Date: 10-3-2011
Status: Draft
Author: Richard Herrington, PhD

Part I. General Method For Calculating the Effectiveness Categories

The population statistics of the UNT SETE scores (scaled from 1-4), for Spring 2011, are based on N=51302 records and 12 items. The mean and standard deviation for the weighted average score are calculated by summarizing weighted items within respondents, then summarizing these weighted averages across respondents. The weights for the items are produced by a measurement model fit to the data (one general factor - G, and three sub-factors - F1, F2 & F3).

The UNT SETE score population statistics for the weighted average (N=52301) are:

mean: 3.36
stdev: .31 (on the original scale, scaled from 1 to 4)

Steps in Establishing the Thresholds for the Effectiveness Categories and Validating the Composition of Item Responses Within Those Effectiveness Categories

Step 1: **The SETE scores (scaled 1-1000) are rescaled to have the same mean and standard deviation as the population statistics of the original scale, scaled - 1,2,3,4.** The is done by using a z-score transformations technique. The shape and distributions of the transformed (or rescaled) SETE scores will not change; however, the SETE score scaling will now be scaled - 1,2,3,4 - with a mean and standard deviation that corresponds to the population statistics of the original scale (see Figures 1A & 1B - notice the perfect correspondence in shape).

Step 2: **Cut-points are then established on the original scale (scaled 1,2,3,4) such that percentages of item responses within the effectiveness categories, as determined by the thresholds, will correspond meaningfully to the descriptions of the anchor points on the original scale, as seen by the respondents (i.e. 1 - Strongly Agree, 2 - Disagree, 3 - Agree, 4 - Strongly Agree).**

These Effectivenss groups were calculated by setting thresholds on the original scale (scaled 1,2,3,4) at 2.97, and 3.58. The rescaled SETE scores were compared to these cut-points to produce 3 effectiveness categories in the following manner (Table 1):

Table 1. Thresholds for rescaled SETE scores

rescaled SETE score < 2.97 ---------->	"Somewhat Effective"
rescaled SETE score > or = 2.97 AND < 3.58 ---->	"Effective"
resclaed SETE score > 3.5 ---------->	"Highly Effective"

Once the cut-points for the Effectiveness categories have been established, we would like to see what the correspondence is between the unweighted item response scores and the rescaled SETE scores (scaled 1,2,3,4), and by extension, the original SETE scores that were scaled from 1-1000.

We would expect to see that the composition of the unweighted item responses in the "Somewhat Effective" category to be mainly "Strongly Disagree" and "Disagree" responses. Likewise, for the "Effective" category, we would expect to see the composition of unweighted item responses to be mainly "Strongly Disagree", "Disagree", and "Agree" responses. And finally, for the "Highly Effective" category, we would expect to see the majority of item

responses being composed of "Highly Agree" Responses. This outcome would establish the correspondence between the orignal scale (1,2,3,4) anchor point descriptors, and the effectiveness category descriptors. To summarize, we expect (Table 2):

Table 2. Correspondence Between Effectiveness Categories and Item Anchor Descriptions

Effectiveness Category	Composition of Unweighted Item Responses
Highly Effective	mostly "Strongly Agree" responses
Effective	mostly "Agree" responses
Somewhat Effective	mostly "Disagree" and "Strongly Disagree" responses

The following steps are aimed at empirically establishing this correspondence.

Step 3: Analysis of Unweighted Response Data For Each Effectiveness Group and Compared to the Rescaled SETE Score Data

Starting with the “Highly Effective”(HE) grouping, individual items are summarized by calculating quantiles for each item within the effectiveness grouping HE. Then individual item quantiles are combined across items to produce an overall set of quantile scores for the HE group.

Table 3. Highly Effective (HE) Group (N=12849 - 25% of the total)

i) Item Quantiles (10 quantiles for each of 12 items):

Item	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
I1	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I2	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I5	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I6	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I7	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I8	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I9	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I10	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I11	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
I12	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

Once the individual item response quantiles are calculated, we need to summarize these individual item response quantiles for the HE group. The following is the set of average unweighted item response quantiles (summarizing across across all 12 item's quantiles using a mean).

ii) Averaged Quantiles for HE Group

5%	10%	15%	20%	25%	30%	35%	40%
2.916667	3.000000	3.000000	3.500000	3.666667	3.916667	4.000000	4.000000
45%	50%	55%	60%	65%	70%	75%	80%
4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000
85%	90%	95%	100%				
4.000000	4.000000	4.000000	4.000000				

In addition to the quantile summaries, we also show the group's overall unweighted item response's average score (averaged across 12 items producing an average score for a respondent, then averaged across all respondents, within the group, to produce an "effectiveness category" group mean:

iii) HE Group Mean

3.743424

Interpretation of Table 3.

