Title Page:

Developing and Analyzing a Scale to Measure the Impact of the Advanced Technological Education Program

January 2013

Wayne W. Welch
University of Minnesota

Author Notes

Wayne W. Welch, Department of Educational Psychology, University of Minnesota (ret.)

Professor Welch is also a consultant for Rainbow Research, Inc., Minneapolis, MN

This material is based upon work supported by the National Science Foundation under Grant No. 1132099. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this report should be addressed to Wayne Welch, 621 W Lake St # 300, Minneapolis, MN 55408. Contact: wwelch@umn.edu

Abstract

Using statements generated by team leaders and other stakeholders, I show they can be used on a Likert-type survey to create a reliable and valid scale to measure the impact of the National Science Foundation's Advanced Technological Education program (ATE). The process, called Peer-Generated Likert Scaling, uses these statements to solicit opinions from other ATE grantees. They are asked to respond with the usual options of strongly agree to strongly disagree. However, there was an option to mark, Not Applicable (NA), if the statement did not apply to their grant. This was because the grants vary in size and duration. It was not clear from a literature review how to score NA option on these kinds of survey.

I used two ways to calculate a scale score; a sum of the responses to the items called the Total Impact scale, and the mean of the responses named the Mean Impact scale. Both approaches yielded scores that met the conditions of normality and had high reliability indices, .88, and .84. Because I was working with a population and not a random sample, I presented all findings using descriptive statistics. Effect sizes were used to help interpret the magnitude of differences or the strengths of correlations.

Both scales were able to detect differences between groups, however, there were variations in the results. In general, the Total Impact Scale resulted in larger differences between groups than did the Mean Impact Scale. Excluding items that were answered "Not Applicable" and coded as zero, tended to flatten out disparities between groups. A discussion of why this occurs and the issue of which scoring system to use is discussed in this report.

Size of grant, and other characteristics related to size, e.g., the center/project comparison, were the best predictors of impact scores. The age of grants, and variables related to age, e.g., whether a grant was active or expired, produced small effect sizes. Both scales could be used to determine what other variables are related to impact, for example, whether the grant was sustained or not, and for testing various theories of program impact.

A key issue in targeted research studies is the extent to which one can generalize the findings to other ATE grant sites. Currently, the findings are limited to the 261 grantees included in the study. Increased generalizability would require some kind of study replication. I suggest one way to do this would to compare the results of this study with data gathered on identical items included on the ATE annual survey carried out at Western Michigan University. Another way would be to repeat the study using a random sample of current ATE grantees. If the results were replicated, it would greatly enhance the generalizability of the findings.

Because many of the impact items are evaluative in nature, for example, they address desired goals of the ATE program such as producing more and better-prepared technicians, I propose they could be used to create a Likert-style evaluation survey. Individual project leaders and NSF could use such a survey to help evaluate individual grants and the ATE program.

The Advanced Technological Education (ATE) program, mandated by Congress and administered by the National Science Foundation (NSF), was designed to improve the education of technicians in high-technology fields (U. S. Senate, 1992). It began in 1994 and NSF has made more than 1,200 awards to two- and four-year colleges and other organizations. The primary focus of the program is on two-year colleges. These types of institutions have received about 75% of the ATE grants. The program supports program, curriculum, professional, and materials development and involves partnerships between academic institutions and employers.

In recent years, the Foundation added a "targeted research" program track that supports research on topics that advance the knowledge base needed to make technician education programs more effective and more forward-looking. These projects address research questions or outline a topic of broad interest and importance to the ATE program and its grantees. The topic of interest in this NSF-supported research was to determine it was possible to develop and validate an instrument for assessing the impact (effect or influence) on the people and institutions involved in the ATE program. If so, NSF could use the instrument to identify ways the program has influenced its grantees, help it evaluate the ATE program, and provide information to grantees on how ATE has influenced them and their institutions. In addition, the scales could be used to provide insights to policy makers on how the recent and substantial influx of federal funds has affected community colleges.

Method

Scale Development

A survey was developed following generally accepted procedures for test development (Borg & Gall, 1983). However, I used a new method called Peer-Generated Likert Scaling to generate items for the survey. This procedure asks project stakeholders to write statements about the impact of their ATE experiences. These statements, placed in quotation marks, are used as items for the survey.

The initial development step was to define the dimensions of the concept, "impact." In test and survey development parlance, this is called the domain of content. A working outline of the domain was developed based on a review of the literature and the advice of an advisory panel. I then asked current and former principal investigators (PIs) and others familiar with the program to describe the impact of their ATE experience. I did this using stakeholder interviews and included the following question on an annual survey of ATE grantees.

> "Please reflect on the impact that the grant is having on your academic program, your institution, the community, or other interested parties." These effects of the grant may be positive, negative, or neutral. They may be intended or unintended. Please describe the most important effects of your project."

The above processes generated 95 statements. These were mapped against the working framework to determine how well they fit with the domain implied by the generated statements. This process yielded the following outline.

I. People: Faculty, Students, Administrators, ATE PIs/Staff

II. Program: Curriculum, Instruction, Educational Materials

III. Organizations: Colleges, Schools, Business/Industry, Communities

After several rounds of review by survey experts and persons familiar with ATE grants, the final survey consisted of 29 impact statements. A detailed description of this process is available in (Welch, 2011b). I placed the statements in quotation marks, and put them on a survey for other ATE PIs to decide if the statements described their situation. Respondents were asked to rate their opinions using a five-point Likert scale. The scale ranged from "strongly agree" to "strongly disagree." There was also an option to circle "not applicable" if they thought the statement in question was not applicable to their situation.

A few examples of the survey statements follow.

- "Our faculty has improved their teaching style because of their involvement in our ATE grant."
- "Our NSF grant has given us the confidence to seek and obtain funding from other sources."
- "The ATE grant has increased our sense of worth by being a part of this national effort."
- "The grant provided the catalyst to establish and/or strengthen collaborations with business and industry partners."

The development process provides evidence that the survey possesses content validity, that is, it is measuring what it purports to measure. Another characteristic of an effective measuring instrument is its usability. This includes such things as readability, clear instructions, low response burden, ease of scoring, and the like. The precise development steps used to ensure usability are described in the report mentioned above (Welch, 2011b).

Scale Analysis

The target population for the study included active ATE grantees that began prior to Jan 1, 2009, and grantees that had expired between Jan. 1, 2007 and Jan. 1, 2010. The population size was 261. I mailed the survey to these ATE principal investigators (PIs). Several follow-up contacts were made and eventually 212 completed surveys were returned. The response rate was 81%. A nonresponse bias study was carried out that indicated the larger centers were somewhat more likely to respond to the survey, but there were no differences between respondents and nonrespondents in the nature of their responses. (Welch & Barlau, 2011).

