

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to the *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited.

Volume 12, Number 19, December 2007

ISSN 1531-7714

A data visualization and data mining approach to response and non-response analysis in survey research

Chong Ho Yu, Angel Jannasch-Pennell, Samuel DiGangi, Chang Kim, & Sandra Andrews
Arizona State University

Survey data based on self-selected samples are inherently subject to the threat of selection bias. In this study, both data visualization and data mining techniques were employed to examine whether nonresponse bias had affected a survey regarding 1:1 computing conducted at Arizona State University. Unlike conventional hypothesis testing, data visualization/EDA attends to pattern recognition instead of probabilistic inferences. In addition, unlike logistic regression, classification trees in data mining are capable of ranking independent variables in terms of their predictive power. In contrast to the findings of other studies, this study reveals that academic level, gender, and race were not identified as crucial factors in determining the response rate. Rather, the nature of the subject matter might be more important for science/engineering and law students seemed more interested in this technology-related survey.

The objective of this article is to illustrate how data visualization and data mining can facilitate analysis of response and non-response rates with survey data. Survey data based on self-selected samples are inherently subject to the threat of selection bias, thus conventional parametric procedures might not be properly applied (Cassell, 2006). While various methods have been developed to compensate for non-response errors in survey research (e.g. Kott, 2006; Haziza & Rao, 2006), these sophisticated methods might be inaccessible to most educational and psychological researchers. Although many researchers are familiar with stepwise logistic regression, and at first glance it seems to be instrumental in constructing a predictive model for response rates, its problems and limitations are well-recognized (Greenland, 1989, Glymour, 2001; Thomsson, 1995). In this study, data visualization and data mining tools are considered more viable for their conceptual clarity, procedural simplicity, and software availability, and are powerful enough to unveil insight that could have been missed by conventional analytical tools.

In this article, a campus wide survey implemented at Arizona State University (ASU) concerning one-to-one (1:1) computing is used as an example to show that demographic variables (e.g. gender), which are believed to be crucial factors to response rate (Porter & Umbach, 2004), did not play a significant role in affecting the response pattern at ASU. Although it is not our intention to use a particular population to make a broad generalization across all other US institutions, it is our firm belief that other institutions can benefit from the experience of ASU, which is the fifth largest university in the country, in the fifth largest city in the US, and is composed of a very diverse student body.

Technology is the center of ASU's vision for building a learning environment that can break the temporal and spatial barriers, and 1:1 computing is part of the technology strategy for the future (Penuel, 2006). In a 1:1 computing environment, each student will have a mobile computer with wireless access to ASU resources and the Internet,

so that they will have the opportunity to be engaged in relevant learning activities.

Based upon the findings of a preliminary survey conducted in Spring 2006, a pilot program at ASU's downtown Phoenix campus was launched in an attempt to enable and support 1:1 computing specific to the needs of downtown students (DiGangi et al, 2007). The downtown campus is equipped with wireless Internet access, with experienced, on-site support staff who can provide users with troubleshooting. Common computing resources are provided at the campus for those students who choose not to own a laptop. There is no doubt that 1:1 computing plays a major role in implementing ASU's vision of ubiquitous access to learning resources, however, the question is: How should technology serve this institution to achieve outlined goals—administratively, academically, in the research enterprise, and in all other aspects of the university?

With this question in mind, ASU's Applied Learning Technologies Institute (ALT[^]I) collected baseline data on the scope and scale of technology use by ASU students. In this survey, we focused on gathering information on use of technologies such as laptops, desktops, wireless access, iPods (and similar MP3 devices), PDAs, tablet PCs, etc. We also addressed the following questions: What tools do students and faculty use? Where and how do they use them? How frequently do they use them? We are particularly interested in the use of such technologies for academic and educational purposes.

THE PROBLEM OF NONRESPONSE AND COUNTERMEASURES

The lack of response in a survey is a strong concern for decision making in higher education institutions, and there have been numerous studies on methods to increase survey participation and to analyze the nonresponse bias. A high non-response rate may occur in mail surveys due to respondent resistance (Sosdian & Sharp, 1980). In addition, survey fatigue from multiple or consecutive surveys of students can also lead to a high nonresponse rate (Porter, Whitcomb, & Weitzer, 2004). The topic of the survey is also a major factor. Previous research indicates that sensitive topics, such as sex, may

cause nonresponse, refusal, and termination in surveys or interviews (Johnson & Delamater, 1976). Likewise, many researchers obtained similar conclusions with health surveys, which are related to use of alcohol, tobacco, and illicit drugs (Wechsler, Lee, & Rigotti, 2001; Kypri & Gallagher, 2003; Cunradi, Moore, Killoran, & Ames, 2005; Rigotti, Lee, & Wechsler, 2000).