We can see that for the HE group, that 65% of all responses correspond to "Strongly Agree" responses, making the composition of the HE group mainly "Strongly Agree" responses with a few "Agree" responses. This distributional pattern for the HE group comes from setting the HE group cut-score > 3.58 for the rescaled SETE scores. For puroposes of comparison, the average of this set of unweighted responses for the HE group is 3.74 which is substaintially larger

than the population mean of 3.36. The distribution of the HE group rescaled SETE scores is displayed in Figure 1C, which has a mean of 3.69, is slightly less than the average of the unweighted average score of 3.74.

Effective (E) Group (N=33,388 - 65% of the total)

Table 4.

i) Item Quantiles (10 quantiles for each of 12 items):

Item	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%
I1	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
I2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
I3	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
I4	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
I5	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
I6	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
I7	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
I8	2	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4
I9	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4
I10	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
I11	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
I12	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	4

Item	95%	100%
I1	4	4
I2	4	4
I3	4	4
I4	4	4
I5	4	4
I6	4	4
I7	4	4
I8	4	4
I9	4	4
I10	4	4
I11	4	4
I12	4	4

ii) Averaged Quantiles for E Group

5%	10%	15%	20%	25%	30%	35%	40%
2.000000	2.500000	2.750000	2.916667	3.000000	3.000000	3.000000	3.000000
45%	50%	55%	60%	65%	70%	75%	80%
3.083333	3.333333	3.583333	3.916667	4.000000	4.000000	4.000000	4.000000
85%	90%	95%	100%				
4.000000	4.000000	4.000000	4.000000				

iii) Effective Group Mean

3.330466

Interpretation of Table 4.

We can see that for the E group, that 35% of all responses correspond to "Strongly Agree" responses, making the composition of the E unweighted response scores mainly "Agree" responses, with a few "Disgree" responses. This distributional pattern for the E group comes from setting the E group cut-score >= 2.97 and < 3.58 for the rescaled SETE scores. For purposes of comparison, the average of this set of unweighted responses for the E group is 3.33 which is close to the population mean of 3.36. The E group rescaled SETE scores is displayed in Figure 1D, which has a mean of 3.34, is very close to the population mean (3.36), and is also close to the unweighted item response mean (3.33).

Somewhat Effective (SE) Group (N=5065 - 10% of the total)

Table 5.

i) Item Quantiles (10 quantiles for each of 12 items):

Item	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%
I1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	3	4
I2	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4
I3	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4
I4	1	1	1	2	2	2	2	3	3	3	3	3	3	3	3	3	4	4
I5	1	1	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4
I6	1	1	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4
I7	1	1	2	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4
I8	1	1	1	2	2	2	3	3	3	3	3	3	3	3	3	4	4	4
I9	1	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	4
I10	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4
I11	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	3	4
I12	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	4

Item	95%	100%
I1	4	4
I2	4	4
I3	4	4
I4	4	4
I5	4	4
I6	4	4
I7	4	4
I8	4	4
I9	4	4
I10	4	4
I11	4	4
I12	4	4

ii) Averaged Quantiles for SE Group

5%	10%	15%	20%	25%	30%	35%	40%
1.000000	1.000000	1.250000	1.666667	2.000000	2.250000	2.333333	2.416667
45%	50%	55%	60%	65%	70%	75%	80%
2.666667	2.750000	2.916667	3.000000	3.000000	3.000000	3.000000	3.333333
85%	90%	95%	100%				
3.666667	4.000000	4.000000	4.000000				

iii) SE Group Mean

2.583646

Interpretation of Table 5.

We can see that for the SE group, that 40% of all responses correspond to mostly "Agree" responses and some "Strongly Agree" responses, making the composition of the SE response scores mainly a mixture of "Strongly Disagree", "Disagree" with a few "Agree" and "Strongly Agree" responses. This distributional pattern for the SE group comes from setting the SE group cut-score < 2.97 for the rescaled SETE scores. For purposes of comparison, the average of this set of unweighted responses for the SE group is 2.58 which is substaintially less than the population mean of 3.36 - the distribution of the SE group rescaled SETE scores is displayed in Figure 1E, which has a mean of 2.69 which is slightly larger than the unweighted item response average of 2.58.

Summary

The preceeding empirical analysis of the UNT SETE faculty evalution scores from Spring 2011, established the validity of setting the current thresholds displayed in Table 1. Analysis of the unweighted item responses indicated a clear correspondence between the gradations of the three effectiveness categories, and the descriptions of the item anchor points, as responded to by the respondents of the SETE 2011 Faculty Evaluation instrument.

