Student Evaluation of Teaching Effectiveness
SETE

SETE Information Handbook

10-3-11

Student Evaluation of Teaching Effectiveness (SETE)
SETE Information Handbook

A handbook developed by the University of North Texas for the University of North Texas Student Evaluation of Teaching Effectiveness (SETE) survey instrument.

Office of Institutional Research and Effectiveness
Dr. Allen Clark, Assistant VP
Elizabeth Fisher, Director for Institutional Effectiveness
Dr. Mary Barton, Director for Institutional Research

SETE Team
Dr. Ronald Carriveau, Assessment Development Manager and PI
Dr. Richard Herrington, Lead Psychometrician
Dr. Craig Neumann, Psychometrician
Paula Iaeger, Lead Researcher and co-PI

TABLE OF CONTENTS

SETE

INTRODUCTION

This handbook provides information for UNT faculty on the development of the UNT Student Evaluation of Teaching Effectiveness (SETE) and on the interpretation and use of SETE scores. Validity evidence to support the SETE score interpretation and use is provided, including information on the development process, the construct and dimensions being measured, and scoring methodology. Also included is psychometric information on the structure of the survey, model fit, and score scale development. For more in-depth psychometric information, please contact the Office of Institutional Research and Effectiveness.

PART 1: SETE DEVELOPMENT

THE NEED FOR A VALID INTER-DEPARTMENTAL ASSESSMENT TOOL

The goal of the SETE development project has been to develop a psychometrically valid faculty teaching evaluation instrument for inter-departmental usage here at UNT. From the outset of the project, the SETE committee's goals were to utilize assessment and measurement best practice in the development and evaluation of a SETE assessment tool. The SETE was to be part of a larger faculty evaluation program overseen by the Provost's office. The SETE development initiative was originally planned to span a three year period that included the following research activities: i) initial development; ii) pilot data collection; iii) item selection/refinement; iv) psychometric model selection/refinement; and finally; v) model validation/calibration supplemented with parallel validation evidence from other relevant methodological perspectives.

The SETE instrument design was developed through an iterative process of focus- group evaluation sessions with subsequent redesign based on focus-group findings. These focus-group meetings occurred over a nine month period in which faculty members, students, and administrator's met to evaluate the initial instrument's design specifications. The eventual goal was to produce a theoretically and empirically based evaluation of the domains of student perceived teaching effectiveness, as indicated by the relevant academic literature, in conjunction with the aggregate experiences and advice of UNT faculty and students. This collaborative outcome produced an acceptable pool of 28 items that addressed three categories of teaching effectiveness behaviors, with a general effectiveness domain subsuming the three sub-domains. It is important to note that these three effectiveness domains have both prior theoretical and empirical support in the academic treatments of teaching effectiveness.

While the main goal of the SETE team was to produce a useful assessment tool to meet the needs of UNT faculty evaluation committees, it was clear that if the SETE were to play a larger role in ongoing UNT campus-wide evaluation standards, then it would need to allow for inter-departmental comparisons of student's perceptions of teacher effectiveness. The approach taken by the SETE team has been to develop a population (or site based) normative instrument, where the across-departmental-influences and student-

demographic-influences are minimized as much as possible in their combined effect on measuring student's perceptions of teaching effectiveness.

An additional consideration is that serious constraints exist in any attempt to standardize SETE administration for a 35,000+ student population. The goals of SETE administration standardization, and the need to maintain a cost-effective, and efficient delivery mechanism have lead the SETE team to adopt a web-delivery platform for SETE administrations. Additional cost-effectiveness and standardization concerns were motivated by the Texas legislative house bill 2504 mandating public web-access to faculty effectiveness ratings (to become legally effective fall of 2011). This development increased the concern of faculty and administration that the SETE team's ongoing effort to produce a valid assessment device be carried out.

PURPOSE, SCORE USE AND LIMITATIONS

The purpose of the SETE is to produce valid scores for measuring teaching effectiveness on a scale that crosses all course sections at the University of North Texas. The intent is that the scores can be applied to a continuous improvement model that shows individual instructor growth over time. Descriptions of the effectiveness factors that are measured are provided to give meaning to the scale scores and to provide information for making instructional decisions. It is recommended that SETE scores be treated as one piece of evidence in evaluating teachers and that they should not be used as the only measure when making teacher evaluation decisions. When using SETE scores, it is important to keep in mind that they are on an equal interval scale and a growth of ten points anywhere on the scale is the same amount of growth as ten points on another part of the scale regardless of the course taught.

BACKGROUND

A University of North Texas Evaluation of Teaching Committee (EOT) which consisted of six faculty, one student, two staff, and three administrators was formed in April 2008 to develop a student evaluation of teaching effectiveness instrument. The committee was to ensure a transparent process that would include faculty and student input and provide reports to the Faculty Senate. The EOT completed its work in September 2008, after which a Validity Study team was assembled.

The EOT was charged with providing to the Provost of the University of North Texas (UNT) a recommendation for an assessment tool to facilitate student evaluations of their instructors and would allow university-wide comparisons in key areas. The committee made its recommendation to the Provost on Oct. 1, 2008, and it was subsequently approved.

After a review of the literature and input from committee members, it was determined by the EOT that the survey should focus on measuring teaching effectiveness and that course effectiveness should be treated separately and by a different committee. It was also determined that the survey instrument should be structured on the dimensions and elements presented on pages 51-53 in Berk (2006) as synthesized by

Davis (1993) from research on good teaching (Chickering & Gamson, 1991; Eble, 1988; Murray, 1991; Reynolds, 1992; Schon, 1987) and on student achievement and success (Noel, Levitz, Saluri, & Associates, 1985; Pascarella & Terenizini, 1991, Tinto 1987). Berk's book, *Thirteen Strategies to Measure College Teaching,* was selected as a handbook and guide for the project.

