammar, spelling and punctuation accurately and correctly.	Contains **virtually no errors** in grammar, spelling and punctuation; any errors in writing mechanics and word usage do not interfere with reading or message.	Demonstrates **some** errors in grammar, spelling, punctuation and/or word usage that **somewhat** interfere with reading or message.	Demonstrates **numerous** errors in grammar, spelling, punctuation and/or word usage. These errors distract from the reading and weaken the message.	Demonstrates **excessive** errors in grammar, spelling, punctuation and word usage. These errors display an inability to communicate the message.	0.00
APA Compliance Follows APA style that includes headings, citations and a reference page.	Employs **very** accurate APA style.	Employs **mostly** accurate APA style.	Employs **mostly inaccurate** APA style.	Employs **little or no** APA style.	0.00
Technology Application Creates accurate electronic document according to specifications of the assignment.	Creates an electronic document that **complies with all** of the assignment specifications.	Creates an electronic document that **mostly complies with** the assignment specifications.	Creates an electronic document that **partially complies** with the assignment specifications.	Creates an electronic document that minimally complies or **shows no evidence of compliance** with the assignment specifications.	0.00

Book Review

Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education (2nd ed.). Alexander W. Astin and anthony lising antonio. Lanham, MD: Rowman & Littlefield and the American Council on Education, 2012. 380 pp. ISBN-13:978-1442213620, paperback, $55.00

REVIEWED BY:
Linda J. Sax, Ph.D.
University of California, Los Angeles

Twenty years ago I was fortunate enough to have read the first edition of *Assessment for Excellence* (Astin, 1991) while enrolled in a graduate course on assessment taught by Alexander Astin. I recall appreciating the book's point of view, conversational tone, clear explication of methods, and thought-provoking commentary. I relied extensively on that book in order to learn the key conceptual and methodological issues surrounding assessment, and for the last eighteen years have used it as the primary textbook in my graduate courses on assessment and evaluation in higher education.

In recent years, however, I noticed that although the basic principles and methods detailed in the first edition were as relevant as they were when the book was first published, the assessment context had changed. The first edition was written just after the assessment movement gained steam in the 1980s, a time when colleges and universities were expanding their student assessment activities in response to mounting criticism of higher education's assessment efforts and growing state and federal demands on colleges to demonstrate student outcomes. In the original book, Astin noted that "although a great deal of assessment activity goes on in America's colleges and universities, much of it is of very little benefit to either students, faculty, administrators, or institutions. On the contrary, some of our assessment activities seem to conflict with our most basic educational mission" (Astin, 1991, p. ix).

In the book's second edition, co-written by Astin and anthony antonio, that very same argument is made. This alone is a statement on how little our philosophical approach to assessment has evolved (even as new methodological approaches have proliferated). The current edition of *Assessment for Excellence* is set in the contemporary context during which interest in student assessment has intensified, especially interest in accountability and assessment of learning outcomes (e.g., the Spellings Report [U.S. Department of Education, 2006]). Astin and antonio argue that the current literature on assessment remains problematic both because it lacks coherence and because it is not useful in practice.

In many ways, this book offers an antidote to such deficiencies. Though heavy with critique (and even a hint of exasperation at how little has changed), the book makes an important contribution to the literature by weaving the philosophy and psychology surrounding assessment together with actionable approaches for data collection, analysis and dissemination. Just like the first edition, the second edition is an invaluable resource for individuals who are involved in planning, conducting or utilizing college student assessment.

In the original book, Astin noted that "although a great deal of assessment activity goes on in America's colleges and universities, much of it is of very little benefit to either students, faculty, administrators, or institutions. On the contrary, some of our assessment activities seem to conflict with our most basic educational mission. In the book's second edition that very same argument is made. This alone is a statement on how little our philosophical approach to assessment has evolved (even as new methodological approaches have proliferated).

