Running Head: SEI IN HIGHER ED

Student Evaluations of Instruction in Higher Education

Michael Preuss

Liberty University

Abstract

The present American higher education setting includes the use of Student Evaluations of Instruction (SEIs). This practice has a history that spans a number of decades and is now firmly established as part of the American system of higher education. However, there is a continuing dispute over the use of the information gathered from SEIs, the legitimacy of the interpretations of the results, the assumptions behind the use of the instruments, and the assumptions foundational to the instruments themselves. This paper considers each of these topics and methods suggested by scholars to alter or augment SEIs.

Table of contents

Student Evaluations of Instruction in Higher Education

One of the areas in which Student Evaluations of Instructors (SEIs) are most often employed is gauging teacher effectiveness. With nearly four decades of use (Hooper & Page, 1986), the impact of SEIs on American higher education should be visible. "One should be able to accurately conclude that today we have more effective teachers and more learned college graduates than we did 30 or 40 years ago" (Olivares, 2003, p. 240) if SEIs have had their desired effect. If one was able to substantiate such an assertion, it "….would provide strong evidence of…validity" (Olivares, 2003, p. 240) for the use of SEIs as measures of teacher effectiveness and the interpretation of their data as a means of improving instruction. However, "there is no empirical evidence to suggest that the widespread implementation of teacher rating has resulted in more effective teachers or more learned students" (Olivares, 2003, p. 240). Since this is the case, one must wonder, what are the identifiable characteristics of SEIs and what result has their use produced?

General state of affairs

"Now, as at no other time in the history of education, there is a push for excellence, with indicators of institutional effectiveness often tied to personnel performance" (Szeto, 1994, p. 4). "Consumers of higher education are demanding with increasing regularity that this excellence originate in the classroom" (Szeto, 1994, p. 5). This is a circumstance that should not be a surprise according to Wergin. He states,

> "Under the old model, the assumption was that if the proper infrastructure was in place, students would learn well enough….But with both the demand for access to

> higher education and its costs at record levels…it should surprise no one that calls for institutional ‘accountability’ for student learning have become ever more strident” (Wergin, 2005).

The result of the push for excellence and accountability for learning, in part, is an emphasis on the use and interpretation of SEIs. Citing the research of others, Obenchain asserts that SEIs represent a significant area of emphasis in college education in America. She states, “American higher education has been experiencing a (paradigm) shift from academic merit to consumerism….In this transformation students have taken the role of gatekeeper…and consumer/customer” (Obenchain, 2001, p. 242). This role is fulfilled, in part, through their participation in the SEI processes at their respective institutions.

Proponents of SEIs are not concerned about this circumstance. In surveys and research, teaching effectiveness consistently is regarded as one of the most important aspects of an institution of higher education (Szeto, 1994, pp. 12-14). And, as Danielson has written, “’For many years, educators have agreed that the fundamental purposes of teacher evaluation are both quality assurance and professional development’” (White, 2002, p. 10). Therefore, SEIs are presented by their proponents as fulfilling these two necessary roles by obtaining feedback from the client regarding the instruction received.

SEIs currently account for a large percentage of the feedback received by institutions of higher education. In fact, as early as fifteen years ago they were one of the two most common sources of information about teacher performance.

> “Seldin (1989) found, over the ten-year survey period of 1978-1988, administrator evaluations were consistently the major source of information on teaching performance, although student evaluations…moved from third place to a virtual

> tie for first with chair evaluations in appraising teaching performance" (Szeto, 1994, p. 19-20).

Almost a decade before Seldin's research Stier found that the most common measure of the effectiveness of instruction "involv[ed] a combination student and administrative evaluation" (Szeto, 1994, p. 20). And by 1999, a decade after Seldin's ten year study, SEIs were such a part of the landscape of higher education that Layne, DeCristoforo, and McGinty saw them as "traditional" (p. 1) aspects of the endeavor. Having been a component of "quality assurance and professional development" (White, 2002, p. 10) in higher education for over three decades and comprising one of the two main sources of teacher effectiveness data, SEIs are firmly entrenched in academia.

However, during the same period, SEIs' detractors have been voicing opinions critical of their use. One key concern is the use of SEIs in the realm of personnel decision making. There continues to be a "debate revolving around what kind of measures should be used for summative evaluation of faculty, in making personnel decisions for retention, promotions, tenure, or salary increases, and of course, to assess their effectiveness" (Hobson & Talbot, 2001, p. 3). Some critics characterize SEIs as a "popularity contest" (Bell & McClam, 1992, p. 3), others question their validity.

> "So there in brief is my brief (not only mine; the points have been made before by many) against student evaluations: They are randomly collected. They are invitations to grind axes without any fear of challenge or discovery. They are based on assumptions that have more to do with pop psychology or self-help or customer satisfaction than with the soundness of one's pedagogy. A whole lot of machinery with a very small and dubious yield" (Fish, 2005, p. 5).

The detractors voicing these opinions have not lacked an audience. “The concern over the usefulness of faculty evaluations in improving teaching has stimulated intense debate, research, and action at various levels, and has led to the development of programs to address the need for improved data collection techniques” (Bangura, 1994, p. 1). But even the reported advancements in the composition of the SEIs themselves can not satisfy the detractors. Their objections have kept pace with the revisions. “True, the evaluation forms have been revised and supposedly refined, but in general the revisions have followed political and sociological trends rather than any advance in our understanding of what is and is not good teaching” (Fish, 2005, p. 2). “’In general, student evaluations can be taken [only] to report honestly student perceptions.... Perceptions are not necessarily accurate representations of the objective facts’” (Hobson & Talbot, 2001, p. 5).

As noted above, the purposes for which SEIs are intended is “’quality assurance and professional development’” (White, 2002, p. 10). The citations above indicate criticisms of the use of SEIs as measures of “quality assurance” (White, 2002, p. 10) and of their validity and applicability in the development of faculty and the necessary skills of a faculty member. Another critical concern to SEI detractors is defining teacher effectiveness which Fish refers to as “what is and is not good teaching” (2005, p. 2). Teacher effectiveness has been defined in multiple ways (Hooper & Page, 1986). And, it is, therefore, not only an uncertain target, as will be demonstrated below, it is a moving target. Thus detractors of SEIs who note the lack of consensus on what constitutes effective teaching ask, “How can one measure a construct that is yet to be defined?”

The use and interpretation of SEIs is an open debate. This debate includes discussions about the implementation of and validity and reliability of SEIs. The remainder of this paper is dedicated to describing primary points in that debate, and identifying a set of recommendations.

Primary purposes

> "Determining the purposes of the evaluation system before it is implemented and continuing to communicate these purposes to affected individuals is paramount to effective evaluation design….Evaluations that lack clearly articulated purpose(s) are essentially meaningless and contribute little to the accomplishment of the institution's goals" (Szeto, 1994, p. 8).

Szeto notes that both Gill (1977) and Locher and Teel (1984) found faculty believed evaluations should be used for

> "assessing training and development needs, improving current performance, assessing past performance, assisting career planning decisions, setting performance objectives, providing feedback on performance to employees, aiding human resources planning, and identifying employees for transfer and lay-off" (Szeto, 1994, p. 9).