It was determined that every effort should be made to find extant surveys and published lists of survey items and to evaluate them for usefulness versus writing new items. Approximately 3,000 survey items were collected and evaluated, including all current UNT department surveys and published surveys and lists that are used by over 100 universities. Primary evidence for the validity of the results of this item selection process included faculty and student input regarding the dimensions, elements, applicability, and quality of the existing statements. This evidence was collected through seven faculty focus groups, four student focus groups, faculty and student interviews, results from a survey sent to all faculty, surveys sent to students, and an item tryout field-test administered to students, plus scoring-rubric results from the committee members' evaluations of items.

After an initial screening process, the 3,000 item pool was narrowed to 1,488 items. Evaluating these items with rating scales reduced this number to 788 items. A second evaluation matching items to specific elements reduced the number to 346. Using specific scoring criteria to qualify items for inclusion, committee members reduced the number of items to 51. These 51 items were then presented to faculty and student focus groups and to students in a developmental field test; based on the results, a final draft set of 38 items was selected. A final review was then conducted using the criteria of student viewpoint, student observable, measurability of the statement, and conformity to the research elements, duplicity, and universality in terms of class size and in terms of online to in-class administration. This process resulted in the final pool of 28 items. Over 400 people were involved in the process.

The second phase of the SETE development included three teams made up of faculty and staff who specialized in assessment development and psychometrics. Team A conducted the psychometrics; Team B administered a spring pilot test of the SETE items and conducted follow-up faculty and student focus groups; and Team C developed open-ended response items. The final 28 SETE items were pilot tested using a stratified sampling across the University. The pilot test was administered at the end of the Spring semester 2009, and a validity study team was assembled to analyze the data, validate the model fit, conduct item reduction studies, and develop a scoring methodology. The result of the psychometric work was a validation of the 28 items as usable and the selection of 12 of these items for the SETE survey that was administered across the university in the fall of 2009. Based on research (Haladyna,) it was determined that the fewer the number of items the lower the occurrence of test taking fatigue and halo effect (more on this in later sections of this manual). It was determined that even though statistically fewer than 12 items would work, twelve items were recommended to ensure adequate domain representation.

The results from the fall 2009 UNT administration, a fall 2009 administration by Texas Woman's University (also in Denton, Texas), and a second UNT administration in the spring of 2010 provided data for the ongoing validity studies conducted by Team A. An ongoing research agenda for the continued study of SETE score-validity and the use of particular methodologies and algorithms assembled and developed by Dr. Richard Herrington of UNT for the SETE is in place and is led by Dr. Herrington.

THE ITEM SELECTION PROCESS

The following process was used to select the items for the SETE.

Phase One: This was an initial sort of the original pool of 3000 items based on whether a particular survey item was measuring instructor effectiveness versus elements associated with course effectiveness. A second sort criterion was whether the item required a written response versus selecting a point on a scale. This process narrowed the original pool to 1,488 items.

Phase Two: Items were evaluated using rating scales to determine the degree to which they fit the construct and dimensions found in the literature on which the SETE was to be modeled. This included a measure of content match, syntax, semantics, and a measure of usefulness. The dimension targets were the three factors that are in the current SETE: organization and explanation of materials; learning environment; and self-regulated learning. This process reduced the item pool to 51 items.

Phase Three: The 51 items were then evaluated by faculty and students in focus groups and interviews. A standardized and IRB approved protocol was used, and two of the primary response categories were whether the faculty and students felt that the statements measured teacher effectiveness and whether students could assess what the statements were presenting. For example, items that asked whether the textbook was appropriate and whether the teacher knew the topic was rejected by both groups as not something students could assess. Additionally, in phase three, the 51 items were administered in a survey format to students from a business management class as a developmental field test (also known as an item tryout test). Students in this sample were at the sophomore, junior, and senior academic levels. The surveys were scored, and the results were projected on a screen and there was open discussion on the quality and measurability of the items. These processes reduced the item pool to a draft set of 38 items. Over 400 faculty and student were participants in this process.

Phase Four: A final review was conducted by the SETE development committee, and the items were ranked on the following criteria: student viewpoint or perception of the behavior described by the item statement; whether the item was something that students could observe; the degree to which the behavior related to the item statement could be measured; conformity to the research model (factors and dimensions); duplication of item content; and universality in terms of whether the items could apply equally to different class sizes and to an online administration as well as an in-class administration. This process reduced the item pool to the final recommend set of 28 items that were submitted for a field test administration so the responses could be used for item validation and model fit.

THE SURVEY ITEM RESPONSE SCALE

For the final survey-form development, it was determined that four anchor points were appropriate on a response scale of 1) Strongly Disagree, 2) Disagree, 3) Agree, and 4) Strongly agree. The following points were considered when making this decision.

Four point scale. Research shows that after five points there are diminishing returns in terms of reliability. Additionally, information may be lost if the scale exceeds the respondents ability to discriminate among the anchor points. (Berk, 2006) A 28 item survey with a 4-point scale can yield high reliability coefficients (Herrington, 2009).

No midpoint position on the scale (i.e. neutral, uncertain, or undecided). Information is lost when a midpoint position is included in a set of bi-polar (i.e. both positive and negative) anchors that are intended to measure the degree (intensity) of a respondent's opinion. The neutral mid-point is also problematic because it will lower the mean for a teacher who receives a high score and adds no compensation for a teacher who received a low score. From a measurement viewpoint, nothing is gained from a neutral response. Berk (2006) states that, "For rating scales used to measure teaching effectiveness, it is recommended that the *midpoint position be omitted* and an even-numbered scale be used, such as 4 or 6 points."

No NA (not applicable) choice. The use of NA was avoided because the teacher effectiveness scale will be used for a class level analysis, and every time a student chooses NA, the student's scale score will be different because one or more of the items will not be part of the score. This is a major problem in terms of measurement, analysis, and validity. Since there are class conditions across the university (even on the teacher effectiveness only scale) that would require an NA option, the committee followed recommended procedures for identifying which items might require an NA so they could be eliminated from the final item selection. These procedures included faculty and student review groups in which Faculty were asked to identify those items which they felt could not be observed by students across all classes and thus would require an NA, and students were asked to identify those items which they felt could not be observed by students across all classes and thus would require an NA. Identified items were eliminated.