In Chapter 1, Astin and antonio argue that although "the basic purpose of assessing students is to enhance their educational development" (p. 5), this goal is generally not met by traditional assessment practice. Instead, they describe assessment as too often driven by narrow definitions of excellence that prioritize an institution's resources and reputation rather than the college's effectiveness at developing the talents of its students. As was a central message of the first edition, this edition calls for assessment practices that enable institutions to know how their specific curricula, practices, and programs make a difference in their students' cognitive and affective development.

Their recommended conceptual approach is the Input-Environment-Outcome model for assessment, which provides a framework for thinking about how college affects students. As detailed in Chapter 2, a central premise of the model is that one must consider the characteristics of students prior to their exposure to college (known as "inputs") before presuming that college environments or experiences have an effect on student outcomes. They emphasize that college "outcomes" include a wide variety of student characteristics ranging from cognitive to affective and from psychological to behavioral so that the multi-dimensional nature of college impact may be considered ("outcomes" elaborated in Chapter 3). They point out that the range of inputs also can be quite broad, including family background, pre-college skills, abilities, goals, aspirations, and values ("inputs" elaborated in Chapter 4). College "environments" also reflect a broad range, from structural characteristics of institutions (e.g., size, type and selectivity, or characteristics of the peer and faculty environment) to student-determined environments, such as academic engagement, interactions with family and friends,

employment, and extracurricular activities ("environments" elaborated in Chapter 5). Explication of the I-E-O model is especially useful for individuals involved in the design of assessment studies because they encourage the reader to think broadly and creatively about what constitutes an educational environment and which outcomes are (and ought to be) valued.

Astin and antonio argue that the current literature on assessment remains problematic because it lacks coherence and because it is not useful in practice.

A particularly useful section in Chapter 2 is the discussion of "Incomplete Designs." These include: Outcome-Only Assessment (e.g., using a final exam to assess what students "know" as opposed to what they have "learned"); Environment-Outcome Assessment (e.g., presuming that between-institution variations in degree attainment are a reflection of each college's "effectiveness" at retaining students); Input-Outcome Assessments (i.e., presuming that student change and growth during college is a function of college attendance); and Environment-Only Assessment (e.g., equating well-resourced colleges or highly productive faculty with "quality" learning environments). Discussion of these incomplete designs is instructive because all readers will presumably recognize these approaches at their own institutions, but perhaps do not realize the limitations they present.

Whereas Chapters 2, 3, 4, and 5 give readers a framework for thinking about assessment, Chapter 6 helps readers understand how assessment data can be analyzed. Astin and antonio describe the purpose and application of basic statistical methods, both descriptive (e.g., two-way cross-tabulations, correlations) and causal (e.g., cross-tabulations with three or more variables, regression analysis). All concepts are presented in a manner that can be understood by people with little or no prior knowledge of statistics. Further, this edition incorporates a new section that distinguishes among five statistical techniques that can be used to conduct I-E-O based analyses: simple multiple regression, hierarchical blocked multiple regression, step-wise blocked multiple regression, structural equation modeling, and multilevel modeling. As an instructor who incorporates all of these methods into teaching about assessment, this new section is invaluable.

Chapter 7 shifts gears by moving away from a technical presentation of methods and into a discussion of how assessment results can be made useful. Astin and antonio emphasize that in order to be useful, assessment results should inform practitioners about the connections between what they are doing in practice and how that relates to student outcomes. In other words, it is not enough to know that students rate favorably on achievement, satisfaction or other college outcomes without understanding whether (and how) such outcomes are the result of anything the college has done to facilitate such positive results. They caution, however, that even when such knowledge is communicated—whether to faculty, administrators or students—it does not always translate directly into practice because of the resistance that may surface within these campus constituencies. As such, they describe eight different strategies (referred to as "academic games") that faculty and administrators use as a means of dismissing assessment results. Whether readers practice assessment or not, they will likely be amused by some of these descriptions since they have probably witnessed one or more such tactics in their own professional lives. Who among us has not witnessed "passing the buck" on an issue by establishing a committee to examine it further? And we have probably all experienced colleagues who effectively discredit results by raising questions about validity and reliability. These are such astute observations of academic life which, although presented in the context of assessment, are transferable to all academic contexts.