The study population was the 212 sites that returned the survey. The sampling process used is called a purposive sample. It is a non-probability sample, which means that inferential statistics such as t-tests and analysis of variance are inappropriate even though they are often used (Berk & Freedman, 2012). Such procedures use the sample statistics to estimate the population parameters. However, I already have the population parameters, at least for the 212

people who responded to the survey. Because I found no evidence of nonresponse bias between the 212 who responded and the 49 who did not as reported by Welch and Barlau, I could cautiously generalize to the total target population of 261 grantees. However, without a replication of the research, I cannot draw inferences about the more that 1200 awardees who have received ATE grants since the program's inception in 1994.

I used two ways to calculate an impact score. First, I summed the responses across the items to create a Total Impact Score. The response options and the way they were coded follows.

0 – Not Applicable
1 – Strongly Disagree
2 – Disagree
3 – Uncertain
4 – Agree
5 – Strongly Agree

Several items were stated negatively. They were recoded so that disagreeing with those statements yielded a higher score. For example, disagreeing to the statement, "Our NSF/ATE grant has had little long term impact on our college" was coded 4 instead of 2. Items marked Not Applicable were assigned a score of zero. The total score is the sum of the responses to the 29 statements, hence, the higher the score, the greater the perceived impact.

I also computed an impact score based on the average response to the 29 items. I called it the Mean Item Impact score. Statements marked "Not Applicable" (NA) and coded zero, were not included when calculating the mean. The mean score is a measure of the average rather than the total score. A more detailed explanation of the thinking behind these two scores is found in Welch (2011a).

Some statisticians have questioned using Likert items to calculate a scale score. However, Norman (2010) and Uebersax (2006) make convincing justifications. For example, Norman wrote "Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of ''coming to the wrong conclusion.'' These findings are consistent with empirical literature dating back nearly 80 years." (p. 632).

After creating the two impact scale scores, I calculated descriptive statisitics for each and examined their distributions to see if there were any problems such as outliers or unusual distributions. In addition, I computed the reliability of the measures to determine if they met accepted standards for survey development.

Total Impact Scores. The distribution of the Total Impact scores is shown in Figure 1.

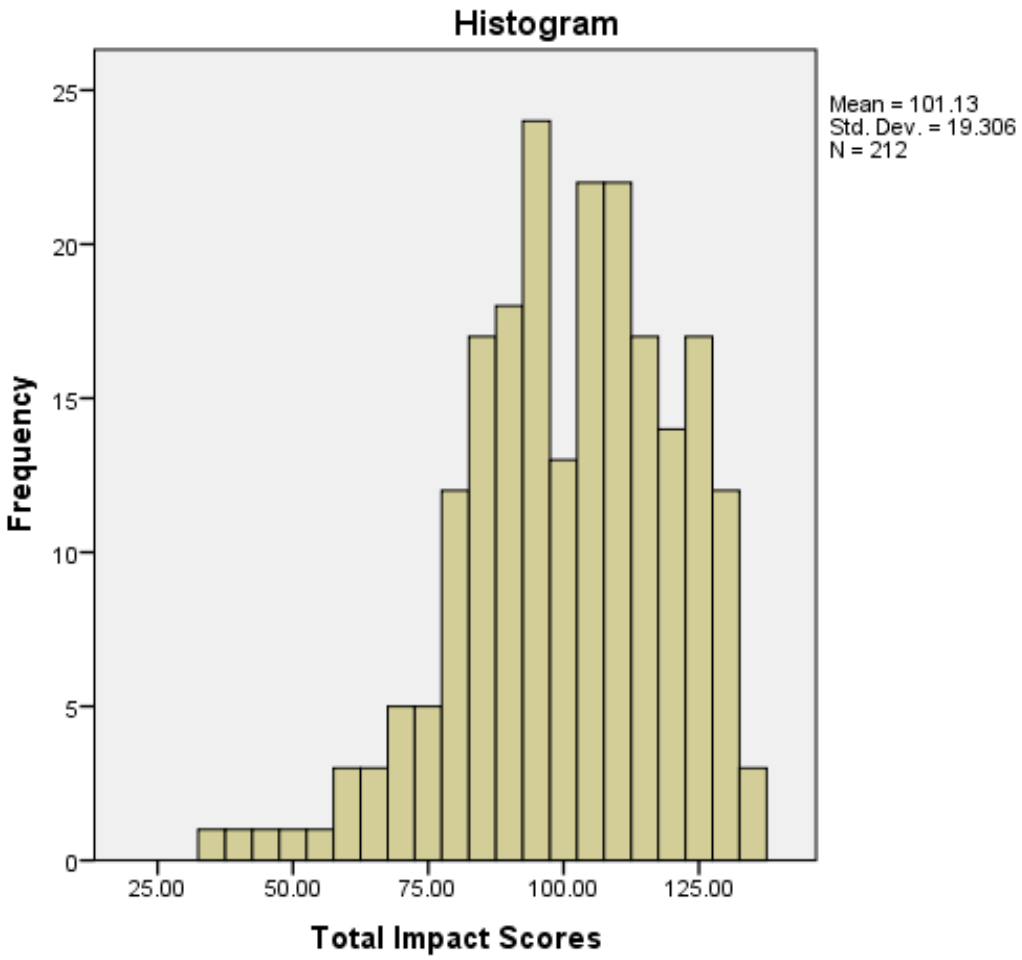

Figure 1. Distribution of total impact scores for 212 ATE grants

The median score was 103. An examination of the frequency showed that 92% of the scores are above the median or 50th percentile. This result and previous research examining item responses (Welch, 2011b) suggests that the ATE program was perceived as having a strong impact on its grant recipients.

The distribution of the scores generally met accepted standards for measuring instruments. The kurtosis of the distribution was acceptable but skewed to the left. I conducted the Kolmogorov-Smirnov test of normality and obtained a value of .089. This is not significant at the .05 level. This means the requirements for normality were attained.

A key requirement for an effective measure is that the scores meet acceptable levels of reliability. I computed a reliability coefficient (Cronbach's alpha) that measures the internal consistency of the items. Missing values were handled using a listwise deletion process. That is, only those respondents that answered all statements were included in the calculation. This process resulted in 68 cases where complete response data were available. I obtained a reliability coefficient of .88. George & Mallery (2003) rate a reliability coefficient between .80 and .90 as

"Good." Nunnally (1978) recommends reliabilities of .70 or higher for preliminary research and .80 or above for basic research (p. 245). The reliability of the total impact scores obtained on the Peer-Generated Likert Survey exceeded these standards.

Mean Impact Scores. A mean item response score was computed for each respondent. This was their average response to the 29 items on the survey. A "Not Applicable" (NA) response was defined as a missing value, which meant that item was not included in the averaging process. I computed the measures of central tendency and plotted the distribution of the mean impact scores. The results are shown in Figure 2.