No item reversal (negative and positive items) to minimize response set bias. This bias is referred to as acquiescence, the tendency to agree or give positive responses regardless of the content of the items (similar to Halo effect). A strategy used to minimize the effect of this survey taking behavior is to word half of the statements positively and the other half negatively (but in random order). However, this method does not eliminate (or reduce) the bias, it simply cancels out the effect of the bias with the result that the effect of the bias is reduced to zero. Berk (2006) recommends that reversals may be appropriate for some scales, but not for teacher effectiveness scales because the positive/negative reversals can be confusing and result in increased response time and response errors. The teacher effectiveness scale is designed to rate the teacher's positive behaviors, not negative ones.

FACTORS AND ITEM SETS

As was stated on page 2 of this handbook, a primary reference is Berk (2006). Over 100 sources are listed in the References for this handbook, but the Berk book was used as a practical guide for survey development. Based on sources cited in Berk (page 51-53), the following researched based clusters were adopted as a starting point for what the SETE should measure.

1. Organizing and explaining material in ways appropriate to students' abilities: knows the subject matter; can explain difficult concepts in plain, comprehensible terms; gauges student's background knowledge and experiences; identifies reasonable expectations for students' progress; selects appropriate teaching methods and materials; devises examples and analogies that clarify key points; relates one topic to another; assesses whether students are learning what is being taught.

2. Creating an environment for learning: establishes and maintains rapport with students; responds to students' needs; communicates high expectations; gives appropriate feedback; respects diverse talents and learning styles; emphasizes cooperation and collaboration; uses strategies that actively engage learners.

3. Helping student become autonomous, self-regulated learners: communicates goals and expectations to students; directs students in making their own connections to course content; views the learning process as a joint venture; stimulates students' intellectual interests.

These clusters evolved into the following three factors: *Organization and explanation of materials; Learning environment;* and *Self regulated learning.*

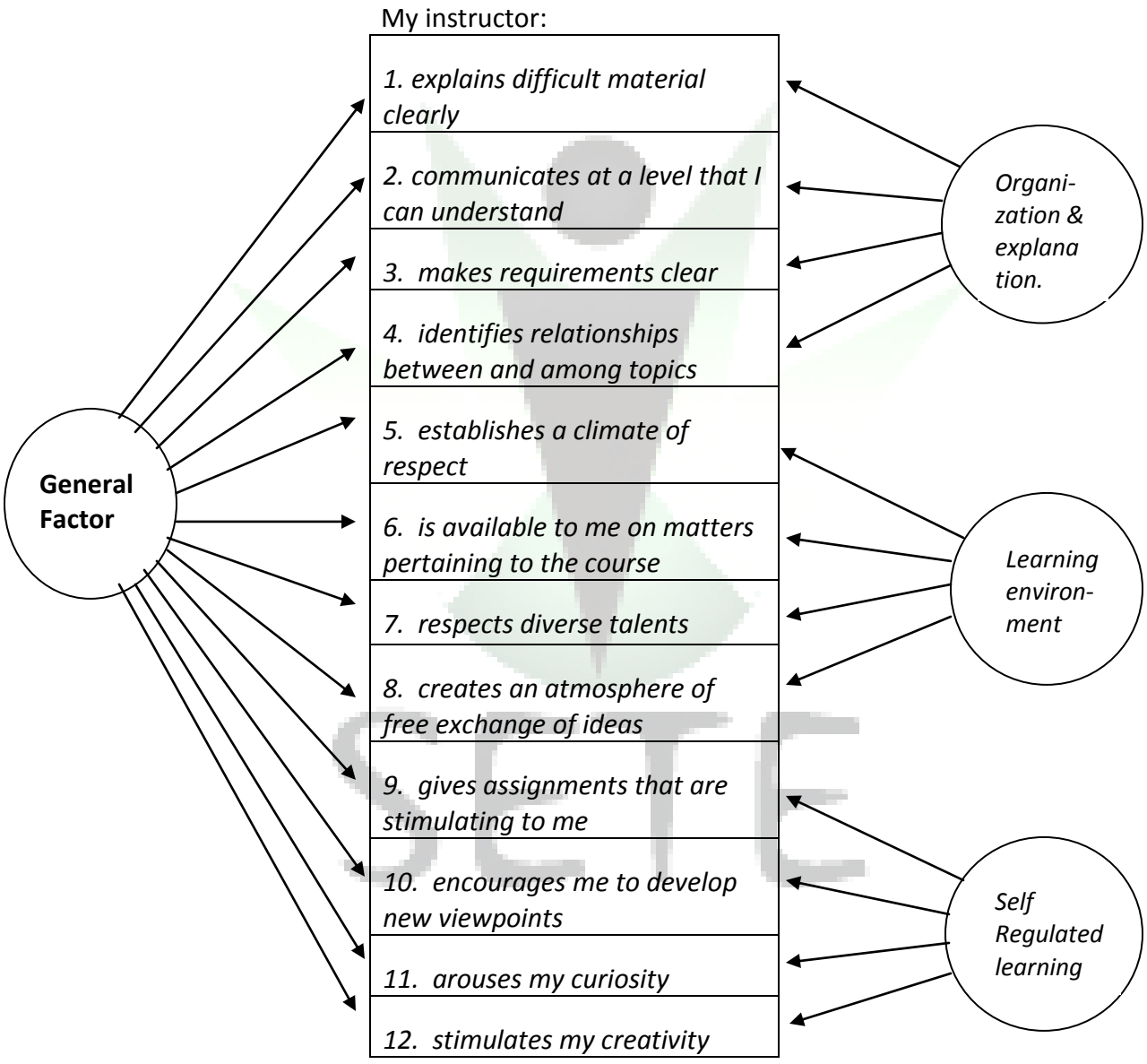
For the first field test in spring 2009, all 28 items in the final item pool were administered so they could be validated. All 28 items were found to be usable for their intended purpose, and analysis showed that they loaded on the three factors of the model as expected. These 28 items and the factors are provided in the table below.