Once previous chapters have set the stage for how assessment can be conceptualized, conducted, and made useful, Chapter 8 provides readers with concrete examples of what might be contained in a student assessment database. Though some readers might be tempted to skip this chapter if they are not directly involved in the technical aspects of assessment, Astin and antonio caution that all parties should understand how these databases are constructed so that they can appreciate the challenges and possibilities of assessment. Importantly, this edition includes a substantial new section that reviews five major assessment programs: the Cooperative Institutional Research Program (CIRP), the National Survey of Student Engagement (NSSE), the Beginning College Survey of Student Engagement (BCSSE), the Collegiate Learning Assessment (CLA), and the Collegiate Assessment of Academic Progress (CAAP). Further, they discuss the usefulness of each of these survey programs in the context of talent development and the I-E-O model.

The book makes an important contribution to the literature by weaving the philosophy and psychology surrounding assessment together with actionable approaches for data collection, analysis and dissemination.

The next two chapters go into greater depth on key points made in Chapter 1, with Chapter 9 expanding on the ways in which assessment can provide direct feedback to the learner (whether students or faculty), and Chapter 10 focusing on how our current educational system reinforces socioeconomic inequities among students by emphasizing resources and reputations over talent development. Astin and antonio make a strong argument for a shift in our collective thinking about the purpose of higher education by encouraging us to consider who it is designed to benefit most. Readers familiar with the first edition may notice that little

has changed in terms of the book's argument that assessment practices have the potential to promote educational equity if they focus on maximizing talent development, but that current practices tend to reinforce inequitable conditions by giving institutions little incentive to admit students who may have the most to gain from college.

As mentioned above, this revised edition is important because it places the discussion about assessment in contemporary context. This is especially evident in Chapter 11, which argues that the assessment movement's growing emphasis on accountability places too much emphasis on student learning outcomes as opposed to the processes that might facilitate learning. Astin and antonio argue that "the true test of any state assessment policy is not whether it makes institutions more accountable but whether it serves to enhance the talent development function of its higher education institutions" (p. 238). Institutional leaders and state-level policy makers will be especially interested in this chapter's description of five assessment approaches and their potential for enhancing talent development. Astin and antonio propose an alternative approach which incentivizes state higher education systems to work cooperatively to maximize student learning and development in all sectors of the system, rather than competing against each other and compromising talent development.

If readers want an abridged version of the major points covered in all chapters, it is recommended that they read the summary contained in Chapter 12. This chapter reiterates the central messages of the book—that longitudinal assessment of students is essential if institutions are to take seriously their commitment to talent development, and that assessment practice must be designed so that students are better informed about what and how they learn, and educators and administrators are better informed about how their teaching practices and educational programming affect students. Though this chapter (and the entire book) is strong on critique, it also conveys hope. It reminds the reader that the future of assessment is in the hands of people who have the power to alter its course: "The capacity for higher education to be a positive change agent in the U.S. society will depend on our ability to transcend our institutional egos, our narcissism, and our self-interest, and to concern ourselves more directly with the impact we are having on our students and communities" (p. 275).

If readers want an abridged version of the major points covered in all chapters, it is recommended that they read the summary contained in Chapter 12.

One of the most valuable features of the book is the appendix, which provides readers with the building blocks for understanding regression, beginning with basic statistics and correlations, then moving through various regression-based techniques, and also offering more sophisticated concepts such as interaction effects and stepwise fluctuations in regression coefficients.

Now that the book is updated, those of us teaching graduate courses in assessment and evaluation have a vital instructional resource for many years to come. The book also remains invaluable for all individuals who will benefit from a philosophical understanding of assessment. Ultimately, this second edition of *Assessment for Excellence*, like the first, encourages us to think about student assessment—What should be assessed? What do we value?—and also gives readers the tools to engage in assessment that is grounded in the talent development perspective.

References

Astin, A. W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. New York, NY: American Council on Education and Macmillan Publishing Company.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education.* Washington, D.C.: Author.

