Title Page:

Developing a Measure of the Impact of the National Science Foundation's Advanced Technological Education Program

Wayne W. Welch
University of Minnesota
July 2012

Author Note

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

The Advanced Technological Education (ATE) program, funded by the National Science Foundation (NSF), is designed to improve the education of technicians in high-technology fields such as biotechnology, advanced manufacturing, information technology, and environmental and energy technologies. The program makes grants to support projects and centers to achieve this goal. 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.

In recent years, the foundation added a “targeted research” program 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 principal investigators (PIs) of ATE projects and centers. The topic of interest in this NSF-supported research was to determine if we could develop and implement a procedure to assess the impact (effect or influence) on the people and institutions involved in the ATE program.

If such a procedure can be developed, NSF could use it to identify ways the program has influenced its grantees, help NSF evaluate the ATE program, and provide information to PIs on how ATE has affected them and their institutions. In addition, the information could be used for program improvement by reducing the negative impacts and enhancing those that are positive.

Method

The initial 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 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.”

I placed their statements in quotation marks and put them on a survey for other ATE PIs to decide if the statements described their own situation. I called this kind of survey a Peer-Generated Likert Scale because 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.

This process yielded 95 statements about impact. These statements were mapped against the working framework to determine how well it 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 few examples of the statements follow. A detailed description of this process is available in Welch (2011a).

- "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 clear instructions, minimal response burden, ease of scoring and the like. I analyzed the survey for readability and obtained a Flesch-Kincaid grade level of 11.5, that is, readable for a high school junior.

I mailed the survey to 261 current and past ATE principal investigators (PIs).[1] 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 that the larger centers were somewhat more likely to respond to the survey. However, no differences were found between respondents and nonrespondents on a scale that measures sustainability (Welch & Barlau, 2011).

Impact Scale Analysis

The purpose of the current study is to determine if the responses to a set of 29 Likert-style items could be used to create a useful, reliable, and valid scale that measures the impact of involvement in an ATE grant. 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).

[1] The research population consisted of active ATE grantees that began prior to Jan 1, 2009 and grantees that had expired between Jan. 1, 2007 and Dec. 31, 2009.

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 ran from zero for a "Not Applicable" response, to five for a "Strongly Affirm" response.[2] The total score is the sum of the responses to the 29 impact items. The higher the score, the greater the impact.

I also computed an impact score based on the average response to the impact items. I called it the Mean Item Impact score. Statements marked "Not Applicable" 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 W. W. 2012).

After creating the two impact scale scores, I computed scatterplots and descriptive statistics, checked for outliers, and plotted their distributions checking for equality of variance. In addition, I computed the reliability of the measures and provided other evidence that supports the validity of the scales.

Total Impact Scores. The distribution is shown in Figure 1.