In the opinion of faculty, the primary purpose of SEIs is use as a "'formative evaluation measure'…for the purposes of self-improvement" (Szeto, 1994, p. 9-10). "Student evaluations are formative when their purpose is to help faculty members improve and enhance their teaching skills" (Hobson & Talbot, 2001, p. 2). Hobson and Talbot believe that SEIs serve a

> "formative purpose only if the following four conditions are met: First, teachers must learn something new from them. Second, they must value the new

> information. Third, they must understand how to make improvements. And, finally, teachers must be motivated to make the improvements, either intrinsically or extrinsically" (2001, p. 2).

To accomplish these purposes would require a planned and regularly evaluated program of faculty development utilizing SEI results as a source of information, but not the only source. While SEIs results are employed in the manner preferred by faculty, formative assessment, they are also interpreted in other ways.

SEIs are also used for summative evaluation of personnel. When they are seen as a "rational, equitable basis for making personnel decisions….[they operate as] 'summative evaluation'" (Szeto, 1994, p. 10). "The summative purpose of SETEs [student evaluations of teaching effectiveness] is for use in evaluating the overall effectiveness of an instructor, particularly for tenure and promotion decisions" (Hobson & Talbot, 2001, p. 2).

These coincidental but distinct purposes for SEIs underline the importance in understanding the characteristics, implementation and interpretation of SEIs. A very important concern and one at the core of the discussions about SEIs is that voiced by Gould regarding "valid measures of teaching effectiveness. Before a department can formalize any system of assessment, it must first establish some consensus about what constitutes good teaching based on the proportional emphasis assigned to" (Gould, 1991, p.2) a set of skills that make up teaching effectiveness. As was noted above, there is yet to be a consensus on these skills or even if such a list should be limited to skills.

Data gathering arguments for SEIs

The arguments put forward for the use of SEIs are often pragmatic and address concerns related to gathering data. Since feedback is necessary and students are the only direct source of audience feedback regarding instruction received, SEIs or some other measure, are seen as necessary. In terms of practicality, it is argued that SEIs are to be preferred over many other forms of data gathering. SEIs can be performed at "…relatively low cost, [and provide] reduction in biasing error, greater anonymity, and considered answers and consultations" (Bangura, 1994, p. 2). Szeto includes a list of additional pragmatic arguments including SEIs serving as a catalyst for a consideration of teaching and learning and fostering participation of students with faculty in devising and implementing educational programs (Szeto, 1994, p. 8). Beyond pragmatic concerns and in terms of educational philosophy, proponents of SEIs believe that if they are

> "employed adequately, as Grasha…pointed out, evaluation procedures can lead to the improvement of teaching and related activities, an increase in faculty and student satisfaction with teaching, personal growth and development of the faculty member as a teacher, and opportunities for faculty advancement within the system" (Szeto, 1994, p. 7).

Each of these arguments could be granted if the assumptions of a valid, reliable, consistently implemented instrument, including accurate interpretation and application of the results consistent with the qualities of the data, could be affirmed. But, as will be demonstrated below, these points are all contested. Even the supposed benefits of the data gathered by SEIs is questioned by their detractors.

Objections to the use of SEIs related to data gathering

Even supporters of SEIs see problems with their use. In reporting the results of his study looking at the relationship of grades to SEI ratings of faculty, Eiszler, a proponent, stated “The conclusion that, although generally valid as measures of teacher effectiveness, college students’ ratings of instruction may be used in ways that raised questions of consequential validity, specifically by encouraging grade inflation, was supported” (Eiszler, p. 1). Szeto, a detractor, wrote “there is a dearth of evidence supporting the viability of principles and practices related to [SEI] evaluation methods” (Szeto, p. 4).

Detractors attack the practicality of SEIs. Layne, DeCristoforo and McGinty argue that they are “time-consuming and costly to administer” (1999, p. 222), produce questionable results because of “the lack of survey administration standardization procedures” (1999, p. 222), and are “often hurriedly completed” (1999, p. 223) in pressure packed and uncertain circumstances, following the final exam for example, minimizing the quality of the data collected.

Detractors point out further limitations of SEIs by citing the characteristics proponents view as positive pragmatic reasons for their use. Among the negative characteristics is the fact that SEIs, as commonly practiced, are “impersonal” (Bangura, 1994, p.2; see also Fish, 2005, p. 5) or anonymous, as a proponent would phrase the same point. Fish, in his “brief” cited above, indicates the concern with anonymity is a lack of accountability on the part of the student. The absence of a check on vengeance and tomfoolery renders SEI results questionable. Bangura also states that the “in-class [SEI]…emphasizes standardized procedures, experimental control, quantitative measures, and statistical analysis, the role of language has not been given full consideration" (Bangura, 1994, p. 4). While proponents of SEIs would see the characteristics listed as positives, Bangura believes they limit the ability of the student to express themselves and the ability of the survey to measure the "considerable individual variation in frames of reference, values, and levels of understanding" (Bangura, 1994, p. 3) among the respondents.

Applying these concepts to multi-ethnic, multi-national or mixed socioeconomic student groups, and the interpretation of the SEI results, detractors point out that SEIs "lack opportunities for probing, and [are] limited to influences that propel students toward passing or failing a course" (Bangura, 1994, p. 2). By this, it is meant that the

> "idea of discourse is suppressed in the in-class questionnaire….the overwhelming majority of the questions in an in-class survey are fixed-choice and yes-no types. The reason these types of questions are widely used…is to facilitate ease in coding" (Bangura, 1994, p. 3)

not accurate assimilation of or interpretation of data. "Yes-no and fixed-choice questions…provide a constraint in finding coherent answers" (Bangura, 1994, p. 3) as they limit the possible responses of the student in situation in which individual perceptions and opinions are being sought. This results in a "pervasive disregard of [the] respondent's social and personal context of meaning…in the questionnaire…and in the modes of interpretive theorizing about responses" (Bangura, p. 3).

Additionally, detractors of SEIs assert that that research has shown "teachers perceived as enthusiastic, good-humored, and warm consistently fare better on student evaluations…[thus] student-completed evaluations are more about the instructor than about the actual course" (Obenchain, 2001, p. 5). Conversely, they point out that "'In general, student evaluations can be taken [only] to report honestly student perceptions.... Perceptions are not necessarily accurate representations of the objective facts'" (Hobson & Talbot, 2001, p. 5). As a result, they believe SEIs reveal more about students than instruction. And, that

> "A missing element in a number of faculty evaluation surveys is coherence: the issue of using results gleaned from these surveys to help improve teaching is often poorly conceptualized and then data are collected that have little or nothing to do with resolving the problem" (Bangura, 1994, p. 11).

The completion of standardized surveys to identify and address concerns related to a unique classroom setting is believed to be suspect.

For every positive assertion made about SEIs, there appears to a counter assertion. As with all assessments, implementation, data gathering and other practical considerations are important. However, the primary concerns related to SEIs orbit around and are reflections of the reliability and validity of the instrument. It is on these important considerations, and one's beliefs about them, that the arguments regarding SEIs turn.