Book Review

Reinventing Higher Education: The Promise of Innovation. Ben Wildavsky, Andrew P. Kelly, and Kevin Carey (Eds.). Cambridge, MA: Harvard University Press, 2011. 288 pp. ISBN-13: 978-1934742877. Paperback, $29.95.

REVIEWED BY:
Lisa J. Hatfield, MA, MAT
Portland State University

The last chapter of *Reinventing Higher Education: The Promise of Innovation* seems to sum up the book's premise best: make sure students are learning, and hold faculty and administrators accountable for that learning. The book's eight chapters are organized into three basic themes and offer numerous examples of innovative practices across the spectrum of private, public, and for-profit institutions. The book's themes include: a look at barriers to innovation in higher education; examples of innovations currently being implemented; and a glimpse into the future of non-traditional universities. The editors incorporate contributions from authors from academia as well as the private sector. The contributors who hail from academia hold posts either in or associated with schools of business, and many of the authors advocate tenets of a business model. The contributors represent institutions such as University of Southern California (USC) and Massachusetts Institute of Technology (MIT).

What exactly does it mean to be innovative? Dominic J. Brewer and William G. Tierney address that question in the first chapter. They define innovation as "a new method, custom, or device – a change in the way of doing things…Innovation is linked to creativity, risk taking, and experimentation, attributes that are often lacking in large, public, or non-profit organizations" (p. 15). Innovation, as exemplified in this text, occurs most readily when following the lead of the private sector.

They define innovation as "a new method, custom, or device – a change in the way of doing things…Innovation is linked to creativity, risk taking, and experimentation, attributes that are often lacking in large, public, or non-profit organizations."

This pro-business ethos of innovation is clearly supported throughout the book's chapters. For example, in the chapter titled "Creative Paths to Boosting Academic Productivity" William F. Massy likens the redesign of courses and their sequencing to business process reengineering used to increase productivity. Using the studio courses developed at Rensselaer Polytechnic Institute (RPI) in the 1990's as an example, Massy explains that the benefits of such courses extend past learning and teaching and into the realm of accounting. At RPI, it was more cost-effective to run one larger studio section utilizing technology than it was to have two smaller sections using a traditional lecture format. The RPI model is akin to the flipped classroom that many secondary and higher education institutions have been exploring. Another example of innovation tied to private business is given in the chapter "For-Profit Sector Innovations in Business Models and Organizational Cultures." Guilbert C. Hentschke writes that, unlike public and not-for-profit colleges and universities, for-profit higher education institutions often work with local and national employer advisory groups that listen to market performance to decide which programs to add and drop.

The private sector also is the foundation for journalist Jon Marcus's showcase of Harrisburg University of Science and Technology in Pennsylvania in the chapter "Old School: Four-Hundred Years of Resistance to Change." The for-profit institution, which has no tenure, utilizes corporate faculty from the high-tech sector as well as faculty who left tenure-track positions at other institutions. Ronald G. Ehrenberg continues this discussion of questioning the current tenure paradigm in the chapter "Rethinking the Professoriate." Capella University's faculty are judged by their students' success in achieving the institution's very specific outcomes; raises are based on performance evaluations rather than tenure status or union salary schedules. These innovative practices of evaluation and lack of tenure can also apply to traditional public and not-for-profit institutions. Ehrenberg gives the example of New York University, which has created a professional class of teaching faculty, a class deemed as equal to their research-focused peers. Public community colleges can also rethink expectations of instructors and use performance measures as one evaluative measure. Paul Osterman concludes that community colleges need to create systems that work not only toward a narrow mission but also are held accountable. "Forward progress," he writes, "requires additional resources that are aggressively linked to performance" (p. 158).

Discussion of evaluations based on the traditional teaching/research/service triumvirate continues in "Creative Paths to Boosting Academic Productivity" where Massy focuses on teaching and learning productivity, or what he calls "instructional productivity" (p. 74). Pursuing prestige through research poses a tension with teaching obligations, and so faculty tend to "satisfice" (p. 78) their teaching, meaning faculty do an average job of teaching to satisfy this piece of the tenure pie and then focus on the larger slice of research. He writes, "The implication of satisficing is 'Good enough is,' which stops continuous improvement in its tracks" (p. 78). The problem, Massy asserts, is that there is great difficulty in measuring the quality of higher educational instruction outputs, and that it is difficult to improve something one can't measure. He makes a strong point, though he himself acknowledges national efforts such as the National Survey of Student Engagement are being undertaken to begin addressing such measurement.