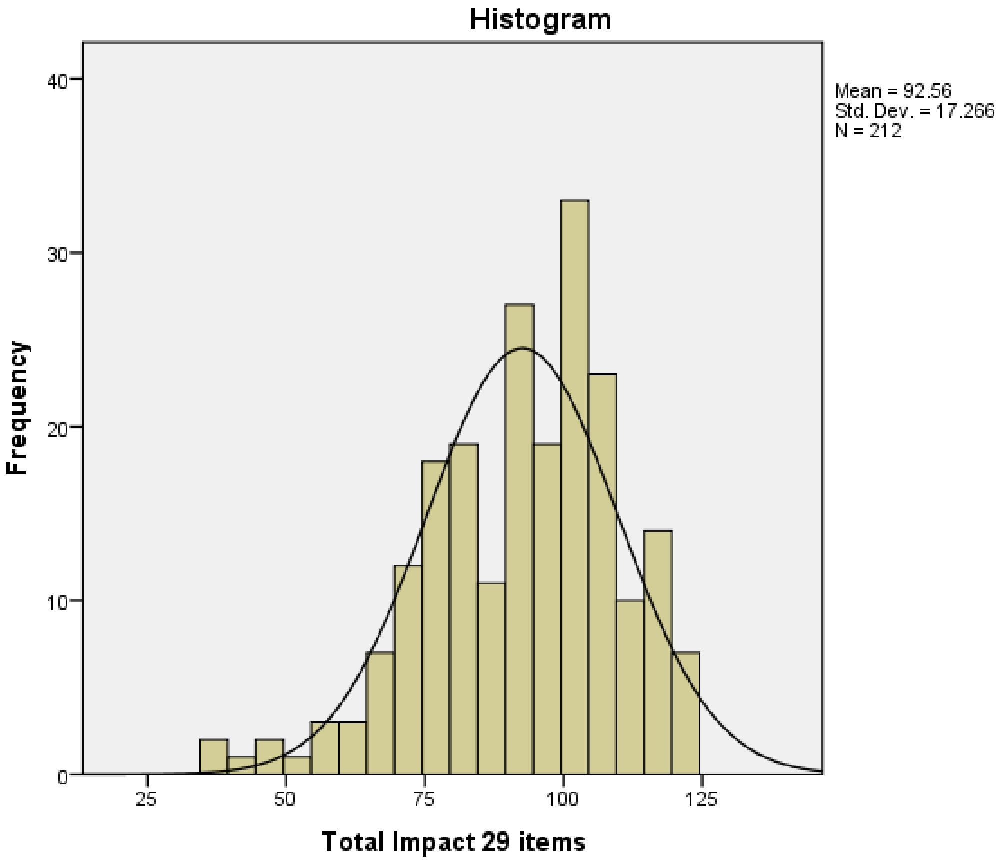


Figure 1. Distribution of Total Impact Scores

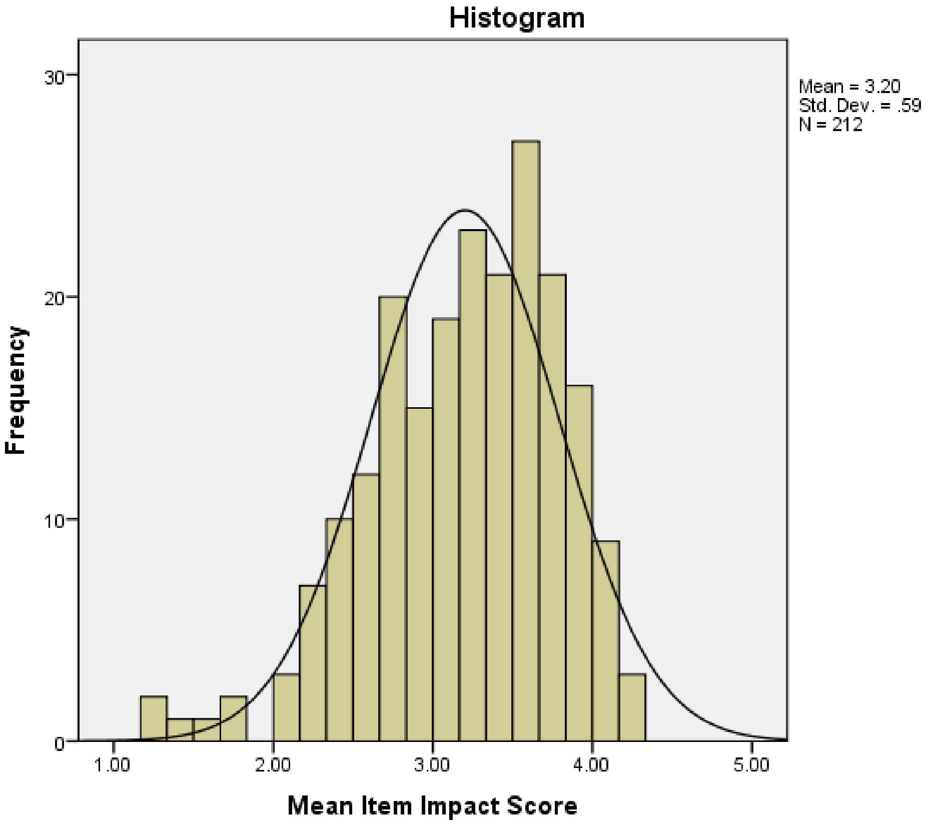
[2] Affirmation means that the subjects agreed or strongly agreed with positively stated items and disagreed with negatively stated ones.

The mean score was 92.56 and the standard deviation was 17.27. The skewness was -.66 indicating the distribution was somewhat skewed to the left. The kurtosis, for flatness of scores, was .37. Both were in acceptable ranges (+/- 2.0) to assume a normal distribution. The scores ranged from 37 to 124.

A second requirement for an effective measure is that the scores meet acceptable levels of reliability. I computed a reliability coefficient (Cronbach's alpha) that is a measure of 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 .84, which meets generally accepted standards. For example, (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 exceeds these standards.

Mean Item 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" response was defined as a missing value, which meant that the item was not included in the averaging process. I plotted the distribution of the Mean Item Impact scores and overlaid a normal curve. The results are shown in Figure 2.

Figure 2. Distribution of Mean Item Impact Scores



The mean score was 3.20 on the 5-point scale and the standard deviation was .59. The skewness was -.68 and the kurtosis was .46, within the standards of +/- 2 required for a normal distribution. The range of scores was 1.23 to 4.28.