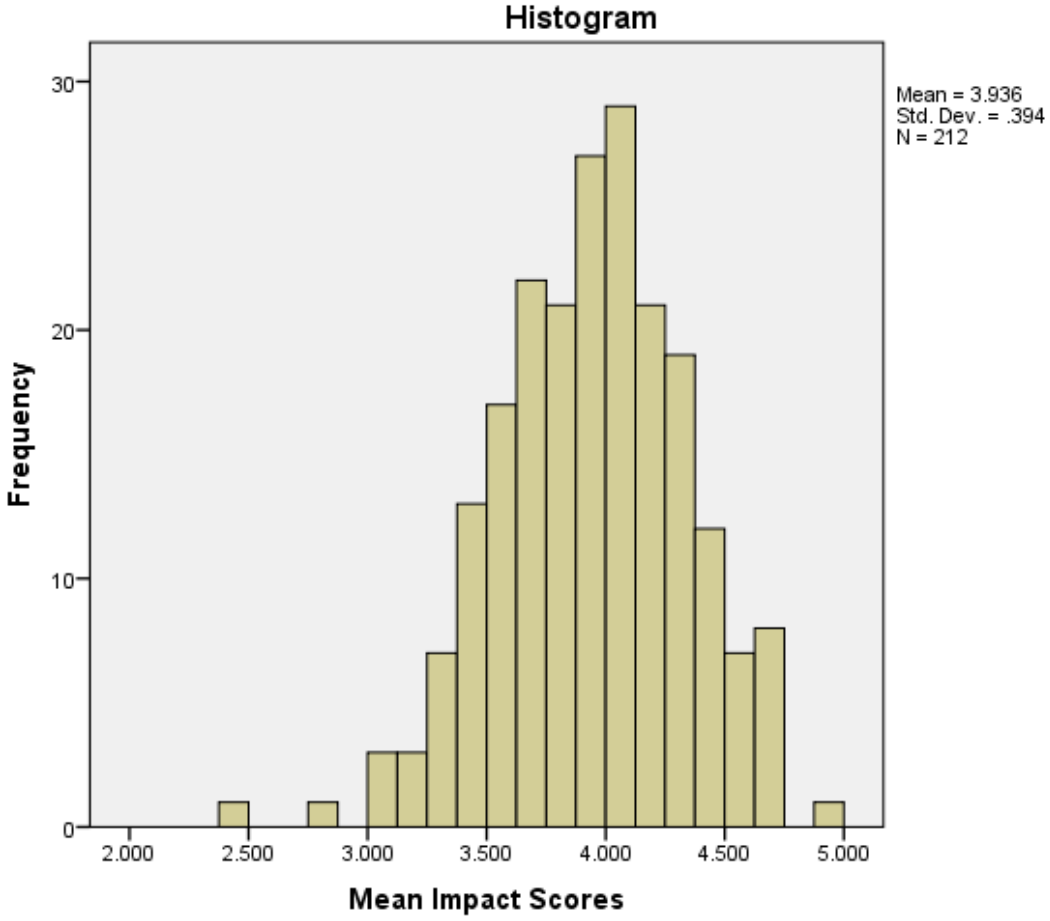

Figure 2. Distribution of scores mean impact scores for 212 ATE respondents

A Kolmogorov-Smirnov test showed the assumption of normality was met. There was one outlier, that is, more than three SD from the mean. This case was removed for the correlational analyses.

The mid-point value for the 5-point scale was 3.0; however, 99% of the respondents were above that value. The average response was 3.94, just slightly below an "agree" or "support the statement" response. The respondents were quite positive in their belief that the ATE experience had positively changed them and their institution. This finding is consistent with previous research on this issue (Welch, 2011a).

Calculating the internal consistency of the mean impact measure is not straightforward because each of the pairwise correlations has a difference sample size. This is mainly due to the use of the NA response. One way to handle this is to calculate all the item inter-correlations using pairwise deletion and compute their average value. I did this and obtained an average item intercorrelation of .152. This can be viewed as the reliability of a one-item scale. An estimate of the reliability (internal consistency) can be calculated for the 29-items scale by applying the Spearman-Brown prophecy formula. The reliability estimate using this procedure was an acceptable .84, exceeding the standards for the reliability of research measures.

Findings

I now turn to the question of whether the scales could discriminate among groups or are related to other variables. Background characteristics of the grantees were found on NSF's FastLane web site. I used this information to examine the scale scores of various groups. These included the following traits.

- Program Track Projects or centers
- Grant Status: Active or expired. An NSF classification for each specific grant
- Grantee Institution: Two-year colleges, four-year colleges, and other
- Size of Grant: Amount awarded in dollars
- Age of Grant: Number of months between initial award and survey date

I report the average scores and their differences for the group comparisons (program track, status, and institution) and used correlation coefficients to describe the extent of relationships for size and age of grant.

I did not use inferential statistics with their attendant t-tests and significant levels because I am working with a population of grantees, not a random sample. I do not need to make inferences about the population means from the sample means because I already know the population means. I used effect sizes (ES) to help interpret the magnitude of any differences. An effect size is a standardized measure of differences usually reported in standard deviation units. I used Cohen's d statistic as my measure for means comparisons (Cohen, 1988). His rule of thumb for interpreting effect size is .20, small, .50, medium, and .80, large.

The correlation coefficients (r) for age and size of grant were reported along with the corresponding effect sizes. Pearson r and Cohen's d statistic are related by the formula $d = \frac{r}{\sqrt{1-r^2}).25}$. Group comparisons and correlation analyses findings are shown in Table 1.

Table 1

Summary of Findings for Total Impact Scores by Background Characteristics.

Group [a]	Mean	Standard Deviation	Mean Difference (Δ) or Correlations (r)	Effect size
Centers (45)	111.64	13.98		
Projects (167)	98.29	19.59	Δ = 13.35	.78
Active grants (131)	102.95	19.86		
Expired grants (81)	98.17	18.11	Δ = 4.78	.25
Two-year colleges(156)	103.54	17.42		
Four-year colleges (39)	98.92	20.97	Δ = 4.62	.24
Colleges (195)	102.62	18.22		
Other (17)	84.00	23.49	Δ = 18.62	.89
Impact score	101.74	18.47		
by size of grant (206)[b]	$758,506	$589,076	r = .30	.63
Impact score	101.22	19.30		
by age of grant (210)[b]	45.99	20.31	r = -.09	-.18

[a] Size of groups in parentheses. [b] Size minus number of outliers

Centers versus projects. An examination of Table 1 shows that a large effect size (ES) was found for program track (.78). Centers perceived their ATE endeavor had a greater influence on them than projects did. This finding is consistent with what one would expect given that centers are generally larger and broader in scope than the smaller and more focused projects. One way to think about this is to express the center mean using the distribution of the project scores. An ES of .78 means that the average center score falls at the 79th percentile of the project scores. I would consider this a significant difference.