However, if institutions truly want to take undergraduate education seriously, they will place an emphasis on the quality of teaching.

This focus on teaching is not only seen in the physical classroom but in digital spaces as well. Peter Stokes, the executive vice president and chief research officer for the private research and consulting company Eduventures, advocates a decentering of faculty and a centering of students in the chapter "What Online Learning Can Teach Us about Higher Education." Stokes's emphasis on the positive disruption of the online environment in forcing educators to reconsider what we know about the traditional classroom and traditional learning is a loud message to hear. Some of this positive disruption is already taking place such as through massive open online courses (MOOCs) and courses administered through a flipped teaching model.

The clearest example of the need to reconsider traditional education models is seen in the book's last chapter, "The Mayo Clinic of Higher Ed" authored by editor Kevin Carey. He highlights the University of Minnesota-Rochester (UMR), "a campus based on the idea that most of what we know about how a public university should operate is wrong, and that it can be done better, for modest amounts of money, right away" (p. 226). UMR demonstrates innovative practices in teaching and tenure practices. UMR faculty from different disciplines collaborated to create a sequenced curriculum map, and the institution has a relationship with the nearby Mayo Clinic so doctors and researchers are guest lecturers, and students have access to surgical mannequins, Mayo Clinic labs, and other facilities. The senior year for UMR students is dedicated to a personalized capstone experience. Tenure at UMR is based on teaching, research in the academic disciplines, and research *about* teaching. These ideas are sound for effective learning and teaching, and, fortunately, some of these ideas are happening at other institutions as well. This concluding chapter brings together all the impactful innovations shared in the ones preceding it and shows that with visionary leadership, such positive impact on student learning can indeed happen in a public university.

If institutions truly want to take undergraduate education seriously, they will place an emphasis on the quality of teaching.

This focus on assessment of student learning, as reiterated in the final chapter, needs to underscore all innovative practices and provokes the reader to consider questions related to assessment of learning outcomes. Specifically readers may ask themselves, what should students be able to demonstrate to show success in learning and teaching or how should faculty be able to demonstrate their growth in learning and teaching? If decisions are indeed based on desired outcomes, higher education would be truly innovative.

Private businesses constantly assess and strive to improve their operations to ensure they earn a profit. This mentality, if applied appropriately to education, can undoubtedly help in nurturing and creating students who are learners. If one substitutes business operations and profits with outcomes and learning, the importance of constant review becomes more understandable. Large public institutions can take the good of a private business model and apply it to credit hours, teaching loads, and research requirements.

This focus on assessment of student learning, as reiterated in the final chapter, needs to underscore all innovative practices and provokes the reader to consider questions related to assessment of learning outcomes.

The foundational ideas of mapping UMR's curriculum that Carey shares may not be pervasive in higher education, but have been a part of K-12 education for years. Similarly, K-12 education focused at length on student learning outcomes, and now with the Common Core State Standards Initiative (2012), 45 states have agreed to work toward outcomes that ensure students are college and career ready. This false dichotomy of college or career also needs to be addressed by those in higher education. In the first chapter, Brewer and Tierney write, "Currently, we know very little about what works in college instruction and curriculum," (p. 38). However, many teaching and learning centers in higher education do know what works, and K-12 models also can be used as guides for what can work.

Assessment practitioners need to understand the practices and trends that exist outside their institutions, and this does not mean simply conducting an analysis of peer schools. As *Reinventing Higher Education* clearly underscores, assessment professionals should look beyond campus, explore what innovations are taking place in the private sector as well as in K-12 education, and apply the best from all sectors to students and their learning. This text seeks to present possibilities of some of these efforts, with the best example of holistic success happening at UMR.