TWENTY-EIGHT STATEMENTS

Factor	28 STATEMENTS	
Organization and explanation of materials	1	My instructor explains difficult material clearly.
	2	My instructor communicates at a level that I can understand.
	3	My Instructor makes requirements clear.
	4	My instructor communicates clearly the expectations for learning in this course.
	5	My instructor assigns activities that are helpful.
	6	My instructor gives assignments that contribute to my understanding of the subject.
	7	My Instructor provides materials that help me understand the subject.
	8	My instructor identifies relationships between and among topics.
	9	My instructor explains new ideas by relating them to familiar concepts.
	10	My instructor evaluates my work in ways that are helpful to my learning.
Learning environment	11	My instructor establishes a climate of respect.
	12	My Instructor is available to me on matters pertaining to the course.
	13	My instructor encourages me toward maximum achievement.
	14	My instructor is skillful in motivating me to do my best work.
	15	My instructor provides useful feedback to guide my progress.
	16	My instructor respects diverse talents.
	17	My instructor creates an environment of mutual respect.
	18	My instructor creates an atmosphere in which ideas can be exchanged freely.
	19	My instructor actively engages me in learning.
	20	My instructor encourages students to actively participate.
Self regulated learning	21	My instructor is skillful in guiding me to be more self-directed in my learning.
	22	My instructor encourages me to connect course topics to a wider understanding of the subject.
	23	My instructor is open to the viewpoints of others.
	24	My instructor encourages me to contribute to the learning process.
	25	My instructor gives assignments that are stimulating to me.
	26	My instructor encourages me to develop new viewpoints.
	27	My instructor arouses my curiosity.
	28	My instructor stimulates my creativity.

SETE ONE GENERAL FACTOR AND THREE SPECIFIC CONTENT FACTORS

The following figure shows the SETE Factor (latent variable) Model for the 12 item set that was selected from the pool of 28 items. These 12 items were used for the fall 2009 and the spring and fall 2010 live administrations.



SETE – Form A, 12 items, final for Fall 2009

This Class is: A Required Course An Elective I am not sure
What grade do you expect to earn in this course? A B C D F

SD = *strongly disagree* **D** = *disagree* **A** = *agree* **SA** = *strongly agree*

	Organization and Explanation of Materials	**SD**	**D**	**A**	**SA**
1	My instructor communicates at a level that I can understand.				
2	My instructor communicates clearly the expectations for learning in this course.				
3	My Instructor provides materials that help me understand the subject.				
4	My instructor identifies relationships between and among topics.				
	Learning Environment				
5	My instructor establishes a climate of respect.				
6	My instructor is available to me on matters pertaining to the course.				
7	My instructor creates an environment of mutual respect.				
8	My instructor creates an atmosphere in which ideas can be exchanged freely.				
	Self Regulated Learning				
9	My instructor is skillful in guiding me to be more self-directed in my learning.				
10	My instructor encourages me to connect course topics to a wider understanding of the subject.				
11	My instructor arouses my curiosity.				
12	My instructor stimulates my creativity.				

	Overall op	**D**	**D**	**A**	**SA**
1	I like this instructor				
2	I am interested in this subject				
3	I think the classroom was appropriate for this class				
4	I think this was a challenging course				
5	I would recommend a course taught by this instructor				

PART 2: SCALE SCORE REPORTING, INTERPRETATION, AND USE

SETE REPORTS

On SETE reports, teachers are provided the following information to help them understand the score scale:

The SETE Scale Score replaces the raw mean score and allows the SETE scores to be put on an interval scale. This process is similar to the standard scores used on the SAT, GRE, and state tests like the STAR. Each of the three effectiveness factors has its own unique scale score. The overall construct of Teacher Effectiveness also has its own scale score, and thus is not simply the average of the factor scores. A measurement model with appropriate external control variables is used in determining how items should be weighted when calculating individual scale scores. This estimation process provides a reasonably fair and unbiased estimate of the individual scale scores as well as providing a high degree of reliability and generalizability to the scale scores. Each scale ranges from 1 to 1000.

Additionally, on the report that provides a teacher's scale score, there are descriptions of the factors that are being measured. Each of these descriptions includes the item content of the four items that are on the survey for each factor plus the content of the rest of the 28 items. In other words, the four items for each factor of the survey are treated as a representation of all the items that could have been used for the test. Thus, the descriptions can be used for self evaluation and improvement. The ability to develop equivalent forms from the 28 items will be discussed in later sections of this handbook.

Organization and Explanation of Materials

This score reflects the student's perception of how well the instructor: makes the course requirements and student learning outcomes clear to the students; gives assignments, activities, and materials that are helpful and that contribute to understanding the subject; explains difficult material clearly; shows the relationships among topics and new concepts; and evaluates student work in ways that are helpful to learning.

Learning Environment

This score reflects the student's perception of how well the instructor: establishes a climate of mutual respect and encouragement; motivates students to work and engage in learning; is available and encouraging; is skillful in actively engaging students in learning; and provides useful feedback.

Self-regulated Learning

This score reflects the student's perception of how well the instructor guides and encourages self-directed learning in which the student is encouraged: to be open to the viewpoints of others; to develop new viewpoints; to connect course topics to a wider understanding of the subject; and to contribute to the learning process.

SETE SCALE SCORE RANGES

For the University of North Texas, three categories of effectiveness were chosen: Somewhat effective; Effective; and Highly Effective. In addition to an overall score scale range, score ranges were calculated for each factor. The factor ranges were calculated from an analysis of raw scores rankings.

***This Class is**:*	*A Required Course An Elective I am not sure*
What grade do you expect to earn in this course?	*A B C D F*

SETE Scale score ranges for effectiveness levels by factors

	Organization and Explanation	Learning Environment	Self-Regulated Learning	Overall Effectiveness
Highly Effective	710 - 981	659 - 972	747 - 998	702 - 998
Effective	438 – 709	347 - 658	495 - 746	406 - 701
Somewhat Effective	167 - 437	35 - 346	243 - 494	111 - 405

ADDITIONAL ITEMS ON THE SETE FORM

On the SETE survey form there are six items in addition to the 12 survey items. Responses to these items are used for research and validation purposes. They are not part of the SETE scale score.