Active versus expired. NSF categorizes grants as active or expired based on the dates stated in the grant authorization document. I selected those grants that had been active for more than one year at the time of data gathering, spring 2010. Expired grants were defined as those who had ended between one and four years before the time of the survey. I asked the question, "Did ATE PIs perceive more (or less) impact if an ATE endeavor was currently active or if it had expired?"

I examined the means of the two groups and found a small effect size of .24. Using the interpretation above, I could say that the mean of the active grants falls at the 60th percentile of the excluded grants for my population of ATE grants. It would be useful to see if my findings could be replicated by doing a similar study of current ATE grants. Such a study, if the findings were similar, would help increase the generalizability of the findings. One way to do this is to repeat the study using a random sample of current ATE grants.

Type of institution. I compared the total impact scale scores for the types of institutions that received grants. These included two-year colleges, four-year colleges, and other. The other category included groups such as museums, professional societies, and educational development organizations. I plotted the means of these groups. These are shown in Figure 3.

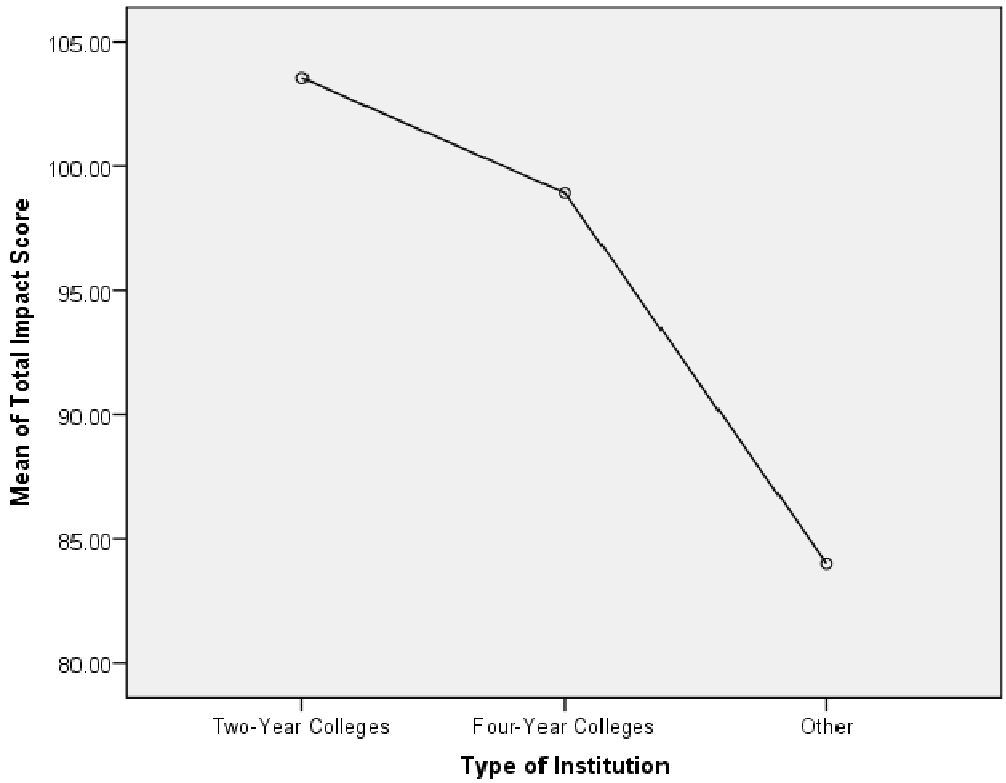

Figure 3. Plot of the average total impact scores by type of institution.

It is possible to compute an effect size when working with three groups using a statistic called eta squared (η^2) which is a measure of variance accounted for. However, it is difficult to conceptualize the results in terms of the means. I followed the recommendation of Howell (2011) and compared two groups at a time; two-year with four-year colleges, and all colleges versus the "other" category. These are the results presented in Table 1.

We noted there was a 4.62 difference in the mean scores between two- and four-year colleges which produced a small (.24) effect size. However, the largest difference for type of institution was between the colleges and the other types of institutions that received grants. The effect size of this comparison was .89, which is a noteworthy difference.

One reason for this result may be the nature of the other organizations. They included such things as museums, professional societies, and educational development organizations. Generally, they received an ATE grant to do a specific task, for example, produce a film series, create a computer-aided performance assessment, or develop a college algebra course. Such projects may not do as well on a measure of total impact because they would have more Not

Applicable (NA) responses.[1] Such responses would not contribute to a total scale score. This is one of the reasons that I also computed a Mean Impact Score. That score is the average response on the items that are relevant to their grant. I report the results for the mean impact scores later in this document.

Size of grant. I investigated the relationship between size of grant and the total impact scores using a correlational analysis. The size of a grant was defined as the amount of money each grantee received. Because several factors can affect correlational studies, I examined the distribution of the grant size variable to check for normalcy and outliers. The distribution of size of grant is shown in Figure 4.