The text would make an even more persuasive argument if it did not consistently make broad general assumptions. For example, some of the writers dismiss current instruction in higher education as purely "traditional" (i.e., lecture) and assume unfairly that students are being taught via rote memorization only. Another often-made generalization is the emphasis on prestige as a powerful driver in maintaining the status quo. True, the elite colleges are prestigious and perhaps always will be; however, according to the National Center for Education Statistics (2011), the

United States is home to nearly 4,600 public and private institutions of higher learning and only a few are considered prestigious. However, all of these institutions, prestigious or not, need to ensure students are learning.

Assessment practitioners need to understand the practices and trends that exist outside their institutions, and this does not mean simply conducting an analysis of peer schools.

Reinventing Higher Education: The Promise of Innovation offers thought-provoking commentary addressing some of the very large elephants in many conversations having to do with improving higher education. The ideas that are presented in the book's eight chapters are not necessarily new; however, they are innovative in that they challenge historical paradigms in a collective manner. As long as all stakeholders, regardless of title or department, keep talking and working toward student learning and measurement of student learning, these conversations will be headed in the right direction.

References

Common Core State Standards Initiative. (2012). *Adoption by state.* Retrieved July 11, 2012, from http://www.corestandards.org/in-the-state

National Center for Education Statistics. (2011). *Degree-granting institutions, by control and level of institution: Selected years, 1949-50.* Retrieved July 11, 2012, from http://nces.ed.gov/programs/digest/d11/tables/dt11_279.asp

RUMINATE: INTEGRATING THE ARTS AND ASSESSMENT

"DISCOVERY"

Photographer: Adam Barnes
Lynchburg, Virginia
www.adambarnes.com

We keep repeating that the world is making progress and that men must constantly be urged to pursue it. But true progress consists in the discovery of something hidden. Frequently it may be something that simply needs to be improved or perfected. No reward is offered for the discovery of something not foreseen; and, in fact, one who tries to bring it to light is frequently persecuted. It would be a disaster if poems were written solely with the hope of winning a state award. It would be better for a poet's vision to remain concealed within him and for the poetic Muse to disappear. A poem should flow from a poet's mind when he is not thinking of a reward or of himself; and even if he wins a prize, it should never make him proud.

Maria Montessori, *The Discovery Of The Child*,
Italian physician and educator (1870 – 1952)

GUIDELINES FOR SUBMISSION

All manuscripts submitted to *Research & Practice in Assessment* should be related to various higher education assessment themes, and adopt either an assessment measurement or an assessment policy/foundations framework:

Assessment Measurement:
a) instrument design, b) validity and reliability, c) advanced quantitative design, d) advanced qualitative design

Assessment Policy/Foundations:
a) accreditation, b) best practices, c) social and cultural context, d) historical and philosophical context, e) political and economic context

Article Submissions:

Articles for *Research & Practice in Assessment* should be research-based and include concrete examples of practice and results in higher education assessment. The readers of *Research & Practice in Assessment* are associated with myriad institutional types and have an understanding of basic student learning and assessment practices. Articles for publication will be selected based on their degree of relevance to the journal's mission, compliance with submission criteria, quality of research methods and procedures, and logic of research findings and conclusions. Approximately fifty percent of submissions are accepted for publication.

Review Submissions:

Reviews (book, media, or software) are significant scholarly contributions to the education literature that evaluate publications in the field. Persons submitting reviews have the responsibility to summarize authors' works in a just and accurate manner. A quality review includes both description and analysis. The description should include a summary of the main argument or purpose and overview of its content, methodology, and theoretical perspective. The analysis of the book should consider how it contrasts to other works in the field and include a discussion of its strengths, weaknesses and implications. Judgments of the work are permitted, but personal attacks or distortions are not acceptable as the purpose of the review is to foster scholarly dialogue amongst members of the assessment community. The *RPA* Editor reserves the right to edit reviews received for publication and to reject or return for revision those that do not adhere to the submission guidelines.