With the advance of information technology and the increase in the number of Internet users, many researchers have conducted experiments to compare between mail and Web-based surveys in an attempt to see whether use of technology can reduce survey nonresponse rates. Several researchers have found that Web-based surveys can obtain higher response rates than their mail counterparts (Kaplowitz, *et al.*, 2004; Kwak & Radler, 2002). Further, in order to improve survey response rates, studies were conducted on the role of incentives in surveys. Several studies have shown that use of incentives consistently promotes response rates (Singer, Van Hoewyk, & Maher, 2000; Singer et al., 1999; Willimack, Schuman, Pennell, & Lepkowski, 1995). In a meta-analysis of mail-based surveys, Yammarino, Skinner & Childers (1991) found that the literature indicates that small incentives given to all participants are most effective, while short (four-page) surveys, cover letters and preliminary notifications also yield higher response rates. However, Porter and Whitcomb (2004) found that the impact of lottery incentives on a student population has minimal effects. A further reason for across the board incentives (rather than lottery incentives) is found in the work of Carini, Hayek, Kuh, Kennedy & Ouimet (2003). They found that differential incentives are widely believed to be in place by respondents, and although these may not affect response rates, the differential incentives are perceived as unfair.

Recent studies of students in higher education confirmed previous findings that high ability students (identified via SAT scores and GPA) were more likely to respond to survey (Porter & Whitcomb, 2005; Porter & Umbach, 2006). In addition, another study showed that students who choose to respond via the web provide more favorable responses for information technology related items than do students who respond via paper (Carini, Hayek, Kuh, Kennedy & Ouimet,

2003). Finally, a large body of research has shown that women are more likely to respond to surveys than men (Porter & Umbach, 2006; Porter & Whitcomb, 2005; Moore & Tarnai, 2002) while students of color are less likely to respond (Porter & Umbach, 2006; Singer, van Hoewyk & Maher, 2000; Singer, Groves & Corning, 1999). But when the medium (mail vs. Web-based survey) was taken into account, a different story concerning the gender and race effects emerged. Sax, Gilmartin and Bryant (2003) asked whether college students respond to Web-based surveys at higher or lower rates than they do to traditional surveys, and whether non-respondent characteristics differed. In this study, race was not significant, but women were more likely to respond than men when either a paper survey or a web survey alone was offered; however, when given a choice, men were more likely than women to choose the web survey. Explanations for these demographic trends, however, have not been thoroughly investigated (Porter & Umbach, 2006).

METHODOLOGY

In order to account for the potential bias resulting from nonresponses, the ASU research team conducted a response rate analysis to examine the profiles of students who responded and those who did not in order to detect whether there was a significant bias in the self-selected sample. To address this issue, it is a common practice to compute inferential statistics to detect whether a significant difference exists between respondents and non-respondents in terms of various student attributes (e.g. academic performance and demographic variables.) However, in this approach, the result is subject to the influence of the sample size. It is important to point out that the sample size for the survey response analysis was 9332 students. To compare the attributes of respondents and non-respondents, however, all 62095 students are included. When a sample size is this large, any trivial difference may lead to a seemingly significant result that is not actually significant. In addition, in spite of its popularity, stepwise logistic regression has certain insurmountable problems. For example, while reporting the odds ratio is a common practice to indicate the ratio between the desirable and undesirable events (e.g. response vs. nonresponse),

the Mantel-Haenszel odds ratio is considered valid only under the assumption that the underlying stratum-specific odds ratios are constant across the strata (Greenland, 1989). Stepwise regression as a tool for variable selection has also been under severe criticism. It was found that stepwise regression tends to yield conclusions that cannot be replicated because this model-building approach capitalizes on sampling error (Thompson, 1995). It is also a well-known fact that the results of stepwise regression are affected by the order of entry of the variables (Glymour, 2001).

As a remedy, linking and brushing in data visualization/exploratory data analysis (EDA) and classification tree in data mining are used here to examine the student profile and the response pattern. It is important to note that data visualization and EDA focus on pattern recognition, hence no probabilistic inferences are involved, whereas classification trees are employed to rank order the factors that affect the response rate.