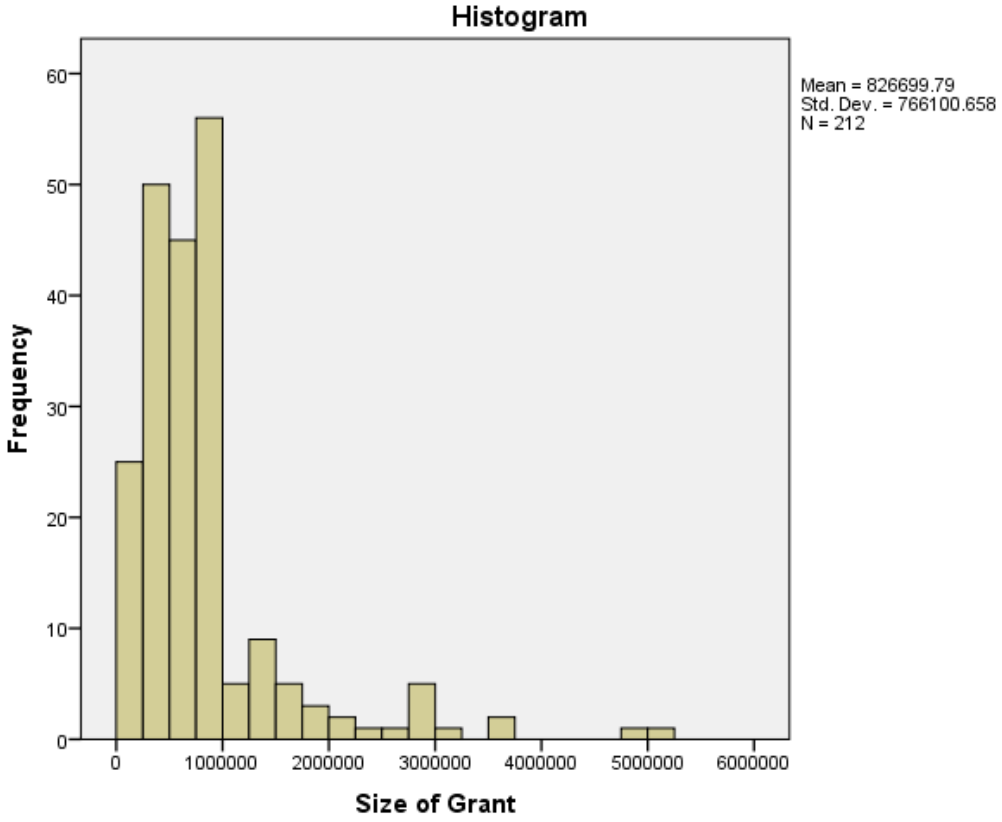

Figure 4. Distribution of the size of ATE grants

We can see this distribution was not normal. The skewness was 2.82 and the kurtosis was 10.00, outside generally accepted limits. There were also four outlier scores, defined as more than three standard deviations from the mean (Tabachnick & Fidel, 1989). I removed those scores leaving my sample size at 208. The new distribution was less skewed and the kurtosis index was reduced to 5.0, still quite large. However, given my large sample size, I did not think it would be a problem to use the Pearson product-moment correlation coefficient.

[1] I computed the number of NA responses and found that the other group averaged 6.9 per survey while the college group had about half that amount, 3.4 per survey.

I also identified two outliers in the distribution of the total impact scores. Because these can distort correlational studies, I removed them as well. This left 206 cases to use to examine the relationship between total scores and size of grant. The value of the Pearson correlation coefficient was .30. A scatterplot of the relationship along with the best-fit line is shown in Figure 5.

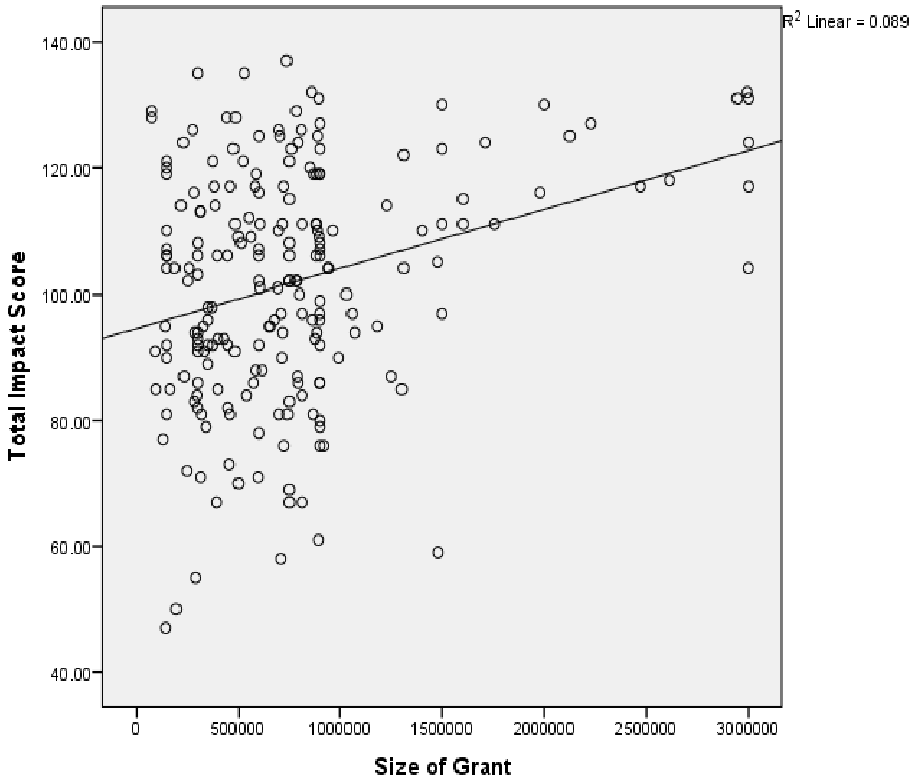

Figure 5. Scatterplot with best-fit line of the relationship between total impact score and size of grant.

One way to think about the strength of correlations is to use r-squared. R-square is the variance in the survey scores accounted for by grant size, in this case 9%. One cannot imply causation here, that is, spending more money will produce a greater impact. However, there is a relationship between these two variables. There may also be other factors affecting this relationship, for example, the project/center differences. Centers score higher but they are also larger on average. One way to address this is to use partial correlation analysis.

First, I computed the partial correlation (pr) between the total impact score and the center/project difference controlling for size. I obtained a coefficient of .14. I repeated the analysis by correlating total score with size of grant controlling for the center/project difference. This yielded a coefficient of .15. Each of these variables accounts for about 2% of the variance in the scale scores. When they were used together, they accounted for about 9% of the variance.

Age of grant. I examined the relationship between the age of a grant and the total impact score. Age of grant was defined as the number of months between initial award and survey date. Two outliers were removed leaving a sample size of 210. The Pearson r between total impact score and age of grant was -.09. The negative sign means the older the grant, the lower the score. I examined the distribution of the age variable to see if there was something that might explain the low correlation. The distribution is shown in Fig. 6.