Special Features:

Each issue of *Research & Practice in Assessment* highlights the work of a noted researcher or assessment professional in a manner that aims to extend the scholarly dialogue amongst members of the assessment community. Special Features are invited by the Board of Editors and often address the latest work of the author.

Notes in Brief:

Notes in Brief offer practitioner related content such as commentaries, reports, or professional topics associated with higher education assessment. Submissions should be of interest to the readership of the journal and are permitted to possess an applied focus. The *RPA* Editor reserves the right to edit manuscripts received for publication and to reject or return for revision those that do not adhere to the submission guidelines.

Ruminate:

Ruminate concludes each issue of *Research & Practice in Assessment* and aims to present matters related to educational assessment through artistic medium such as photography, poetry, art, and historiography, among others. Items are encouraged to display interpretive and symbolic properties. Contributions to Ruminate may be submitted electronically as either a Word document or jpg file.

Manuscript format requirements available at: www.RPAjournal.com

RPA BOOKS AVAILABLE

Balzer, W. K. (2010). *Lean higher education: Increasing the value and performance of university processes.* New York, NY: Productivity Press. pp. 312. $51.95 (paper).

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2010). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs.* Sterling, VA: Stylus Publishing. pp. 224. $27.50 (paper).

Butt, G. (2010). *Making assessment matter.* New York, NY: The Continuum International Publishing Group Ltd. pp. 160. $27.95 (paper).

Cambridge, D. (2010). *Eportfolios for lifelong learning and assessment*. Hoboken, NJ: Wiley, John & Sons, Incorporated. pp. 288. $38.00 (hardcover).

Carey, K., & Schneider, M. (Eds.). (2010). *Accountability in American higher education.* New York, NY: Palgrave Macmillan. pp. 355. $95.00 (hardcover).

Christensen, C., & Eyring, H. (2011). *The innovative university: Changing the DNA of higher education from the inside out.* San Francisco, CA: Jossey Bass. pp. 512. $32.95 (hardcover).

Collins, K.M., & Roberts, D. (2012). *Learning is not a sprint*. Washington, DC: National Association of Student Personnel Administrators. pp. 216. $34.95 (hardcover).

Côté, J. E., & Allahar, A. L. (2011). *Lowering higher education: The rise of corporate universities and the fall of liberal education.* Toronto, ON: University of Toronto Press Publishing. pp. 256. $24.95 (paper).

Dunn, D. S., McCarthy, M. A., Baker, S. C., & Halonen, J. S. (2010). *Using quality benchmarks for assessing and developing undergraduate programs.* San Francisco, CA: Jossey Bass. pp. 384. $45.00 (hardcover).

Flateby, T. L. (Ed.). (2010). *Improving writing and thinking through assessment.* Charlotte, NC: Information Age Publishing. pp. 238. $45.99 (paper).

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered: Institutional integration and impact*. San Francisco, CA: Jossey Bass. pp. 224. $30.00 (paper).

Joughin, G. (Ed.). (2009). *Assessment, learning and judgment in higher education.* New York, NY: Springer Publishing Company. pp. 445. $289.00 (hardcover).

Kramer, G. L., & Swing, R. L. (Eds.). (2010). *Higher education assessments: Leadership matters.* Lanham, MD: Rowman & Littlefield Publishers, Inc. pp. 288. $49.95 (hardcover).

Makela, J.P., & Rooney, G.S. (2012). *Learning outcomes assessment step-by-step: Enhancing evidence-based practice in career services*. Broken Arrow, OK: National Career Development Association. $35.00 (paper).

Maki, P. L. (2010). *Assessing for learning: Building a sustainable commitment across the institution*. Sterling, VA: Stylus Publishing. pp. 356. $32.50 (paper).

Martin, R. (2011). *Under new management: Universities, administrative labor, and the professional turn.* Philadelphia, PA: Temple University Press. pp. 253. $69.50 (hardcover).

Noyce, P. E., & Hickey, D. T. (Eds.). (2011). *New frontiers in formative assessment*. Cambridge, MA: Harvard Education Press. pp. 260. $29.95 (paper).