Two items are used to validate the bi-factor model fit.

Four items are used for model fit purposes. They are not part of the SETE scale score, but they are reported on the teacher reports because teachers indicated that they would find the information useful. The results are reported as percentages for each scale point.

	Overall opinions	***SD***	***D***	***A***	***SA***
1	*I like this instructor*	O	O	O	O
2	*I am interested in this subject*	O	O	O	O
3	*I think the classroom was appropriate for this class*	O	O	O	O
4	*I would recommend a course taught by this instructor*	O	O	O	O

Overall opinion results are reported to teachers in percentages.

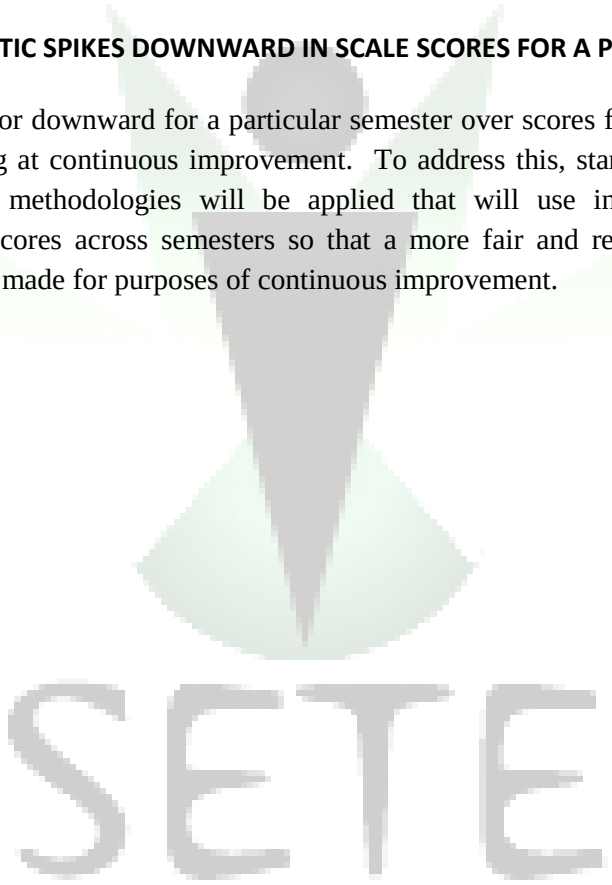
Percent of respondents who said they like this instructor	__%
Percent of respondents who said they were interested in this subject	__%
Percent of respondents who said the classroom was appropriate for this class	__%
Percent of respondents who said they would recommend this instructor	__%

APPLICABILITY OF SETE ITEMS TO COURSES DELIVERED ONLINE

Application of the SETE items to online courses was a major consideration of the committee. Expertise in delivering online instruction was well represented in the committee. Additionally, input was gathered from faculty and student groups. Several online courses were included in the SETE field test in order to do a comparison of online versus not-online student responses. The structural equation modeling used to confirm the structure of the student responses included online courses and confirmed the usefulness of the SETE survey items for online courses. Faculty and student review groups were convened at the beginning of the fall semester 2009 to confirm the usefulness of SETE survey items for online courses.

CONCERN FOR DRAMATIC SPIKES DOWNWARD IN SCALE SCORES FOR A PARTICULAR SEMESTER

A dramatic spike upwards or downward for a particular semester over scores from several semesters can be a concern when looking at continuous improvement. To address this, starting with the Sprint 2011 administration, prediction methodologies will be applied that will use information from previous semesters to smooth the scores across semesters so that a more fair and reasonable interpretation of effectiveness scores can be made for purposes of continuous improvement.



PART 3: BASIC INFORMATION ON TECHNICAL ASPECTS OF THE SETE

GENERAL MODELING CONSIDERATIONS

Good practice in modern statistical modeling necessitates fitting a range of interesting depictions of the data (e.g. statistical models) and then using formal model comparison techniques (e.g. likelihood methods, information theory methods – AIC, BIC) along with theoretical considerations to evaluate, rank, and select useful models. A major concern with the validation of the SETE was assessing the degree to which the model might be "miss-specified." The idea of model miss-specification is connected to the practices of internal and external validity assessments. Methods used for these kinds of validity checks were: cross-validation, jackknife, and other related re-sampling techniques (permutation and bootstrap re-re-sampling).

In general, "bi-factor solution" will usually provide a superior model fit but with loss of parsimony over simpler models (e.g. random effect intercept factor model). That is, a bi-factor solution can have a larger AIC or BIC index (and usually do, because this bi-factor technique involves estimating a larger number of parameters in comparison to similar models). For this reason, the SETE modeling process included fitting multiple competing models and selecting those models which accord with theoretical intentions and have more desirable statistical properties such as a lower AIC or BIC index. Various stratification strategies such as "non-metric propensity score weighting" combined with "independent random groups (RG)" replications were used. "Re-sampling" based strategies (e.g. the "Bootstrap" and "permutation based approaches") were used to get more accurate estimates of model fit, and more accurate coverage of confidence intervals for item weights and the "latent scaled scores" which are partly based on estimated item weights.

The factor structure of SETE is representative of more than 100,000 responses collected over three terms and is currently representative of two different institutions, University of North Texas and Texas Woman's University. Confirmatory Factor Analysis Fit (CFA) results show RMSE<.04 (less than .05 is considered excellent), and GOF>.97 (greater than .05 is considered excellent).

TARGET POPULATIONS, SAMPLES, AND BIAS

The SETE validation process applies survey sampling practices that continually address the nature of the representativeness of survey findings by addressing the relationship between the "sample" and the "population". The basic notion is that with proper selection (some random selection mechanism – i.e. simple random sampling), a subset of the whole (all individuals that could be measured or polled), can reasonably reflect, with some established precision, the attitudes, beliefs, or perceptions of the entire group of individuals who haven't actually been measured. A SETE sample estimate is more than likely not equal to the exact measurement that we could have if we could actually measure everyone in the

population. Survey researchers will frequently work hard to make sure that a planned sample collection is going to be "appropriately" representative of some target population before-hand.