I used a pairwise deletion process to calculate Cronbach's alpha reliability. This is a measure of the internal consistency of the items, that is, how well are the items measuring the same thing? Because of missing values and the use of the NA response, the number of cases used to measure the inter-item correlations varied from one pair to the next. One way to handle this is to compute and average inter-item correlation. This is an indication of reliability for a single correlation. The Spearman-Brown formula can be used to estimate the reliability of a 29-item survey. This process is similar to the intra-class correlation process found in the IBM SPSS statistics software program.

A 29-item array produces 406 different item correlations. The average of those values was .143. Applying the Spearman-Brown prophecy formula yielded an alpha value of .83. This meets the standards for acceptable measuring instruments.

Group Comparisons

A valid measure should be able to detect differences between groups if such differences exist. For example, one could assume that music majors would score higher on a music appreciation survey than would college students in general. If you give the survey to both groups, and it is a valid survey, you would expect the music students to have higher scores. If this does occur, then one has provided evidence of the validity of the survey. If this is not the case, then one must question the underlying assumption about music attitudes or decide the instrument was not valid.

I applied this reasoning to the current study by predicting that the larger and longer lasting centers would have a greater impact than the smaller and more focused projects. I found that centers did have higher scores on the Total Impact scale. The mean for centers was 103.16 while the mean for projects was 89.71. This difference is shown in Table 1 along with comparisons on four other variables. These variables are described below.

- Grant Status: Active or expired. Whether the grant was active or had ended.
- Program Track: Projects or centers
- Grantee Institution: Two-year or four-year colleges
- Size of Grant: Average amount awarded in dollars
- Age of Grant: Average number of months between initial award and survey date

Total Impact Scores. I compared the group means for the first three traits using independent t tests. I used Pearson's correlation to determine the relationship between the survey scores and size and age of grant. The findings are shown in Table 1.

Table 1
Total Impact Score by Group Background Characteristics

Group [a]	Mean	Mean difference	*t* and *r* values	*p* value	Effect size
Active grants (131)	94.13				
Expired grants (81)	90.02	4.11	1.69	.09	.24
Centers (45)	103.16				
Projects (167)	89.71	13.45	6.25	.00	.92
2-Yr colleges (156)	94.30				
4-Yr colleges (39)	90.97	3.33	1.13	.26	.19
Total impact score by Size of grant	___	___	r = .25	.00	.52
Total impact score by Age of grant		___	r = -.04	.58	.01

[a] Size of group in parentheses

The last column is the effect size (ES). It is the name given to a number of indices used to express the magnitude of the score differences. For the comparison of means, I used Cohen's d. It is the mean difference divided by the pooled standard deviation. It is a standardized estimate of the size of the differences between groups (Cohen, 1988). He defined effect sizes as "small," d = .2, "medium," d = .5, and "large," d = .8 or above.