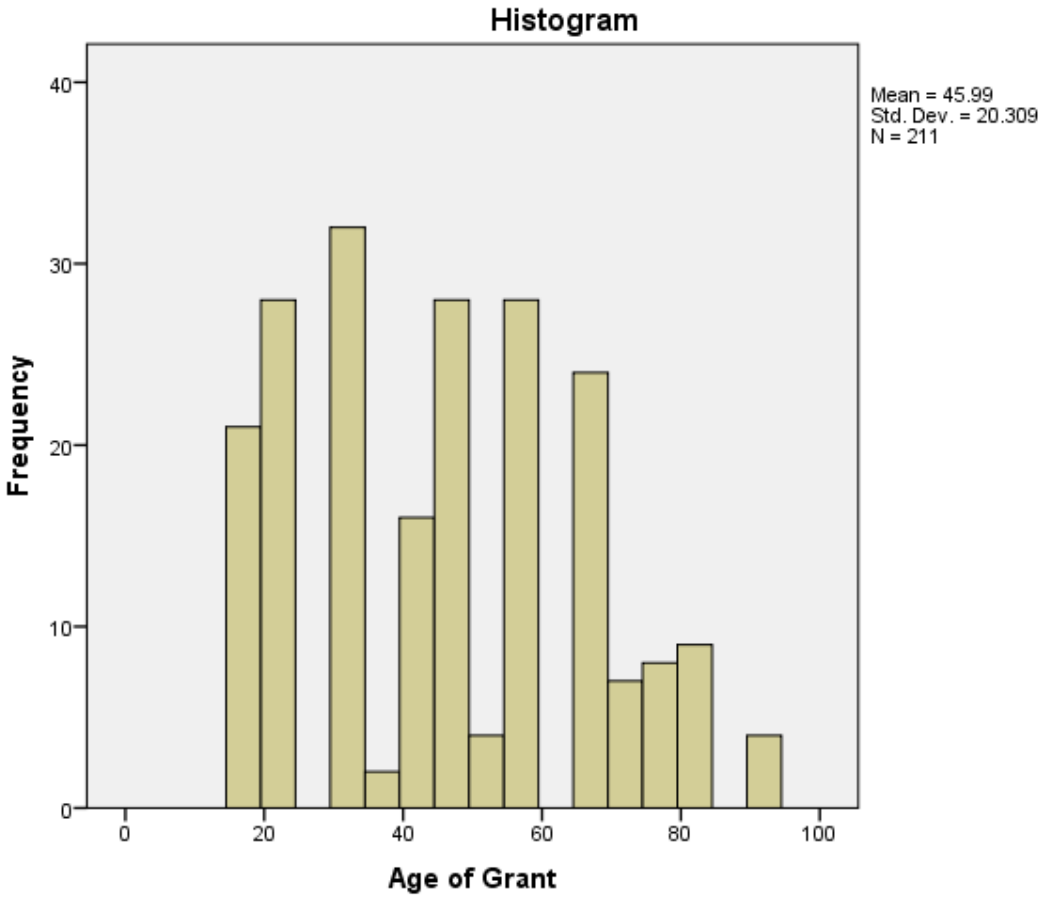

Figure 6. Distribution of Age of Grant in Months

This unusual distribution reflects the way that NSF makes its grants. It approves a set of grants at a given time, usually in late spring or early summer. In addition, the grants are made for a specific period, from one to four years. The result of this process is mirrored in the figure with the vertical columns at about 20 months, 32, 45, 57, 68, and 80 months. These are about one year apart reflecting the usual funding process. However, nothing in the distribution was found that might be causing spurious findings.

There are two variables in the set of background traits I used that were related to the age of the grant. One was age and the other was grant status, active or expired. Expired grants were usually older than active ones. The averages were 34 months for active and 66 months for expired grants.

Again, I used partial correlational (pr) analysis to determine the relative contribution of size and grant status. I obtained a pr of .002 between the total impact score and the age of the grant controlling for the active/expired difference. However, when I computed partial r between total scores and the active/expired difference, I obtained a value of -.082. This means that practically all of the age variable relationship is accounted for by the active/expired difference. The older expired grants scored lower than the younger active ones that created the negative correlation.

I suspect the situation is a result of the way that NSF classifies it grants. When a grant is authorized, NSF assigns a starting date and an expiration data. These were the dates we used when classifying the grants. However, these dates do not always match what is going on in the field. They usually require some start up time before work actually begins and most of the time; they receive various kinds of extensions. Even more important, many sites apply for and receive new grants so the site is still active but the specific grant was not. Further work is needed before putting much credence in this finding to determine more exactly the meaning of "active" and "expired" from the PI's perspective.

I have shown that it was possible to develop a total impact scale created by using peer-generated statements as items on a Likert-type survey that met the standards for reliability and validity. I now turn to a different way of scoring the survey based on the mean responses rather than the total to see if this approach yields a scale that also meets the required psychometric standards. I also compare the results for the mean impact scale with those found for the total impact scale.

Mean Impact Scores

I repeated the above process using the Mean Impact score as the dependent variable. Recall this is a measure of the average response on the 5-item scale that runs from "Strongly Agree" coded 5, to "Strongly Disagree" coded as 1. The NA response was coded as a missing value. It is a relative measure of impact rather than a summated measure. It was developed to try to account for differences in the scope of activities among the various grants.

I used the same background characteristics to determine if the mean impact scale could detect differences among groups as I did for the total impact scores. These findings are summarized in Table 2.

Table 2

Summary of Findings for Mean Impact Scores by Background Trait

Group [a]	Mean	Standard Deviation	Mean Difference (Δ) or Correlations (r)	Effect size
Centers (45)	4.07	.42		
Projects (167)	3.90	.38	Δ = +.17	.42
Active grants (131)	4.00	.39		
Expired grants (81)	3.83	.40	Δ = +.17	.44
Two-year colleges(156)	3.94	.39		
Four-year colleges (39)	3.96	.43	Δ = - .02	-.05
Colleges (195)	3.95	.41		
Other (17)	3.81	.34	Δ = +.14	.38
Mean impact score	3.93	.40		
by size of grant (206)[b]	$758,664	$586,355	r = .24	.51
Mean impact score	3.94	.39		
by age of grant (210)[b]	46.27	20.69	r = -.12	-.24

[a] Size of group in parentheses. [b] Size minus number of outliers.

I computed the difference (Δ) between the means by subtracting the second value from the first. For example, the difference between the scores for centers was .17 scale points higher than for projects. These differences are expressed as effect sizes in column four. The correlations are reported as the relationship between the first and second value. A negative result means there is an inverse relationship. As before, I removed outliers prior to conducting the correlational studies. Six were excluded for the correlation between the mean impact scores and size of grant and two were excluded for the age of grant correlation.

Again, I found the mean impact scale was able to detect differences between groups and was correlated with the background characteristics that were available. In general, the magnitude of the mean impact score differences and relationships were lower than were obtained using the total impact scale. For example, the mean effect size between centers and projects was .78 for the total scale and .42 for the mean impact scale. This was not unexpected because the NA responses were not coded as zeros. Essentially, it leveled the playing ground because each grantee was scored only for those activities it actually did.

However, it is important to point out that centers still perceived themselves as to have been more impacted than were projects. To set this in the previously used percentile rank terms, the center average score now falls at the 66th percentile of the project mean impact scores rather than at the 79th for the total impact scores.

I found a somewhat surprising result for the comparison of the active grant scores compared to those of the excluded group. The effect size was greater using the mean scores, .44 compared to .25. That is a change from a small to a moderate difference. I suspect this may be related to the age variable as I pointed out in the age/total score analysis discussed above. I will return to this discussion later in this report.

I show a graphic representation of the differences in the mean score as a function of the institution type that received the grant, two-year colleges, four-year, and other. See Figure 4.