A solution used for the SETE is to randomly sample within specific strata or natural groupings within the population (i.e. "stratified sampling" or "cluster sampling"). That is, the researcher apportions the random samples that are to be collected, within some variable or groupings that are known to vary as subpopulations within the larger population. The researchers hope is to have the sample's respondents accurately reflect a proportional allotment in comparison to a target population's allotment of respondents (e.g. membership of respondents within the levels of "department", within the target population).

SETE sampling makes use of various stratification strategies - e.g. "non-metric propensity score weighting" combined with "independent random groups (RG)" replications. "Re-sampling" based strategies (e.g. the "Bootstrap" and "permutation based approaches") are used to get more accurate estimates of model fit, and more accurate coverage of confidence intervals for item weights and the "latent scaled scores" which are partly based on estimated item weights.

SAMPLE SELECTION BIAS AND INVARIANCE

One of the main goals of the SETE project is to appropriately account for sample selection bias while utilizing modern latent variable models to better estimate SETE in the UNT population. This requires the establishment of a valid measuring instrument that displays a fair degree of "measurement invariance" with respect to the domain under consideration (teacher effectiveness). That is, if we have good measures of SETE, then our measured values should not systematically vary across strata or natural groupings; but should measure the "true values" of the construct across individuals well, while being relatively insensitive to contexts or groups within which individuals find themselves. From this perspective, a framework which allows us to emulate the conditions under which an experiment could have been performed (quasi-experimental) was used, the idea being that any information (design information or otherwise) that can be used in conjunction with the SETE can help establish a well defined relationship between the sample and the population for which the SETE is being used to measure. Rubin's (1983, 1984) propensity score models provided a convenient framework, and set of tools, for estimating unknown selection probabilities that can produce non-random samples in survey data. SETE research findings are that sizes of factor loadings do not vary substantially across sub-samples (e.g. demographic variables – department, student major, gender, etc.). Additionally, factor loading patterns do not vary substantially across sub-samples.

ASSESSING RELIABILITY AND GENERALIZABILITY

The Omega h coefficient was used to assess SETE score reliability and generalizability. "Omega" coefficients such as the Omega h are important indices that are routinely used for this purpose. The Omega coefficient is a multipurpose coefficient in that it assesses both reliability and the generalizability of a set of items. Of interest in this study is how well the scales measure its general factor. Omega h provides a quantitative measure of how well the general factor is estimated by items. Moreover, the

Omega coefficient can be generalized for both multiple factor and hierarchical factor solutions. More importantly, for the purposes of estimating generalizability coefficients, the Omega coefficient, as estimated through SEM parameter estimation methods (e.g. maximum likelihood), allows separate error terms to be estimated for items, whereas a "generalizability theory" approach to generalizability coefficients, based on ANOVA models, uses pooled error terms to estimate marginal effects and joint effects. It is understood that pooled error terms in parameter estimation is important for identifying potentially important contextual effects. Research has shown that the Omega coefficient for the 12 item SETE varies between .90 and .95.

ASSESSING DIMENSIONALITY AND GOODNESS OF FIT

Dimensionality assessments based on "eigenvalue>=1" criterion (Kaiser-Guttman criterion) and scree-plots are really not sufficient to assess multidimensionality. These "rule of thumb" approaches are poor substitutes for more formal models that model the sampling variability of eigen-values. Additionally, there are other dimensionality assessment methods for which item clusters can be based on a variety of different distance metrics that are not tied to correlation as a measure of distance (e.g. cluster analysis). Techniques such as cluster analysis can provide additional sources of information in assessing inventory dimension.

It was determined that simple heuristics should not be used to assess the dimensionality of the SETE inventory. A potentially "high profile" inventory such as the SETE inventory should use a combination of methodologies and look for convergent evidence across these methods. Formal tests of dimensionality were used in conjunction with other procedures that utilize simulation or re-sampling to model eigen-value sampling variability (e.g. parallel analysis), and or clustering algorithms that extract dimensionality assessments based on clustering algorithms (e.g. ICLUST algorithm).

A probability model (IRT model) was used to give a convenient metric for prediction purpose (e.g. predicted probability). Effects sizes that are in terms of odds ratios (e.g. risk factors) are usually more meaningful in terms real world outcomes. Furthermore, IRT methods provide more information than item factor analyses (e.g. information curves, item curves, DIF tests) for purposes of item selection.

In general, goodness of fit indices (GOF) "hide" the nature of potential "misfit" in the sense that they are global measures of fit. Which items/cases contribute to low GOF values is identified by a thorough analysis of model residuals (e.g. factor model residuals, IRT fit residuals). The initial appraisal of fit indices are informative as long as they are followed by residual analysis to identify poorly performing items and/or cases that have "odd" response profiles in relationship to the bulk of the response profiles.

SHORT FORM ITEM SELECTION AND ANT COLONY OPTIMIZATION

Short form survey is used whenever the administration of the long form is problematic (e.g. fatigue)

The development of the 12 item short form version of SETE is a common practice. The developer of the short form should demonstrate the relative equivalence between the long form (e.g. 28 items version of SETE) and short form (e.g. the 12 item version of SETE) in terms of reliability and generalizability. The 12 and 28 item SETE both have reliability and generalizabilty greater than .90.

The "iterative" use of factor analysis (e.g. three factors with 4 items per factor) in the selection of short forms can be problematic (e.g. fitting – removing item(s) – refitting - etc.) because each successive fit ignores all possible/potential multivariate relationships between items, which produces sub-optimal subsets of items. The 12 item SETE was selecting using a heuristic optimization method called Ant Colony Optimization (ACO) where Optimal subsets were selected from the final list of 28 items such that certain properties were maximized or minimized. Maximized were: large item loadings (correlation between observed score and latent score minus measurement error); large correlations of factor scores with other external measures; and large model fit indices (e.g. goodness of fit: 0-1). Minimized were: small RMSE (distance between model and data).