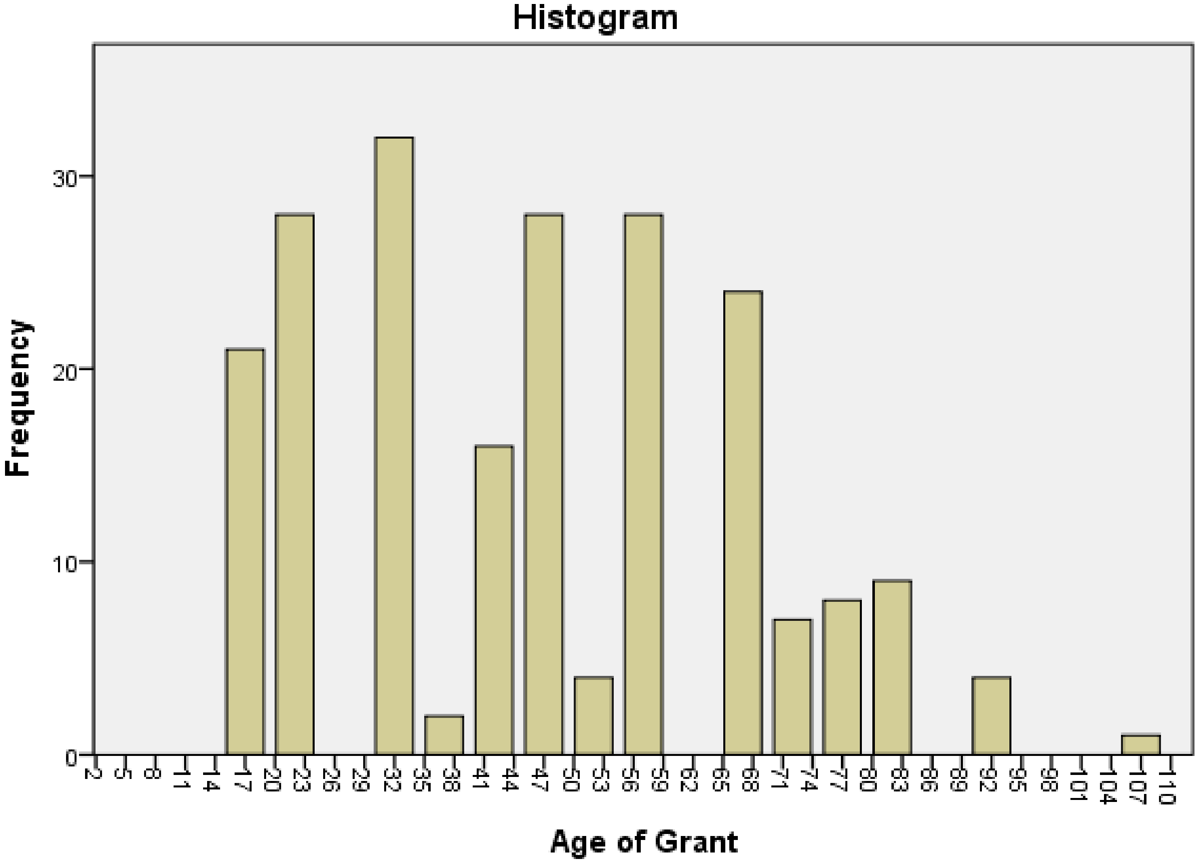
The effect size for correlational studies is the value of r, the correlation coefficient. Cohen's d and r are related by the formula $d = r / \sqrt{} (1 – r^2)$ .25. For example, an r of .37 is equivalent to a d of .80.

One can see the findings support the predicted differences between centers and projects. In addition, the impact scale was able to discriminate between groups on other variables. The capability of the instrument to detect group differences provides validity evidence for the survey.

Contradictory results were obtained for the active/expired difference and the impact of age of a grant. Even though expired grants had significantly higher mean ages, 65.9 months for expired versus 34.1 months for active, the correlation between age of grant and impact scores was only -.04, that is, essentially zero. I decided to examine the age variable in more detail to determine if there was something in the distribution accounting for the low correlation value.

There are several things, in addition to the actual relationship between two variables, which can affect the size of a correlation. These include a restricted range of one or both variables, the sample may be a combination of dissimilar samples, or there may be outliers in one or both variables. To test for these, I plotted the frequency distribution for the age of the grant. This is shown in Figure 3.

Figure 3. 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 table with the vertical columns at about 20 months, 32, 45, 57, 68, and 80 months. These are about one year apart reflecting the normal funding process.

I computed the duration by computing the number of months between the initial funding date and the time the survey was returned. This created about six groups of grants funded at the same time. I examined these six groups and found the distribution to be similar for each group. I believe the sample comprises similar (homogeneous) groups.

There was one outlier but when I removed it and re-calculated the correlation, it changed very little. Finally, the distribution does not indicate any restriction in range. Hence, I conclude that there is little relationship between my measure of age of grant and Total Impact score.

To further test the validity of the age of grant measure, I correlated it with whether the grant was active or had ended (expired). Expired grants are normally older than active ones. The Pearson correlation between the two variables was .75, a very high correlation (r [212] = .75, p = .00). The means for the two groups were Active = 34.12, Expired = 65.93. The difference is in the expected direction. Expired grants are older than active ones. This provides addition evidence that the scale is valid.

Mean Impact Score. I repeated the above process using the Mean Impact score as the dependent variable. This is a measure of the average response to a 5-item scale that runs from “Strongly Affirm” coded 5, to “Strongly Deny” coded as 1. It is a relative measure of impact rather than a summated measure. I computed the relationship between the background characteristics I obtained from NSF’s FastLane site. The results of these comparisons are presented in Table 2.

Table 2

Mean Item Impact Score by Group Background Characteristics

Group [a]	Mean	Mean difference	*t* and *r* values	*p* value	Effect size
Active grants (131)	3.66				
Expired grants (81)	3.51	.15	3.33	.001	.48
Centers (45)	3.76				
Projects (167)	3.56	.20	3.80	.000	.63
2-Yr colleges (156)	3.59				
4-Yr colleges (39)	3.63	-.04	-.73	.47	.12
Mean impact score by size of grant	___	___	r = .29	.00	.61
Mean impact score by age of grant	___	___	r = -.11	.12	.24

[a] Size of group in parentheses

Here, again, we see that the Mean Impact scale distinguishes between groups. The pattern of the findings is similar to those found when using the Total Impact score as the dependent measure. These results provide evidence of the validity of the mean impact scores.