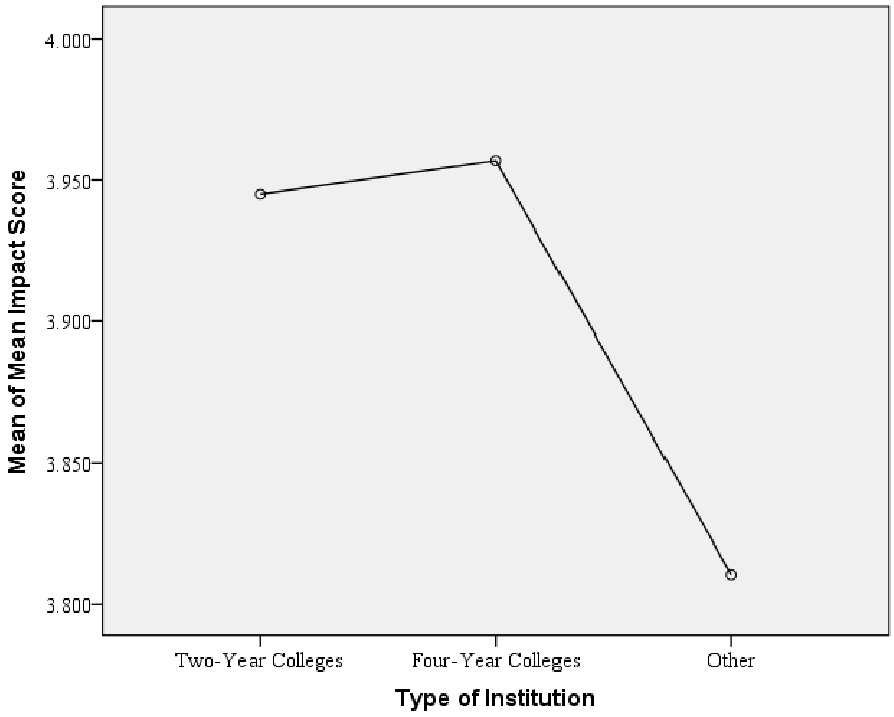

Figure 4. Plot of the average mean impact scores by type of institution.

As discussed above, it is difficult to interpret effect size for three or more groups. When this plot is compared to Figure 4, we see that the two year/four year college difference practically disappears. In addition, the effect size for the differences between colleges and other drops from .89 to .38. This is a substantial difference and suggests that analyses involving the impact of grants to other type of organizations and institutions would need to choose carefully the way they want to score the impact survey.

The correlation between the mean impact scores and the size of grant were similar for the two scoring procedures, r = .30 for total .24 for the mean scores. However, as noted above, the size variable is strongly correlated with the center/project difference so I computed the relative contribution of each by conducting a partial correlation analysis. Using the sample size of 206 respondents, I calculated the correlation between the mean impact scores and size controlling for the center/project difference. I obtained a partial correlation coefficient of .17. When I computed the partial correlation between the mean impact scores and the center/project difference controlling for size, I obtained a pr of .03. This means that the size variable accounts

for most of variance in the mean impact scales. This was a different result than I obtained using the total scores. In that instance, size and the project/center contributions were similar.

The correlation between the mean scores and the age of a grant was a -.12. This negative sign means that the older the grant, the lower the score. This was similar to the finding using the total impact score. However, again the age of grant is confounded by the active/expired relationship. Active grants score higher but they are on average, younger than the expired. I used a partial correlation analysis to compare the mean scores with age controlling for the active/expired variable.

I obtained a partial r of .06 when controlling for active/expired where active was coded as one and expired was coded as two. The pr obtained between the mean scores and the active/expired difference when controlling for age was -.18. Because expired was coded as a higher number, this means that the expired grants had lower scores when controlling for age. The active/expired difference was accounting for the majority of the variance, 3.2% to 0.4%. These results are similar to those obtained when using the total impact scores.

As was the case with the total scale scores, I was able to detect differences between groups using the mean scale scores. This provides additional support for the usefulness of the peer-generated process for developing surveys. I think it could be used to measure the impact or the sustainability of any NSF support program (Welch, 2012). I believe it also has value as a way of evaluating the ATE and other NSF programs as well as individual grants.

Total impact scores compared to mean impact scores. The two scores are correlated (r = .58) but they resulted in somewhat different findings. One can see the effect of the two different approaches by examining an illustrative case. Grant X had a large number of NA responses coded as zeros. Their score on the total scale was 67, a 5th percentile score. However, their score on the mean scale was 3.94, a 50th percentile score. They answered fewer items, but when those were excluded, their score was much higher relative to the other scores.

I used an analogy from baseball to help understand the two approaches. Consider a situation where a manager wants to compare the hitting ability of two baseball players. Josh hit 27 home runs (HR) in the 80 games he played during the season and Justin hit 35 home runs during the 120 games he played. Each missed games during the season because of injuries and other issues. Which player showed the greater home run hitting proficiency?

If we use total home runs as our measure, the winner would be Justin because he out hit Josh, 35 to 27. However, if we use a measure based on the average number of home runs scored during the games played, the leader would be Josh with his average of .34 HR per game. Justin's average was .29 HR per game. Which measure should be used? I think an argument can be made that both measures are indicators of home run hitting ability and the same is true for the total and means scores. However, a strong argument can be made that Justin made the more valuable contribution to his team in that he did hit 35 home runs, 8 more that Josh did. When Josh did not play, he was not contributing to the team and its success or failure. His failure to play for whatever reasons meant that he was not helping the team. I think the same argument can be made for the Total Impact scores; it is an indicator of the total scope of impact. In addition,

those scores were more sensitive to group differences. The choice depends, in part, on the purpose to be served by measuring ATE impact.

For example, I am developing an ATE evaluation instrument using select items from the survey. A positive response means that the grant has been successful in doing things that help to achieve the project and program goals. Consider the goal of ATE as described in the Congressional act that established the program, "to develop and disseminate model instructional materials (U. S. Senate, 1992). A statement on the survey directly addresses that goal. It is "The grant has permitted us to develop educational materials that otherwise would not be available." If a grantee agrees with the statement, that is an evaluative indication that the program is meeting its goals.

Now the issue is what one does about a NA response. Should we judge a project that has developed materials to have more value than one that has not done materials development? Which is better, a project that promised to do 10 activities and does them well or one that promised to do 20 activities and also does them well? It is similar to the home run hitting analogy. It will depend on the purpose of the analysis.

Concluding Remarks

A major purpose of this research was to determine if a Likert-type survey based on peer-generated statements met normal psychometric standards. The answer is yes. Both scales met generally accepted criteria for effective social science research measuring instruments. Because of this, it was argued that the peer-generated procedure could be adapted for use by NSF and other federally funded programs to measure such things as impact, sustainability, and evaluation. In addition, I believe PIs or other grant leaders could use this approach to investigate their individual grants.