In this study both JMP (SAS Institute, 2006) and Splus Insightful Miner (Insightful Inc., 2006) are employed to construct classification trees. Because different software modules use different algorithms, it is essential to explore the data from multiple angles, unveiling insights via triangulation. Specifically, the classification tree in JMP uses Entropy (Quinlan, 1993) as the tree-splitting criterion. In order to verify the results through triangulation, another classification tree approach in Splus Insightful Miner is used. Insightful Miner provides us with sophisticated cross validation features and also another splitting criterion, namely, Gini (Breiman et al., 1984). Splitting criteria are measures of node “impurity” that determines where to make a split. It is based on the estimated probabilities from the node proportions. Ideally, we would like to partition data in a way that each partition is pure, which means that in a partition data vectors, in which each element represents a variable, should come from a single, homogeneous class. However, it hardly happens in reality and thus some degree of “impurity” must be expected (Han & Kamber, 2006). Gini and Entropy are different measures of impurity. While Entropy, the default criterion in JMP, favors balanced or similar splits,

the Gini option provided by Splus tends to favor the largest split or branch of the tree.

To counter-balance the potential problem of generating an overly complex model and to verify model stability, our survey data set was partitioned into five subsets for cross-validation. Specifically, the data were randomly divided into five equal sized groups and a model was built on the data with one group left out. This approach resembles Jackknife in resampling methodology except that in this case the unit is the group instead of individual observations. The missing group is predicted and a prediction error is calculated. This procedure is repeated for each group in turn and the average over all five instances is used as the model error rate.

Further, in Splus the minimum complexity option is chosen so that at each run the least complex model is retained by the minimum complexity criterion. Fitness versus parsimony is pervasive in every type of modeling, but there is a strong rationale for favoring simplicity. To explain an observed phenomenon based upon the data at hand, the best mode is the one that reaches the highest degree of model-data fit for its ample explanatory power. However, the merit of predictive models, such as classification trees, is tied

to its accuracy on unseen data. For the same or similar accuracy, smaller numbers of nodes, which means that the tree is less complex, can work better with unseen data. Conversely, a tree uses splitting variables with large numbers of values, thus yielding more nodes that can result in a negative impact on unseen data (Rosella, 2007).

RESULTS

Since the survey is concerned with computing, it is a reasonable assumption that students with strong math and science backgrounds may be more likely to respond to the survey. If this is the case, this self-selected sample may produce very biased results. Indicators for student math and science backgrounds are therefore included in the analysis. In Figure 1a, students who responded to the survey are “brushed” in dark green and the corresponding observations in the histogram showing SAT-Quantitative scores are highlighted in dark green. In Figure 1b, students who did not respond to the survey are linked and brushed in both panels as dark green. Based on these linked and brushed panels, it does not appear that there is a relationship between survey responses and SAT-Q scores.

Figure 1a: Distributions of student response pattern and SAT Quantitative scores

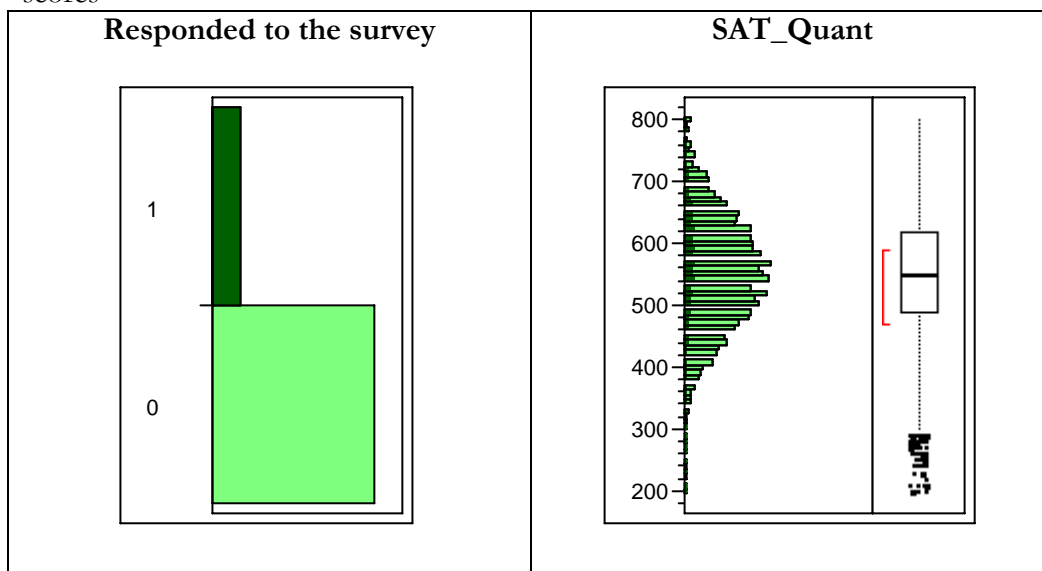
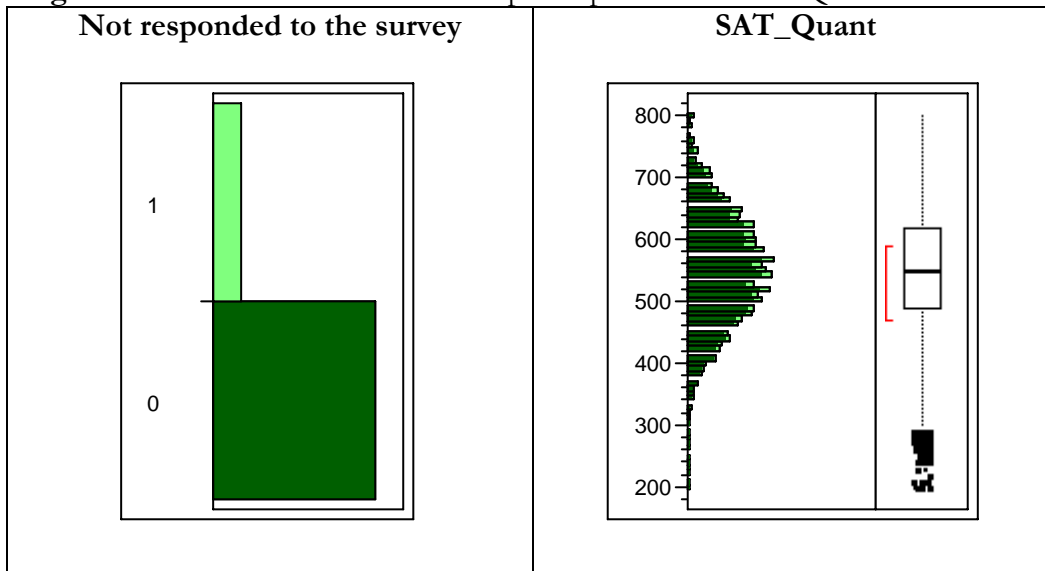