Reliability

The reliability or "consistency of…results" (Linn & Gronlund, 2000, p. 74) of SEIs is debated. Hooper and Page wrote in 1986 that "The conclusion is almost inescapable that students can evaluate faculty teaching with reasonable reliability" (p. 58). They also wrote, "The research on SETEs [Student Evaluations of Teacher Effectiveness] has provided strong support for their reliability, and there has been little dispute about it" (Hooper & Page, 1986, p. 4). But Hobson and Talbot responded in 2001, "The inter-rater reliability of SETE instruments appears to vary depending on the specific instrument being used and on the class size" (p. 3). If inter-rater reliability is variable, the conclusion that SEIs "evaluate faculty teaching with reasonable reliability" (Hooper & Page, 1986, p. 58) is not accurate. One must ask, how can there be such a strong difference of opinion on such an elemental point? Olivares answers this question. "Supporters [of SEIs] have focused on *justifying their conclusions* regarding the validity of SRT's [student reviews of instruction]; critics have focused on examining the *validity of inferences* stemming from student ratings research" (2003, p. 239). An example of this is the following quotation from Obenchain.

> "Mixed results from tests of the reliability and validity of student evaluations of faculty underscore…concern (Haskell 1997). Simmons (1996) reported two separate studies, both of which found poor reliability for student-completed evaluation instruments. Instrument problems included ambiguous items, positively or negatively skewed items, and items that had no correlation to classroom teaching performance. Earlier research on the validity and reliability of student-completed course evaluation instruments reported the instruments to be reliable, stable, and relatively valid (Marsh 1987; Marsh and Bailey 1993; Peterson and Kauchak 1982). However, a review of the research revealed reliability and stability over long periods of time and in multiple courses for the same instructor. None of the research examined the reliability of the individual students" (Obenchain, 2001, p. 1).

In a study designed to “explore the reliability of the student as an evaluator” (Obenchain, 2001, p.3) rather than “instructor's ratings over time…class-average student ratings…or…the instrument itself” (Obenchain, 2001, p. 3), reliability of SEIs was not supported. Using a methodology that allowed for multiple student ratings of the same group of courses, it was found that individual student’s ratings of the most effective course and least effective course matched “for less than one-third of the students sampled….if, as this study found, individuals are not consistent in their evaluations, then aggregated reliability measures are giving faculty a false sense of security” (Obenchian, 2001,p. 4). Obenchain states clearly that “individual students’ evaluations are not reliable from instrument to instrument, as found in this study” (2001, p. 5). This study illustrates that “What remains unclear is the reliability of individual students in evaluating faculty teaching effectiveness, whether the reason is the unreliability of the student, an invalid instrument, or an inconsistent system” (Obenchain, 2001, p. 6). The present state of affairs is that instruments appear to be reliable in consistently producing similar results for the same professor when evaluated by different audiences. However, it can be demonstrated that the persons completing the SEIs are not reliable. They do not provide the same evaluation of an instructor in an isolated setting from one instrument to the next.

Proponents of SEIs have been demonstrating that ratings for one instructor are stable over time and across populations, that class average ratings are stable, or that the same instrument yields similar results (Obenchain, 2001; Olivares, 2003) as means of establishing the reliability of SEIs. Detractors would argue that these are really examples of the reliability of the SEIs in measuring the perceptions of students about instructors’ practices and personality (Hobson & Talbot, 2001), in representing the law of averages

(Olivares, 2003), and their reliability as a measure of some characteristic that is yet to be determined as the SEIs have low validity (Hobson & Talbot, 2001). As noted above, this is not a dispute regarding a defined construct, rather it is a dispute involving beliefs about SEIs, which researchers have attempted to justify.

As “reliability is a necessary…condition for validity” (Linn & Gronlund, 2000, p. 75), the absence of evidence for the consistency of individual students’ evaluations of instructors and inter-rater reliability causes one pause. This is especially the case when considering the present and historical uses of SEIs which the presence of reliability in “instructor's ratings over time…class-average student ratings…or…the instrument itself” (Obenchain, 2001, p. 3) does not dispel. It is evidence of consistent measures of instructors’ effectiveness provided by the same student when rating the same courses on different related instruments that is sought and which is missing.

Validity

The validity of SEIs “refers to the appropriateness of the interpretation of the results” (Linn & Gronlund, 2000, p. 75). “Specifically, this issue addresses a level of confidence that student evaluations are reflections of an instructor's effectiveness rather than artifacts of a particular course” (Hobson & Page, p. 4) or some other construct. Much of the research covering SEIs deals with these issues.

Some of that research comparing SEIs to other measures has been inconclusive for the purposes of establishing the validity of SEIs as measures of teacher effectiveness.

> “Feldman (1988) and Marsh (1984) both concluded that the correlation between peer ratings and student ratings is unacceptably low. Marsh stated the following:

> Peer ratings based on classroom visitation do not appear to be substantially correlated with student ratings or with any other indicator of effective teaching. Although these findings neither support nor refute the validity of student ratings, they clearly indicate that the use of peer evaluations of university teaching for personnel decisions is unwarranted" (Hobson & Talbot, 2001, p. 5).

Some of the research done has been interpreted as exhibiting the validity of SEIs as measures of teacher effectiveness. Lamberth and Kosteski performed research in 1981 comparing the ratings of classroom instruction gathered from teaching assistants and students. Their results showed a significant correlation between the two measures leading these researchers to conclude that SEIs are valid measures of teacher effectiveness. However, these researchers do not define teacher effectiveness nor do they demonstrate that positive ratings from students or student assistants on the instrument employed is a criterion related to teaching effectiveness.

Further evidence of validity is thought to be provided by comparisons of SEIs and tests of student knowledge.

> "Researchers often compare SET [Student Evaluations of Teachers] ratings to objective test measures of student knowledge and/or learning to see whether the measures are positively correlated (Cohen 1981; Feldman 1989a). These studies generally show that student evaluations and learning, as measured by objective tests, are positively correlated but generally not higher than simple r measures of 0.7" (Bosshardt, 2001. p.1).

"Because objective tests do not fully measure learning, or how much of learning is due to the instructor's influence, SET ratings have also been compared to other types of teaching evaluations” (Bosshardt, 2001, p. 1). These include “instructors' self-evaluations and administrators' evaluations” (Bosshardt, 2001, p. 1). There is, in fact, a large body of research that has been performed considering this question.

> “Feldman (1989b) analyzed over 40 earlier studies comparing teaching evaluations done by instructors, instructors' former students and current students, instructors' colleagues, external observers, and administrators. He found the strongest correlation between current and former student evaluations. Instructors' self-evaluations were not highly correlated with those from other sources, but the highest correlation was found to be between self-evaluations and ratings by current students” (Bosshardt, 2001, p. 1).

As the reader may note, the case for the validity of SEIs as a measure of teacher effectiveness is not firmly established by the research cited above. There are many common characteristics between teaching assistants and students that could affect the ratings of instructors provided by both groups. The connection of positive ratings on the instrument to effective teaching has not been established. The correlation of knowledge acquired with rating of teaching effectiveness is positive but not strong. That the “strongest correlation [existed] between current and former student evaluations” (Bosshardt, 2001, p. 1) when considering many possible means of validating SEIs as measures of teacher effectiveness only indicates that there is long-term consistency in ratings of instructors. It provides no explanation for this consistency. And, that the “highest correlation was found to be between self-evaluations and ratings by current students” (Bosshardt, 2001, p. 1) is no surprise. Both parties are intimately acquainted with what has happened in the classroom. However, this is again not proof positive of validity for SEIs in the realm of teacher effectiveness.