NOTE: **The following is not yet complete, but I decided to leave it in to give some idea of where the continuation of this discussion is headed.**

Part II. Reliability Analyses of the SETE Instrument

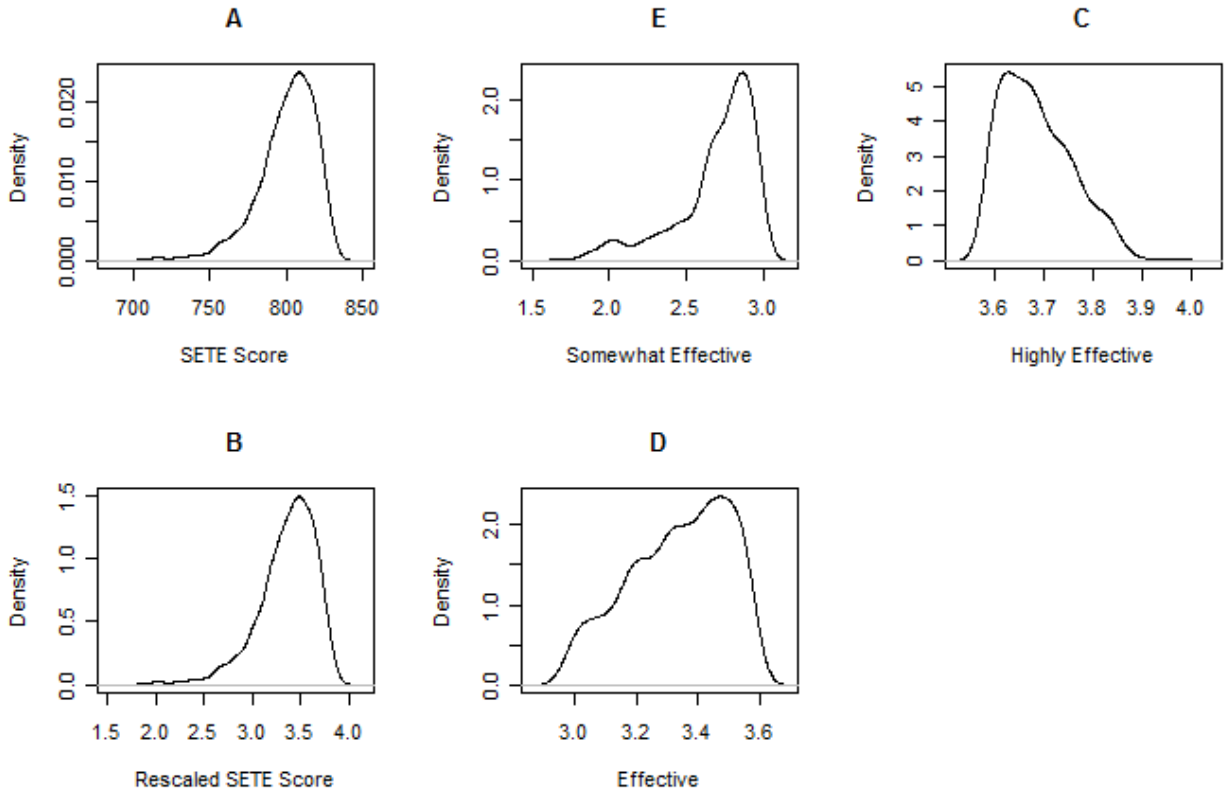
i) Scale Reliability

Scale reliability for the SETE instrument was established by the estimation of and validation of a latent trait measurement model - specifically a linear bifactor, factor analysis model. The bifactor model models a higher order general factor (G) that is "uncorrelated" with the subfactors. Furthermore, these subfactors are (for the most part) uncorrelated amongst themselves. This is in contrast to a higher order, hierarchical factor model, where the general factor G and the subfactors are correlated. An appealing aspect of the bifactor model is that interpretation for the general factor and the subfactors can be improved by minimizing the overlap in information provided by the indicators for these domains is
minimized.

ii) Reliability of the SETE scores in Differentiating Scores Across Assigned Levels of Effectiveness

iii) Reliability of the SETE scores in Differentiating Scores Across Arbitrary Course-Units

Figure 1.


A
Density
0.000
0.010
0.020
700
750
800
850
SETE Score
E
Density
0.0
1.0
2.0
1.5
2.0
2.5
3.0
Somewhat Effective
C
Density
0
1
2
3
4
5
3.6
3.7
3.8
3.9
4.0
Highly Effective
B
Density
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
4.0
Rescaled SETE Score
D
Density
0.0
1.0
2.0
3.0
3.2
3.4
3.6
Effective