ACO selection is automated and produces *near optimal* results – based on estimating all possible configurations of items and selects those configurations that optimize important psychometric criterion. Multiple subsets (parallel forms) can be obtained automatically as well; these multiple subsets can be rank ordered in terms of fit.

SAMPLING DESIGN FOR WEB-BASED DELIVERY AND INVERSE PROBABLILITY WEIGHTING

Delivery of course surveys via web based access presents sampling issues. The sample must include all units in the population of interest. Coverage error occurs whenever the sample does not include parts of the population of interest, for example, when students are asked to evaluate an instructor and not all of the class participants respond. Nonresponse error occurs when sampling units (e.g. respondents) are contacted for the survey but either provide no data or only partial data. Under coverage and nonresponse both contribute to missing data from units that should be in the survey and can result in biased estimates if those units differ systematically from units that are in the sample and that respond to the survey. A predominant approach in the survey sampling literature is to use case weighting and re-sampling methods to estimate and reduce this bias in coverage (e.g**.** post-stratification using inverse probability weighting). Poststratification can partially alleviate coverage bias, but one does not know whether these adjustments truly compensate for this coverage bias unless one obtains data on the persons not covered by the sample. Post-stratification adjustment forces estimates from the reweighted sample to equal the population estimates for the different demographic classes. Survey sampling methodology provides algorithms that are useful in estimating *non-responder bias* in survey samples. Inverse probability weighting (IPW) can be used with auxillary background information to estimate "probability response classes" **(**e.g. background demographic information – dept., college, major, gender, class size, grade assigned, grade expected, etc). The inverse of these probabilities down-weights high probability response classes, and gives more weight to low response probability classes. The effect of IPW is to reduce the relationship of the background variables with the principal outcome measure of interest (teaching effectiveness). IPW also reduces the effect of bias on the model being estimated (e.g. relatively unbiased item loadings in the factor analysis).

EXTERNAL CONTROL VARIABLES (BACKGROUND VARIABLES)

The 12 item SETE modeling process looked at course level, faculty level, and student level background variables. Nine background measures were selected from a total of 17 variables. Prior to this, the 17 variables were pre-selected from 30 variables based on relevance. The Course level variables are: course size, in class vs. internet, instruction type, time course held. The Faculty level variables are: status (e.g. lecturer , assist. prof., full prof.), age, number of years employed at institution, gender of instructor. The Student level variables are: anticipated grade, actual grade assigned, mean GPA, academic level, current course load, students gender, total credits earned, pre-requisites present (yes/no). Department and student major were handled by using "multi-level ANOVA" methods (contextual modeling).

Selecting the "best" external control variables is important since non-relevant variables increase error variability and reduce the predictive validity of the SETE items. Bayesian Model Averaging was used (BMA) to select the best subset of variables from 17 in predicting general teaching effectiveness (G). BMA model selection strategy can select models that have relatively better prediction accuracy, compared to models with smaller posterior probabilities.

The most relevant background variables accounted for about 9% of the total variance in general teaching effectiveness (G): 6 student level variables, 2 course level variables, and 1 faculty level variable. A *relative importance metric* was generated for the 9 variables. This metric decomposes the variance accounted for in G (9%) into non-overlapping components of variance, and the relative importance allows a rank ordering of the importance of variables.

The following are the proportion of variance explained by the model: 9% metrics are normalized to sum to 100%. IPW with these 9 external control variables reduced the 9% variance accounted for in G to about 2-3% variance accounted for in G.

Relative Importance Metrics

Student anticipated grade	.82
Class size	.07
Student course load	.02
GPA mean	.02
Course internet (other)	.02
Student pre requirement	.01
Faculty age	.01
Student credits earned	.01
Student gender	.01

Potential SETE External Control Variables used to control for bias

1. Student ID
2. Faculty ID
3. Unique Course ID
4. Course College
5. Course Size

6. Course Time
7. Course Internet Other
8. Course Department
9. Room
10. Faculty Gender
11. Faculty Age
12. Faculty Rank
13. UNT Years
14. Instruction Type
15. Instruction Department
16. Student Credits Earned
17. Student Course Load
18. Student Pre Requirements
19. Student Major
20. Student Gender
21. Student Academic Level
22. Student GPA Mean
23. Student Anticipated Grade
24. Course Grade Assigned to Student

CONTEXTUAL EFFECTS: DEPARTMENT AND STUDENT MAJOR

Context effects refer to the differential influence of "level" specific variables (contextual variables) on outcome measures of interest. For example:

- students are nested within courses
- students can be nested within student majors
- courses are nested within departments
- departmental influences on teaching may vary across departments
- students will respond more "similarly" (as compared to other students in other courses) since they are exposed to the same instructor
- different courses may have different courses sizes
- the same course can vary in size across semesters

The nested structure of response units creates what is known as "within-class correlation" (known as intra-class correlation). Within-class correlation may not bias mean estimates, but will likely create large bias for confidence intervals (conf. intervals too narrow). The predictive accuracy of course means will also be lower. Contextual effects can be modeled with **"multi-level" ANOVA** methods (see next topic below).

MULTI-LEVEL ANOVA

Analysis of variance (ANOVA) models between group and within group variation were used. Ideally, we would expect that between-department variation in SETE ratings would be small as compared to within

department variation on SETE ratings. This creates within course correlation (intra-class correlation – also known as intra-class reliability). Courses are nested within departments and vary in course size, which contributes to differences in response consistency across courses (i.e. differences in reliability across courses). A strategy in dealing with varying reliabilities and varying course sizes is to base course ratings on a weighted average. This weighted course average is based on:

weighted_course_mean = (*r*) * *course_mean* + (1 – *r*) * *pop_mean*

Where *r* is the course reliability, c*ourse_mean* is the mean of the course, and *pop_mean* is the mean of all course means.

Reliabilities are a function of course size and response consistency. Courses averages with low reliabilities are moved toward the population mean. Course averages with high reliabilities do not move as much toward the population mean. Weighted averages calculated in this manner are sometimes referred to as "Hierarchical Bayes" estimates or "Empirical Bayes" estimates. Heirarchical Bayes estimates "borrow strength" across groups (pooling information across groups) to estimate group means. Heirarchical Bayes estimates are very good at reducing error (i.e. prediction error).