Among the studies regarding SEIs, there are several general themes. Shevlin, Banyard, Davies and Griffiths note that in 1997 d'Apollonia & Abrami published a paper stating

> "A number of extraneous variables have been examined that may confound the measurement of teaching effectiveness. The relationships between ratings of teaching effectiveness and variables related to student characteristics, lecturer behaviour, and the course administration have been examined" (2000, p.2).

In terms of student characteristics, "Marsh & Roche (1997) reported similar relationships between ratings [of instructors] and the prior subject interest of the student and the reason for taking the course" (Shevlin, Banyard, Davies & Griffiths, 2000, p.3). Cruse performed research in 1987 "that suggests that grade expectancies, along with other factors, bias student ratings" (Hobson & Talbot, 2001, p. 4). Other examples could be provided. It is sufficient to note that the characteristics of students can be seen to account for the ratings given instructors on SEIs. Even unusual comparisons, like that done by Bangura considering focus-group results and survey results which found dissatisfied student groups still rated professors as "good" on SEIs, indicate that some characteristic of the student is influencing SEIs (Bangura, 1994).

It is not only the characteristics of student which influence SEIs, the characteristics of the instructors do also.

> "The variable related to the lecturer behaviour that has received the greatest research interest is that of grading leniency. Using a large sample of American

> students Greenwald and Gillmore (1997) demonstrated that grading leniency had a strong positive relationship with ratings of teaching effectiveness" (Shevlin, Banyard, Davies & Griffiths, 2000, p.3).

In a study looking at the effect of instructor charisma on SEI ratings, Shevlin, Banyard, Davies and Griffiths stated "the charisma factor explained 69% and 37% of the variation in the 'lecturer ability' and 'module attributes' factors respectively. These findings suggest that student ratings do not wholly reflect actual teaching effectiveness" (2000, p. 1).

> "With regard to the effect of course administration, there is, for example, a weak relationship between class size and student ratings with the largest and the smallest classes giving the most positive ratings (Fernàndez et al., 1998)" (Shevlin, Banyard, Davies & Griffiths, 2000, p.3).

As Shevlin, Banyard, Davies and Griffiths concluded, "overall, research on the effects of extraneous variables on the validity of SET suggests the need for caution in the interpretation of…data" (Shevlin, Banyard, Davies & Griffiths, 2000, p.3) from SEIs. These examples call into question the foundational constructs in the use of SEIs.

The concerns with validity in the use of SEIs to measure teacher effectiveness extend far beyond identifying potential influences on the results. They include questions regarding the very essence of the SEIs, their content, criterion and construct validity. As Hobson and Talbot wrote, citing multiple researchers, "Validity [for SEIs]…is especially difficult to establish because researchers concede that there is no universally accepted criteria for what constitutes effective teaching" (2001, p. 4).

> "Central to any type of validity evidence…is the need to clearly define the variable of interest….regardless of the validation method(s) used, if teacher effectiveness has not been adequately defined, then it logically follows that any inferences drawn regarding the validity of data or processes to assess teacher effectiveness are seriously compromised" (Olivares, 2003, p. 236).

Research articles published to demonstrate factors influencing SEIs are of limited value, as are the SEIs themselves, if the content, criterion and construct validity of SEIs is reduced. The studies mentioned above reduce the prospect of SEIs having validity and these are not isolated instances.

Content validity

"The goal in the consideration of content validation is to determine the extent to which a set of assessment tasks provides a relevant and representative sample of the domain of tasks about which interpretations of assessment results are made" (Linn & Gronlund, 2000, p. 78). Fish refers to the content validity of SEIs. He notes that some

> "questions might be relevant in some teaching situations, but not in others. 'Did the instructor give lectures that facilitated note taking?' (Even in lecture courses this might not be a suitable measure; note taking is not what some lecturers, for reasons they could articulate, wish to provoke)" (Fish, 2005, p. 4).

Considerations of this type involve the particulars of individual SEIs. Each prompt must be examined for "a relevant and representative sample of the domain of tasks about which interpretations of assessment results are made" (Linn & Gronlund, 2000, p. 78).

Such a consideration is not the focus of this paper. Although, examples cited above would indicate that some instruments in use do have limited content validity.

Criterion and Construct Validity

Another broad and foundational concern beside content validity is a clear understanding of the “domain of tasks” (Linn & Gronlund, 2000, p. 78) about which one is seeking to gather data. In the case of SEIs, this is a “function of, first and foremost, the adequacy of the definition of teacher effectiveness” (Olivares, 2003, p. 234). Citing Scriven’s writing from 1981, Olivares asserts that this definition “should reflect a set of teacher behaviours that are universally acceptable ‘across the whole range of subjects, levels, students, and circumstances’ and reflect an equally acceptable definition of teacher effectiveness” (2003, p. 234). This is criterion and construct validity as it applies to SEIs.

Criterion validity is important when using assessments to “predict future performance or to estimate current performance on some valued measure other than the test itself” (Linn & Gronlund, 2000, p. 88). SEIs are advocated as a means of identifying teacher effectiveness. In this process, the results of the SEI administration is thought to provide information regarding “future performance or to estimate current performance” (Linn & Gronlund, 2000, p. 88) in instruction, a criterion concern.

“A construct is an individual characteristic that we assume exists in order to explain some behavior” (Linn & Gronlund, 2000, p. 83).

> “When we interpret assessment results as a measure of a particular construct, we are implying that there is such a construct, that it differs for other constructs, and

> that the results provide a measure of the construct that is little influenced by extraneous factors" (Linn & Gronlund, 2000, p. 83).

Both of these matters are worthy of attention when discussing the validity of SEIs.

Criterion and Construct Validity: Criterion Suitability

SEIs are employed because it is believed that they can predict future performance and identify current performance levels for instructors. This is criterion validity. It is believed that certain criteria exist and their presence is indicative of quality instruction and teacher effectiveness. Studies have been performed seeking to establish or disprove this belief. The study done by Orpen (1981) serves as an example. Orpen proceeded on the assumption that student test performance equaled student learning which was then taken as an indicator of teacher effectiveness (1981). Rodin and Rodin state this argument as the "objective criterion of teacher effectiveness is based on what students have learned from the teacher" (Rodin & Rodin, 1973, p. 5). However, one must ask, among other questions, how was learning measured? Does performance on subjective evaluations equal learning and does it represent only learning associated with the instruction? Are there not extraneous factors like student motivation? Does no learning from text or learning outside of class related to the subject matter occur? Researchers have stated that "student achievement, [as is employed by Orpen,] may be a suspect criterion because it is as clearly a function of the students as it is of the instructor" (Drews, Burroughs & Nokovich, 1987, p. 23). Olivares also states that research has shown that "student learning [used as] a predictor of student ratings and a proxy for teacher effectiveness….is questionable" (Olivares, 2003, p. 235). He cites "Marsh and Roche (1977) [who] suggested that ‘no single criterion of effective teaching is sufficient’ and they ‘categorically rejected the appropriateness of this narrow criterion-related approach to validity’” (Olivares, 2003, p. 235) when discussing SEIs. He also notes that “Scriven (1981) stated, ‘The best teaching is not that which produces the most learning, since what is learned may be worthless’” (Olivares, 2003, p. 235). In fact, Rodin and Rodin found that “students may not wish to maximize learning [, as is being assumed,] as much as balancing learning and effort needed to obtain it” (Rodin & Rodin, p. 8). Williams and Ceci noted research which shows that judgments about instructors can be made in seconds and from image only videos of lectures (no audio) that provide similar results to SEIs administered at the end of the semester (1997). They also performed a study that established a link between enthusiastic presentation and student ratings. Their study found “that factors unrelated to actual teaching effectiveness (such as variation in a professor’s voice) can exert a sizable influence on student ratings of that same professor’s knowledge, organization, and basic fairness” (Williams & Ceci, 1997 p. 23). While one might dispute whether tone and enthusiasm are or are not characteristics of teacher effectiveness, the results of this study are striking. That variance in tone and enthusiasm correlates with ratings of the instructor’s knowledge, organization, and fairness is remarkable. There are apparently a great many possible criteria for effective instruction, seemingly including voice tone qualities and enthusiastic presentation. Williams and Ceci believe that their “findings do not imply that we should abandon the use of student ratings – only our uncritical acceptance of their validity” (1997, p. 23). Based on the evidence above, this should certainly be the case as regards criterion validity.