Another contribution of this report was the analysis of how to score a Likert-type scale that included a "Not Applicable" option. Little was found in the literature on the topic but it was clear from the study they measure somewhat different things. Several examples of the differences were presented. For example, if one wanted to measure the impact of grants that went to institutions other than colleges, it would be more appropriate to use the mean impact scale because of the focused nature of these grants. Similarly, if one was interested in determining what grants best addressed a broad scope of activities, the total impact score would be more appropriate.

Because I was working with a population and not a random sample, I presented all findings using descriptive statistics. Effect sizes were used to show the magnitude of differences or the strengths of correlations.

Several grantee characteristics were identified that helped explain what is entailed by the construct, impact. For example, I found that impact was related to the size of a grant for both scales while age was not. Perhaps a theory of dissemination and utilization can be found, for example (Rogers, 1995) or (Fullan, 1991), to determine if the elements of their model is consistent with the factors related to impact as presented in this research. For example, what do

the theories predict about the age of a project and is it consistent with the results found in the current study.

A limitation of the study was the inability to generalize to all ATE grants. We know the results for the 261 awardees included in the population but there have been more than 1,300 grants made since the program's inception. This situation leads to a suggestion for further research. Such research would address the question, "Are similar results obtained if the study was replicated using a random sample of grants?" In addition, some of the statements used in this study were used on the annual survey of ATE grants conducted by Western Michigan University. These provide an opportunity for another kind of replication that may enhance the generalizability of the results.

Many of the impact items are evaluative in nature. For example, here are four items that loaded highest on the impact scale. The italicized words indicate what or who was changed, e.g., faculty, and what was the nature of the change, e.g., improved teaching style.

"*Student interest* in technology careers has *increased* because of our ATE grant."

"Our *faculty* has *improved* their teaching style because of their involvement in our ATE grant."

"*Businesses* in our area *have benefited* from having a more qualified pool of job candidates from which to choose."

"Our ATE grant has helped us *produce more* science and engineering *technicians* than we would have done without the grant."

These items are all evaluative in nature as are a majority of the other statements on the impact survey. It should be possible to select the evaluative items and use them to create an ATE evaluation survey that would be useful for NSF and for grant leaders.

References

Introduction to Regression. (2011, July). Retrieved from Princeton University, Data and statistical services: http://dss.princeton.edu/online_help/analysis/regression_intro.htm

Exploratory Factor Analysis. (2012). Retrieved from Wikiuniversity: http://en.wikiversity.org/wiki/Exploratory_factor_analysis

Multicollinearity. (2012, December 10). Retrieved from Wikipedia, the free dictionary: http://en.wikipedia.org/wiki/Multicollinearity

Wikiversity. (2012, December 12). Retrieved from Eta-squared: http://en.wikiversity.org/wiki/Eta-squared

Berk, R. A., & Freedman, D. A. (2012). *Statistical assumptions as empirical commitments.* Retrieved from stat.berkley.edu: http://www.stat.berkeley.edu/~census/berk2.pdf

Borg, W. R., & Gall, M. D. (1983). *Educational Research: An Introduction.* New York and London: Longman.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd. ed.).* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Fullan, M. &. (1991). *The new meaning of educational change. 2nd ed.* New York: Teachers College Press.

George, D., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference. 11.0 update (4th ed.).* Boston: Allyn and Bacon.

Gullickson, A., & Welch, W. W. (2006). *The sustainability of advanced technological education supported efforts; An evaluation.* Retrieved 2012, from Evalu-ate Resource Library: http://evaluation.wmich.edu/evalctr/ate/ATESustainabilityReport.pdf

Howell, D. C. (2009, March 9). *Treatment of missing data*. Retrieved 2012, from www.uvm.edu/~dhowell/StatPages: http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

Howell, D. C. (2011). *Fundamental Statistics for the Behavioral Sciences.* Belmont, CA: Wadsworth.

http://dss.princeton.edu/online_help/analysis/regression_intro.htm. (n.d.). Retrieved from Princeton University, Data and statistical services.

IBM SPSS Statistics 20. (2011). Using reliability measures to analyze inter-rater agreement. *SPSS Statistics Help*. Armonk, New York 10504-1722, New York, U.S.: IBM Corporation.

Internet Scout. (2012, March). *ATE Outreach Kit.* Retrieved from ATE Central Advanced Technological Education: http://atecentral.net/index.php?P=OutreachKit

Mallery, G. M., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference. 11.0 update (4th ed.).* Boston: Allyn and Bacon.

National Science Foundation. (2002). *Advanced Technological Education (ATE)*. Retrieved 2012, from Program Solicitation NSF-02-035: www.nsf.gov/pubs/2002/nsf02035/nsf02035.html

National Science Foundation. (2011). *Advanced Technological Education (ATE)*. Retrieved from Program Soliciation NSF 11 - 692: www.nsf.gov/pubs/2011/nsf11692/nsf11692.htm

Norman, G. (Published online: February 10, 2010, February 10). Likert scales, levels of measurement and the ‘‘laws’’ of statistics. *Advances in Health Science Education*.

Nunnally, J. (1978). *Psychometric Theory (2nd ed.).* New York: McGraw-Hill Book Company.

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors Quality and Quantity 41(5)673-690. *Quality and Quantity*, 673-690.

Rogers, E. (1995). *Diffusion of Innovations. (4th ed.).* New York: The Free Press.

Tabachnick, B. G., & Fidel, L. S. (1989). *Using multivariate statistics. (2nd edition).* Glenville, IL: HarperCollins.

U. S. Senate. (1992). *Library of Congress Thomas.* Retrieved from S.1146 : Scientific and Advanced-Technology Act of 1992: http://thomas.loc.gov/cgi-bin/bdquery/z?d102:SN01146:|TOM:/bss/d102query.html|

Uebersax, J. (2006, August 31). *Likert scales: dispelling the confusion.* Retrieved November 06, 2011, from Statistical Methods for Rater Agreement website: http://john-uebersax.com/stat/likert.htm

Welch, W. W. (2011a, December). *A Study of the Sustainability of the Advanced Technological Education Program (Revised)*. Retrieved from Evalu-ATE: http://evalu-ate.org/resources/sustainability_of_ate/

Welch, W. W. (2011b). *Research Report 2: The Impact of the Advanced Technological Education Program*. Retrieved from Evalu-ATE: http://evalu-ate.org/resources/sustainability_of_ate/

Welch, W. W. (2012). *Measuring the sustainability of the advanced technological education (ATE) program.* Retrieved from Evalu-ATE: http://www.colorado.edu/ibs/decaproject/pubs/

Welch, W. W., & Barlau, A. N. (2011). *Addressing Survey Nonresponse in Science Education Research.* Minneapolis, MN: Rainbow Research, Inc.