SCALE SCORE DEVELOPMENT

The SETE factor model produces estimates of the unobserved true scores for teaching effectiveness (factor scores). These factor scores are linearly related to the observed scores and can be linearly rescaled to an arbitrary mean and arbitrary standard deviation. Factor scores are transformed into "z-scores" - mean of 0 and standard deviation of 1. These "z-scores" are further rescaled by multiplying by the desired standard deviation and adding in the desired mean: scaled score = mean + (s.d. * z_scores).

Factor scores are calculated for each respondent within a course. Average factor scores (within a course) are estimated using weighted means (Hierarchical-Bayes means). Average factor scores are linearly scaled to have a mean of 800 and standard deviation of 50 (this standard deviation captures the lower and upper values such that scores are greater than 0 and less than 1000)– label this **S1**. Additionally, the average factor scores are linearly scaled to have the same mean and standard deviation of the original 4 point scale (raw data) – label this **S2**. Scores on the **S1** scale are identified that correspond with the anchor points on **S2** (4, 3, 2, 1).

Cutoffs for *Highly Effective*, *Effective*, and *Somewhat Effective* are established and correspond to anchor point boundaries of 4-3, 3-2, 2-1 on **S1.**

MISSING VALUES

Accounting for non-response (missing values) is important for reducing bias in model estimates (e.g. means, factor loadings). Simple (but inadequate) methods for dealing with missing values include: removing records with missing data, and mean substitution. Better methods exist that take into account

the multivariate patterns in the complete and missing data when making a **"**data imputation" (e.g. maximum likelihood, multiple imputation).

Missing data patterns in SETE data are estimating using "k-nearest neighbor imputation" within a course. Nearest neighbors are records that have similar completed data patterns. Within a course, the average of the k-nearest neighbor's completed data are used to impute the value for a variable that is missing its value. k-nearest neighbors assumes missing at random (MAR) – i.e. missing data only depends on the observed data; able to take advantage of multivariate relationships in the completed data. The drawback of k-nearest neighbors is that does not include a component to model random variation, consequently uncertainty in the imputed value is underestimated.

FURTHER REFINEMENTS FOR IMPLEMENTATION

Future implementations of SETE will focus on using SETE scores to predict teaching effectiveness one time period ahead (e.g. following semester). This is a dynamic measurement model. Current course ratings become a *prior distribution of ratings* used in predicting the following semester ratings. The current course ratings are calculated as a weighted average of the previous semester's ratings (prior) and the current semester's ratings (data) to produce a *posterior estimate of the current semester's rating.*

This posterior distribution of ratings is essentially a weighted average that lies between the previous semester rating and the current semester rating. This posterior distribution of ratings become the prior distribution for the following semester. This estimation procedure creates a *moving average across semesters (weighted average).* This moving average reduces unwanted variability across semesters due to fluctuating course size and also minimizes the effect of outlying semester ratings

REFERENCES

Arreola, R., Theall, M., & Aleamoni, L. M. (2003). *Beyond Scholarship: Recognizing the Multiple Roles of the Professoriate*. Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL

Berk, R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians*. Sterling, Va: Stylus Pub.

Chickering, A. W., & Gamson, Z. F. (1991). *Applying the seven principles for good practice in undergraduate education*. New directions for teaching and learning, no. 47. San Francisco: Jossey-Bass.

Davis, B. G. (1993). *Tools for teaching*. The Jossey-Bass higher and adult education series. San Francisco: Jossey-Bass.

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002). College students' attitudes toward methods of collecting teaching evaluations: In-Class versus on-line. *Journal of Education for Business*. 78 (1), 11.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38,* 327–333.

Ewell, P. T. (2001). *Accreditation and Student Learning Outcomes: A Proposed Point of Departure. CHEA Occasional Paper*. Council for Higher Education Accreditation, Washington, DC.

Feeley, Thomas H. (2002). Evidence of halo effect in student evaluations of communication instruction. *Communication Education, 5,* 225-236.

Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician* 61, 139-147.

Haladyna, T., & Hess, R. (2000). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment*. 6, 129-153.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with Discussion). *Statistical Science,* 14, 382--401.

Hyndman, Koehler, Ord, Snyder (2008). *Forecasting with Exponential Smoothing,* Springer Series in Statistics, Springer, New York.

Jackman, Simon (2009). Bayesian Analysis for the Social Sciences, John Wiley & Sons, United Kingdom.

Kerlinger, F. N. (1973). *Foundations of behavioral research: 2d ed*. New York: Holt, Rinehart and Winston.

Leite, Huang & Marcoulides (2008), Item Selection for the Development of Short Forms of Scales Using an Ant Colony Optimization Algorithm, *Multivariate Behavioral Research* , v43 n3, p411-431.

Mangan, Katherine. (2009). Professors compete for bonuses based on student evaluations. *The Chronicle of Higher Education*, 55(21), A.10. Retrieved July 22, 2009, from Research Library. (DOI 1643398441).

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal* , 16, 57-69.

Marsh, Herbert W. & Hocevar, Dennis. (1991). Student's evaluation of teaching effectiveness: The stability of mean rating of the same teachers over a 13-year period. *Teaching & Teaching Education, 7*, 303-314.

Mateo, M. A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational and Psychological Measurement*, 56, 5, 771-78.

McConnell, C., & Sosin, K. (1984). Some determinants of student attitudes toward large classes" *The Journal of Economic Education*. 15 (2), 181-190.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Roderick P. McDonald (1999). *Test Theory: A Unified Treatment.* Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

Särndal**,** Carl-Erik**,** Swensson**,** Bengt & Wretman**,** Jan (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics, Springer, New York.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. The Jossey-Bass higher and adult education series. San Francisco: Jossey-Bass.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*. 44 (4), 409-432.

Theall, M. and Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instructors? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (New Directions for Institutional Research, No. 109, 45-56. San Francisco: Jossey-Bass.

Wolter (1997). *Introduction To Variance Estimation*, Springer Series in Statistics, Springer, New York.