Criterion and Construct Validity: Definition of the Construct

"One of the most important questions about SET [Student Evaluations of Teachers] instruments is their construct validity--the degree to which they measure what they are designed to measure" (Bosshardt, 2001, p. 1). "Paramount to the establishment of construct validity is the definition of the construct" (Olivares, 2003, p. 235). "Regardless of the validation method(s) used, if teacher effectiveness has not been adequately defined, then it logically follows that any inferences drawn regarding the validity of data or processes to assess teacher effectiveness are seriously compromised" (Olivares, 2003, p. 236).

What is understood to be teacher effectiveness? "Scriven (1981) suggested…[that] a set of teacher behaviors that are universally accepted across a wide range of students, contexts and pedagogical methods and reflect an equally acceptable definition of teacher effectiveness" should be employed on SEIs (Olivares, 2003, p. 238) to measure teacher effectiveness. However, as noted above, Medley traced the definition of teacher effectiveness through four stages (Hooper & Page, 1986). These stages included focuses on different sets of characteristics to define effective instruction. Even if one is limited to the present emphasis on discovering the characteristics of an effective teacher, there is a wide variety of views.

> "Despite the perceived importance of SET [student evaluations of teachers] there are theoretical and psychometric issues related to the assessment of teaching effectiveness that are yet unresolved. First, there appears to be little agreement on the nature and number of dimensions that represent teaching effectiveness" (Shevlin, Banyard, Davies & Griffiths, 2000, p.2).

> “For example, Swartz et al. (1990) identify the two factors of effective teaching as (1) clear instructional presentation, and (2) management of student behaviour, whereas Lowman and Mathie (1993) identify them as (1) intellectual excitement, and (2) interpersonal rapport. There is no obvious mapping between these two pairs of dimensions. Further studies identify more and different factors of teaching effectiveness. For example Brown and Atkins (1993) identify the three factors of effective teachers as (1) caring, (2) systematic, and (3) stimulating, whereas Patrick and Smart (1998) identify the three factors of teaching effectiveness as (1) respect for students, (2) organisation and presentation skills, and (3) ability to challenge students” (Shevlin, Banyard, Davies & Griffiths, 2000, p.2).

The state of South Carolina employs the “Assessments of Performance in Teaching (APT)” (_______, 1981, p. 1). The goal of this assessment is to measure “minimal competence” (_______, 1981, p. 1) in teaching. “The instrument addresses five performance dimensions: planning; instruction; classroom management; communication; attitude” (_______, 1981, p. 1). Orpen identified “seven teaching dimensions: skill, achievement, interaction, overload, structure, rapport, and feedback” (Orpen, 1981, p. 6). “Other researchers have suggested…seven factors (Ramsden, 1991) or nine factors of effective teaching (Marsh & Dunkin, 1992)” (Shevlin, Banyard, Davies & Griffiths, 2000, p.2). The nine suggested by Marsh are “learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, examination/grading, assignments, and workload/difficulty” (Bosshardt, 2001, p. 2). Centra also lists nine

> “qualities…most often cited as related to effective teaching. These attributes are…communication skills – good speaking ability…favorable attitudes toward students…knowledge of subject…good organization of course and subject…fairness in examinations and grading…flexibility…encouragement of student thought…interests in and enthusiasm for subject…[and] thorough preparation for class” (Hooper & Page, 1986, p. 57-58)

“Feldman (1988) identified twenty-two ‘instructional dimensions’ of effective teaching in his research on SETEs [student evaluations of teacher effectiveness” (Hobson & Talbot, p. 2).

On the opposite extreme, there are also researchers who acknowledge that teaching is multi-faceted but doubt the ability of SEIs to measure the presence of these facets.

> “Abrami and d'Apollonia (1997), while acknowledging the multidimensionality of teaching, were skeptical of the use of ratings on multiple items dealing with particular dimensions of teaching to make summative decisions on overall teaching effectiveness. They argued that the specific attributes of good teaching vary across courses and instructors and recommended using global evaluation items whenever summative judgments about teaching effectiveness are called for” (Bosshardt, 2001, p.2).

Similarly, there are scholars who believe that the multidimensional nature of teaching should be expanded from a primarily cognitive emphasis to include “social, civic and personal outcomes” (Shavelson & Huang, 2003, p. 12). This is the case since on present SEIs “only a small portion of the cognitive and motivational goals of a course may be tapped” (Drews, Burroughs & Nokovich, 1987. p. 23).

There are also scholars who question whether the use of SEIs to measure teacher effectiveness is not a circular and self-perpetuating system. “One of the issues to consider is whether we are measuring the most important variables of teaching effectiveness or whether some variables are becoming more important just because they are measurable” (Shevlin, 2000, p. 1).

There is agreement that teaching effectively is a multidimensional construct. However, there is not agreement as to the various components of that construct or the number of characteristics of the enterprise which should be considered when seeking to measure teaching effectiveness. All of this is related to one point. “Supporters and critics of SRT’s [student ratings of teachers] concur that ‘teacher effectiveness’ has not been adequately defined and operationalized” (Olivares, 2003, p. 237) by researchers and educators.

Yet students are asked to employ a definition of teaching effectiveness when completing a SEI. Therefore, there is also the need to consider the definition of the effective teaching being employed by the students. Research demonstrates that the multidimensional listings of teacher effectiveness components are not one and the same as the definitions employed by students who do the rating. And, there is evidence that there is not a uniform view of effective teaching across student populations. In fact, “there is considerable evidence that suggests that students do not hold a common view of teacher effectiveness (Chandler, 1978; McKeachie , 1979) and students are prone to judgmental biases (e.g. Scullen *et al.*, 2000; Stanfel 1995)” (Olivares, 2003, p. 237).

> “Students’ holistic rankings represented their own perceptions of quality teaching with no parameters set by a standardized evaluation instrument. Kishor (1995) contends that students base their evaluations on an implicit personality theory of a good instructor….recall previous information and infer other information to from their personality theory” (Obenchain, 2001, p. 4).

> “Murphy and Cleveland (1995) propose….raters make evaluations of [teacher] performance with their own self-interests in mind and within an organizational context. Similarly, McKeachie (1997) suggests that students evaluate their academic experiences based on the extent to which their experiences satisfy their academic goals” (Olivares, 2003, p. 237).