A summary of the findings for the two scales is shown in Table 3.

Table 3
Comparison Effect Sizes for Total Impact and Mean Item Impact Scores

Variables	Effect Sizes	
	Total Impact Score	Mean Item Impact Score
Active vs. Expired	.24	.48
Centers vs. Projects	.92	.63
Two year vs. Four year	.19	.12
By Size of Grant	.52	.61
By Age of Grant	.01	.24

We see that the impact of an ATE program was related to whether the grant was for a center or a project and to the size of the grant. Both have large effect sizes. This is consistent with what one would expect. A greater impact would be expected for a $3M grant than one might expect for a grant of $750,000.

These two variables are related which confounds the analysis. That is, one cannot determine whether the scale results are due to grant size or to project/center differences. Centers generally are funded at larger amounts. A multiple regression analysis suggested the major factor is the size of the grant rather than the center/project difference. However, additional analyses will be carried out and the results presented in a subsequent report.

Another variable that is related to impact but at a more moderate level is the active versus expired comparison. Active grantees report a larger impact than do expired ones. There was a small effect of the age of grant using the mean item score as the dependent variable. Older grants had lower impact scores. Here, again, there is a confounding situation because age of grant is correlated with the active/expired difference.

A multiple regression analysis of the relative contribution of age and active/expired difference suggests it is the active versus expired comparison that is most related to the survey scores. This, too, will be investigated in a future report.

Both measures of impact showed differences among some of the groups. I revisited the issue of whether to use both or just one of them. I used an analogy to explain how a total score compares to an item mean score. Consider a situation where a manager wants to compare the batting 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. They measure somewhat different things but both are indicators of impact and both could be used depending on the purpose of the analysis.

The two measures are related (r [212] = .53, p = .00). This is considered a high correlation coefficient but it also mean the two scores are not measuring exactly the same thing.

Impact Reconsidered

In the preceding analysis, evidence was provided that we have a viable measure of the impact of an ATE experience. It may be informative to examine the construct of "impact" to see what is being measured. In the introduction, synonyms for impact were effect and influence. Words such as outcome and consequence also could be used. However, I think a useful operational definition is that an impact is a change in the system or a part of the system. As hypothesized during survey development, the change might occur for people, programs, or organizations.

We can illustrate the kind of change being discussed by examining the following sample items for each of the three categories. What was changed, and how it was changed, are shown by the italicized words.

Change in People: (Faculty, Students, Administrators, ATE PIs/Staff)

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

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

Program Changes: (Curriculum, Instruction, Educational Materials)

The grant has permitted us to *develop* educational *materials* that otherwise would not be available.

We were ~~not~~ able[3] to *develop* all the curriculum *materials* that we had planned to do."

Institutional Change: (Colleges, Schools, Business/Industry, Communities

The ATE grant helped us to *establish relationships* with professionals from four-year colleges that will continue in the future.

Our NSF/ATE grant has *had* ~~little~~ a long-term *impact* on our college.

[3] I changed the direction of the statements so they all represent positive impacts.

Phrases such as “faculty improved,” “student interest increased,” and “relationships established,” describe the nature of ATE impact. Summing across these statements yields an indicator of the extent of the change, that is, the impact of the grant.

Concluding Remarks

The purpose of this study was to determine if we could develop and implement a procedure to assess the impact (effect or influence) of Advanced Technological Education (ATE) grants. I followed recommended procedures for scale or test development to create a Peer-Generated Likert scale to measure impact. I defined the domain of content using statements made by ATE PIs and other stakeholders, selected my sample, and mailed out the survey. After three follow-up contacts were made, I obtained 212 responses, an 81% response rate.

I computed a Total and a Mean Impact score and examined the distribution of the scores. The measurement properties met acceptable standards for further analysis. I computed scale reliabilities and provided validity evidence of the scales.

The impact scores were related to size of grant, and whether it was a center or project. Centers and larger grants reported higher impact scores. These findings are confounded because centers are funded at higher levels than projects.

I also found that scores on the mean impact scale were related to age and to whether the grant was active or expired. Active grants had higher impact scores and there was a small positive correlation with age. Here, again, the two variables are related in that expired grants are generally older grants. Future analysis will use regression analysis to understand better this relationship. In addition, a factor analysis of the scale will be carried out to determine if there is an underlying structure for the concept, impact.

References

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

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.

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.

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