As a result of the above, SEIs can not be viewed as instruments based on logic and established constructs which then produce consistent measures of teacher effectiveness. Olivares writes that

> “given the quiddities of human nature, the nature of SRT’s [student ratings of teachers], and the method by which ratings are generated and instructor effectiveness evaluated, it is highly improbable that student ratings are good measures of teacher effectiveness. Furthermore, evidence by both supporters and critics of SRT’s, as well as the principles of validation and logic therein, suggest that SRT’s are of questionable validity and, therefore, are not appropriate for drawing inferences regarding ‘teacher effectiveness.’ Hence, the continued resistance to SRT’s as valid measures of teacher effectiveness appears to be well founded” (2003, p . 240).

Criterion and Construct Validity: Irrelevant Variance

"Two questions are central to any construct validation: (1) Does the assessment adequately represent the intended construct? (2) Is performance influenced by factors that are ancillary or irrelevant to the construct?" (Linn & Gronlund, 2000, p. 83). The first of these questions is referred to as construct representation or under-representation (Linn & Gronlund, 2000, p. 83) and has been discussed above. Since the construct of teacher effectiveness "has not been adequately defined and operationalized" (Olivares, 2003, p. 237), there can be no certainty that it is represented on SEIs and, therefore, the validity of SEIs as measures of teacher effectiveness is called into question. A second characteristic of validity is "construct-irrelevant variance" (Linn & Gronlund, 2000, p. 83), which considers the question, "is performance influenced by factors that are ancillary or irrelevant to the construct?" (Linn & Gronlund, 2000, p. 83).

When SEIS are used as measures of teacher effectiveness, "it is assumed…that:

- rating forms adequately capture the domain of teacher effectiveness across instructional settings, academic disciplines, instructors and course levels and types;
- students know what effective teaching is, hold a common view of teacher effectiveness, and are objective and reliable sources of teacher effectiveness data;
- relatedly, ratings are, for all intents and purposes, unaffected by potential biasing variables; and, collaterally;
- teacher effectiveness is being measured as opposed to, for example, course difficulty or differences in disciplines, student characteristics, grading

> leniency, teacher expressiveness, teacher popularity or any number of other variables.

Thus, the validity of SRT's requires that each criterion above be met" (Olivares, 2003, p. 236). The material presented above has demonstrated that the first two sub-points can not be said to be the case for SEIs. The following examples will illustrate that the last two sub-points, which represent concerns with irrelevant variance, cannot be said to apply either. "Anthony G. Greenwald and Gerald Gilmore of the University of Washington finding that easy grading professors receive better evaluations" (Wilson, 1998, p. 1). In another study "at Salford University….[it was found] that the scores [on SEIs] are also related to less desirable features such as lecture size and subject matter" (Fox, survey 25). Williams and Ceci found that enthusiastic presentation improves ratings of teacher effectiveness (Williams & Ceci, 1997). They also discussed studies which have been done comparing student judgments of teacher effectiveness completed after 30 seconds, or less, without audio, to SEIs completed after an entire semester of classroom interaction. In these studies the results of the two measures were found to be similar (Williams & Ceci, 1997). Everett demonstrated that the difficulty of the subject, preparation of the students, and an emphasis on higher level thinking skills in instruction effects SEI's (1977). But perhaps most striking are the findings of Felton, Mitchell and Stinson. They compared ratings of the sexiness of professors on an internet website with SEI results.

> "When grouped into sexy and non-sexy professors, the data reveal that students give sexy-rated professors higher quality and easiness scores. If these findings reflect the thinking of American college students when they complete in-class

> student opinion surveys, then universities need to rethink the validity of student opinion surveys as a measure of teaching effectiveness. High student opinion survey scores might well be viewed with suspicion rather than reverence, since they might indicate a lack of rigor, little student learning, and grade inflation" (Felton, Mitchell & Stinson, 2004, p. 1).

Criterion and Construct Validity: Summary

The validity of SEIs as measures of teacher effectiveness is not supported by current research in terms of criterion validity. "Teacher performance is a dynamic criterion predicated on an ill-defined notion of teacher effectiveness" (Olivares, 2003, p. 237). The construct validity of SEIs in regards to irrelevance variance is also low. There are multiple factors that influence SEIs.

> "Student ratings have been shown to be influenced by precourse interest (Marsh, 1982), student emotional states (Small *et al.*, 1982), student expectations (McKeachie, 1979, 1997), grading leniency (Greenwald & Gillmore, 1997b), course difficulty (Mason*et al.*, 1995), academic discipline (Cashin, 1990), non-verbal teacher behaviours and teacher likability (Ambady & Rosenthal, 1993), congruence of student-teacher cultural backgrounds (Amin, 2000), teacher expressiveness (Williams & Ceci, 1997), and teacher charisma (Shevlin *et al.*, 2000). Evidence suggests that it is fair to conclude that students do not objectively rate teachers on characteristics found on omnibus rating forms….there are confounding factors that influence or bias ratings" (Olivares, 2003, p. 238).

The use of SEIs to measure teacher effectiveness is also not supportable as regards the definition of the construct. This is true whether one is considering the definition of teacher effectiveness used in constructing the SEI or that used by the rater. “Supporters and critics of SRT’s concur that ‘teacher effectiveness’ has not been adequately defined and operationalised” (Olivares, 2003, p. 237) by educators and scholars.

> “Research suggests that students do not hold a common view of teacher effectiveness nor do they make judgments of teacher effectiveness in a vacuum; rather, there are myriad factors that may influence students’ perceptions of ‘teacher effectiveness’ and the ‘objectivity’ of students’ evaluative judgments. Husbands and Fosh (1993) suggest that that subjectivity in student ratings of teachers is illimitable. To think that students, who have no training in evaluation, are not content experts, and possess myriad idiosyncratic tendencies, would not be susceptible to errors in judgment is specious” (Olivares, 2003, p. 237).

The debate over the validity of SEIs will continue. This is due, primarily, to the proponents and detractors generating research on two different but related tracks. “Supporters have focused on *justifying their conclusions* regarding the validity of SRT’s [Student Reviews of Teachers]; critics have focused on examining the *validity of inferences* stemming from student ratings research” (Olivares, 2003, p. 239).

SEIs will also continue to impact American higher education. Notable impacts discovered through the use of surveys were “reduction in coursework demands on students…lowering grading standards...[that] 50% of respondents [indicated that they] had attempted to improve their ratings in ways they considered inappropriate” and grade inflation (Olivares, 2003, p. 241). “Data suggests that the institutionalization of SRT’s [Student Reviews of Teachers] as a method to evaluate *teacher effectiveness* has resulted in students learning less in environments that have become less learning- and more consumer-oriented” (Olivares, p. 243).

However, “a lack of validity does not mean the SRT’s are not useful; rather, it just suggest that SRT’s are not measuring what they are intended to measure and therefore inferences regarding teacher effectiveness or student learning should be constrained” (Olivares, 2003, p. 240).

Other Methods Suggested

In response to the research conducted on SEIs, scholars have suggested the use of additional forms of evaluation to measure teacher effectiveness. This is necessary since “An important goal for educational institutions…[is] the construction of comprehensive, objective, individualized, systematic, public, and fair personnel evaluation systems that are consistent with the law and cost effectiveness for that particular institution” (Szeto, p. 22).

Bangura’s research considered focus-groups. He found that “focus groups provided additional information related to improving teaching” (Bangura, 1994, p. 10). As noted above, this additional information included the insight that professors who students openly criticized in a focus-group discussion were still rated as “good” by the same individuals on SEIs (Bangura, 1994). Cosser wrote about the use of teaching portfolios in a university setting (1997). Layne, DeCristoforo and McGinty have investigated and advocated the use of anonymous digital feedback for professors during the course of the semester (1999). Fish has suggested a standing faculty committee to review complaints, concerns and, presumably, accolades (2005). “Miller (1974) suggests five types of evaluation of classroom teaching: Student evaluations; classroom visitations, review of teaching materials and procedures, special incident reports, and self-evaluation” (Szeto, 1994, p. 13). Szeto has suggested the evaluation of research and scholarship produced by faculty while simultaneously voicing a number of concerns and caveats in this area (Szeto, 1994), and the evaluation of institutional service on the part of the faculty member (Szeto, 1994). These final categories would significantly expand the categories of teaching effectiveness and include some of the aspects mentioned by Shavelson and Huang (2003).

When considering these suggestions, one must refer to the research done by Feldman and the discussion of criterion and construct validity above.

> “Feldman (1989b) analyzed over 40 earlier studies comparing teaching evaluations done by instructors, instructors' former students and current students, instructors' colleagues, external observers, and administrators. He found the strongest correlation between current and former student evaluations. Instructors' self-evaluations were not highly correlated with those from other sources, but the highest correlation was found to be between self-evaluations and ratings by current students” (Bosshardt, 2001, p. 1).

Considering these comparisons and the presence of many factors related to the learner, the environment, the instructor, the curriculum, etc., mentioned above which influence SEIs causes one to view additional methods of determining teaching effectiveness with reservations. In addition, such a consideration reinforces the evidence that we may be "measuring the most important variables of teaching effectiveness or…some variables [which] are becoming more important just because they are measurable" (Shevlin, 2000, p. 1). And, there are practical concerns with each of the above methods. For example, both Wilson (1998) and Deegan (1974) refer to a significant concern in a peer review or administrator review system. "Results of the survey showed that there was a decisive consensus among all groups that lack of administrative time and lack of faculty time were the major problems encountered in implementing a faculty evaluation program" (Deegan, 1974, p. 11) which involved reviews by peers or administrators.

However, of the methods discussed in the literature reviewed for this work, there are a number which appear to have merit. In addition, there is a consensus that multiple sources of information is the most desirable circumstance when evaluating teacher effectiveness.

> "According to Whitman and Weiss (1982), 'If there exists one conventional wisdom in the field of faculty evaluation, it is that using multiple data sources is desirable.' Likewise, Kronk and Shipka (1980) contend that a combination of methods provides a check-and-balance system" (Szeto, p. 20).

Gould agrees. He states, "no single source is appropriate for assessing a teacher's effectiveness in all [areas]. Student evaluation, peer evaluation, and self-evaluation all have strengths and weaknesses when used to evaluate teaching effectiveness" (Gould, 1991, p. 1).

In light of these facts, this reviewer believes a combination of the following methods would be most productive for evaluating teacher effectiveness. Bangura's approach using focus-groups seems to show a great deal of promise. The ability of the work he did to probe for information, consider the idiosyncrasies of the individual and the situation, and reveal coincidental but seemingly contradictory results speaks well of the potential for this approach. Bangura points out that focus-groups allow “specification of certain relationships between ideas….[and] delineat[ion of] logical connections between causes and effects” (1994, p. 11). Further, it allows one to “make deductions from ideas we otherwise would not have thought about” (1994, p. 12) and to “pinpoint what our investigation was and what we needed to ferret out” (1994, p. 12). These are all highly desirable characteristics in assessments.

The limitations of focus-groups are, however, significant. They could only be used with a limited number of students, providing a potentially representative sampling rather than a broad and nearly global sampling. In addition, establishing the rapport and probing for insight in focus-groups requires considerable time investment in comparison to administering written surveys. Finally, to be done well, focus-group data gathering requires the involvement of a trained facilitator who is also an exceptional interpersonal communicator. Adding this approach to assessment of teacher effectiveness would require the addition of a part time focus group facilitator to a college’s staff if not several full time employees specializing in this form of instructional follow up.

Cosser’s article (1996) discusses the application of a portfolio process in teacher effectiveness evaluation at the university level. This approach has many positive attributes. It would allow the faculty person to demonstrate growth and advancement, to be evaluated for strengths and weaknesses, to demonstrate measurable or certifiable competencies rather than receive anonymous critique, and to be provided personalized development plan. It would provide the evaluators access to many aspects of instruction and preparation for instruction that a survey does not and can not include and it can be an umbrella under which other forms of evaluation are subsumed. For example, "Miller (1974) suggests five types of evaluation of classroom teaching: Student evaluations; classroom visitations, review of teaching materials and procedures, special incident reports, and self-evaluation" (Szeto, 1994, p. 13). Each of these forms could be included in a portfolio. This approach also has limited administrative involvement eliminating the need for additional personnel to facilitate the process. However, the evaluation of these portfolios would require considerable expertise and need to be vetted for philosophical biases.

Benjamin, DeCristoforo, and McGinty (1999) describe the use of digital evaluations. In their study, digital feedback provided at the student's discretion within a prescribed period following a course is compared to SEIs administered in classroom settings. This introduces an interesting change to SEIs and eliminates one of the objections voiced by critics. There is no pressure packed, time limited administration of the SEI. And, the flexibility of this approach, facilitated by the use of the Internet, opens up many avenues to consider in using digital feedback for instructors. An example is the system employed at Concordia University (Judith Preuss, personal communication, February 20, 2005). Students are allowed the opportunity to provide SEI data digitally in the first third of the semester. This data is processed and provided to the instructor. The instructor is to use this information to modify the course materials, supplements, and processes in order to aid in advancing student learning in that specific course and context. Final SEIs are also employed but these are structured to provide feedback about the response of the instructor to the earlier feedback provided by students. In this manner, information is gathered early in the instructional process, processed efficiently, and used to alter the instruction thereby benefiting the student. Other forms of rapid, itinerant, or formative SEIs could be structured using technology available on every college campus and widely available to students off campus. As in many other areas of modern life, the Internet holds promise for improvement of the processes and pace of evaluation of instruction.

Another important component of evaluation of instruction is that discussed by Ciscell (1987) and Bangura (1994). Both of these authors focused on developing evaluation instruments tailored to the specific teaching situation. Ciscell (1987) also discussed the timing of the evaluations while Bangura (1994) developed a theme of responsiveness to the social, cultural, economic, and philosophical perspective of the student. This emphasis on the specificity of the setting was also addressed by Fish (2005). There is much to be said for SEIs which are tailored to the setting. They could eliminate some of the irrelevant factors influencing SEI results by focusing on a narrower, more selective, and situationally applicable set of characteristics. They would also allow for the involvement of students in the process of developing the SEI instrument or its content. This would provide the opportunity to define the construct or constructs behind the SEI for both the evaluators and the evaluated. This would begin to address the construct validity concern with SEIs although, it would stop far short of resolving the dispute over the nature and extent of the characteristics of effective teaching. This entire process could be documented in a portfolio, even repeatedly documented demonstrating different skills or advancing comprehension or analysis of a concept or concern. In addition, it is highly compatible with the use of the Internet in itinerant, formative SEI data gathering.

It should be noted that while each of these approaches appears to have merit, research has demonstrated difficulties with many other approaches to evaluating teaching effectiveness. Feldman’s analysis of studies is once again applicable.

> “Feldman (1989b) analyzed over 40 earlier studies comparing teaching evaluations done by instructors, instructors' former students and current students, instructors' colleagues, external observers, and administrators. He found the strongest correlation between current and former student evaluations. Instructors' self-evaluations were not highly correlated with those from other sources, but the highest correlation was found to be between self-evaluations and ratings by current students” (Bosshardt, 2001, p. 1).

Every new method of gathering information about instruction must be validated through research. In addition, what has already been demonstrated in the literature must be considered when devising alternative approaches to evaluation of instruction. While the use of multiple forms of evaluation is a safe-guard, it is not a solution to the problems associated with SEIs. Supplementing general SEIs with focus-group results, regular or itinerant digital feedback, situation specificity in use of other SEIs, and gathering all the documentation in a portfolio along with faculty selected, administration prescribed, and developmentally important evidence will provide a much broader, deeper and more beneficial pool of information with which to evaluate instruction. It will not, however, resolve the concerns related to defining teacher effectiveness or the interpretation and application of the information.

References:

_____, (1981). Assessments of performance in teaching field study instrument. *State of South Carolina*. Abstract retrieved from ERIC on the Ebscohost database.

Bangura, A.K. (1994). The focus group approach as an alternative to collecting faculty evaluation data to improve teaching. *Center for Educational Development and Assessment Conference on Faculty Evaluation, November, 1994.* Retrieved February 4, 2005 from ERIC on the Ebscohost database.

Bell, T.L. & McClam, T. (1992). Peer review of teaching at UTK: An assessment.

Tennessee University Learning Research Center, Spring, 1992. Retrieved February 5, 2005 from ERIC on the Ebscohost database.

Benjamin, H.L., DeCristoforo, J.R. & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*(2), 221-234. Retrieved February 5, 2005 from the Professional Development Collection on the Ebscohost database.

Bosshardt, W. & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *Journal of Economic Education*, *32*(1), 3-18. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Ciscell, R.E. (1987). Student ratings of instruction: Change the timetable to improve instruction. *Community College Review I*, *I15*(1), 34-38. Abstract retrieved February 4, 2005, from ERIC on the Ebscohost database.

Cosser, M. (1996). Introducing the teaching portfolio in the university: A preliminary investigation. *South African Journal of Higher Education*, *10*(2), 130-137. Retrieved February 4, 2005 from ERIC on the Ebscohost database.

Deegan, W.L. (1974). Evaluating community college personnel: A research report. Abstract retrieved from ERIC on the Ebscohost database.

Drews, D.R., Burroughs, W.J. & Nokovich, D. (1987). Teacher self-ratings as a validity criterion for student evaluations. *Teaching of Psychology*, *14*(1), 23-25. Retrieved February 4, 2005, from PsychARTICLES on the Ebscohost database.

Eiszler, C.F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483-503. Retrieved February 4, 2005, from the Professional Development Collection on the Ebscohost database.

Elliot, C. (1988). Review of relevant research on teaching evaluations for selected research variables. *AECT-RTD Newsletter*, *12*(3). Abstract retrieved from ERIC on the Ebscohost database.

Everett, M.D. (1977). Student evaluation of teaching and the cognitive level of economic courses. *The Journal of Economic Education*, Spring, 1997, 100-103. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Felton, J., Mitchell, J. & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, *29*(1), 91-109. Abstract retrieved February 4, 2005, from the Professional Development Collection on the Ebscohost database.

Fish, S. (2005). Who's in charge here? *ACADEME TODAY: The Chronicle of Higher Education's Daily Report for subscribers*, February 4, 2005.

Gould, C. (1991). Converting faculty assessment into faculty development: The director of composition's responsibility to probationary faculty. Paper presented at the Annual Meeting of the Conference on College Composition and Communication March 21-23, 1991. Abstract retrieved from ERIC on the Ebscohost database.

Hobson, S.M. & Talbot D.M. (2001). Understanding student evaluations. *College Teaching*, *49*(1), 26-32. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Hooper, P. & Page, J. (1986). Measuring teacher effectiveness by student evaluation.

Issues in Accounting Education, 1(1), 56-65. Retrieved February 4, 2005, from the Professional Development Collection on the Ebscohost database.

Lamberth, J. & Kosteski, D.M. (1981). Student evaluations: An assessment of validity. *Teaching of Psychology, 8*(1), 8-11. Retrieved February 4, 2005, from PsycARTICLES on the Ebscohost database.

Layne, B.H., DeCristoforo, J.R. & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*(2), 221-235. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Linn, R.L. & Gronlund, N.E. (2000). *Measurement and Assessment in Teaching*. Upper Saddle River, NJ: Prentice Hall

Obenchain, K.M., Abernathy, T.V. & Wiest, L.R. (2001). The reliability of students' rating of faculty teaching effectiveness. *College Teaching*, *49*(3), 100-105. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Olivares, O.J. (2003). A conceptual and analytical critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning. *Teaching in Higher Education, 8*(2), 233-245. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Orpen, C. (1981). Student evaluation of lecturers as an indicator of instructional quality: A validity study. *Journal of Educational research*, *74*(1), 5-7. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Rodin, M. & Rodin, B. (1973). Student evaluations of teachers. *The Journal of Economic*

Education, Fall, 5-9. Retrieved February 4, 2005, from the Professional Development Collection on the Ebscohost database.

Shavelson, R.J. & Juang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change*, January/February, 2003. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Shevlin, M., Banyard, P., Davies, M., & Griffiths M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment and Evaluation in Higher Education*, *25*(4), 397-408. Retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.

Szeto, W.F. (1994). Perceptions of faculty performance evaluation among faculty across academic disciplines at a selected university. *Annual Meeting of the Mid-South Educational Research Association, November 9-11, 1994*. Retrieved February 4, 2005 from ERIC on the Ebscohost database.

Tomasco, A.T. (1980). Student perceptions of instructional and personality characteristics of faculty: A canonical analysis. *Teaching of Psychology*, *7*(2), 79-83. Retrieved February 4, 2005, from the Professional Development Collection on the Ebscohost database.

Wergin, J.F. (2005). Giving credit to accreditation. *ACADEME TODAY: The Chronicle of Higher Education's Daily Report for subscribers, 2/2/05*.

White, M.E. (2002). Evaluating the bilingual teacher: The monolingual administrator’s challenge. *An Imperfect World: Resonance From the Nations Violence (Proceedings of the annual meetings of the NAAAS, NAHLS, NANAS, and IAAS in Houston, Texas, February 11-16,2002).* Retrieved February 4, 2005, from the Professional Development Collection on the Ebscohost database.

Williams, W.M. & Ceci, S.J. (1997). How ‘m I doin’. *Change, I29*(5), 13-25. Retrieved February 4, 2005, from the Academic Search Premier collection on the Ebscohost database.

Wilson, R. (1998). New research casts doubt on the value of student evaluations of professors. *Chronicle of Higher Education*, *44*(19), A12-A15. Abstract retrieved February 4, 2005 from the Professional Development Collection on the Ebscohost database.