Figure 1b: Distributions of student response pattern and SAT Quantitative scores



Logistic regression, as mentioned before, is a common approach to regress a continuous-scaled predictor variable (e.g. SAT scores) against a dichotomous outcome variable (e.g. “responded” vs. “did not respond”). But, to demonstrate its weakness, a logistic regression model is run and its results indicate that SAT-Quantitative scores have a significant impact on the response rate (Odds ratio=1.003, $p < .0001$), At first glance, Figure 2 concurs with the logistic regression model because as the test scores go up, the number of respondents increase (observations highlighted in red). However, it is important to realize that at the two extremes (high and low scores) the number of observations is less dense than in the middle. When the students are partitioned into three groups by the SAT-Q score distribution (below the first quartile = low, above third quartile = high, between = middle), a different story emerges (Table 2). The response rate of students with high SAT-Q scores is 18% whereas that of students with low SAT-Q scores is 15%. In the middle score group the response rate is lower (12.3%) than both the high and low score groups. Looking at the table, a logistic function does not seem to fit the data. The question concerning whether higher SAT-Q leads to a higher response rate is inconclusive.

Figure 2: Logistic Fit

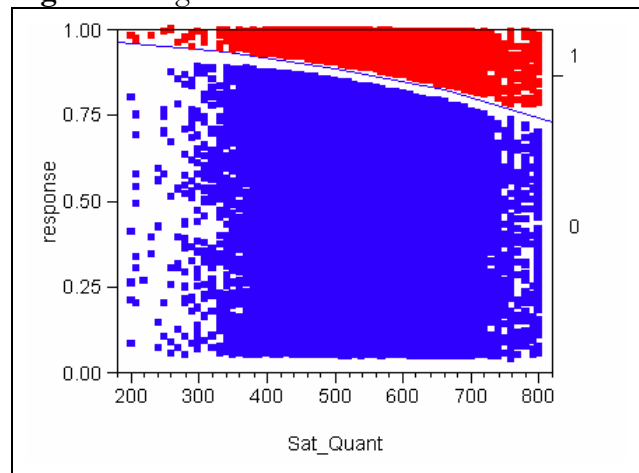


Table 2: SAT-Quant scores by responses

SAT-Quant	Responses		
Frequency	No	Yes	Total
(Row %)			
High	6,073 (81.35%)	1,393 (18.65%)	7,469
Middle	12,450 (87.70%)	1,746 (12.30%)	1,4196
Low	34,945 (84.95%)	6,193 (15.05%)	41,138
	53,471	9,332	62,803

There are other indicators of ASU student math and science backgrounds, such as ACT-Science Reasoning (ACT-SR), ACT-Natural Sciences (ACT-NS), and ACT-Math scores. Since these scores may be correlated, the data are presented in a scatterplot matrix, as shown in Figure 3. In the left panel of Figure 3a, respondents are highlighted in dark green and corresponding observations on the right panel

are bolded (bigger dots.) No cluster concentrated on high or low ACT scores is found. Figure 3b also does not have any discernible pattern. If SAT and ACT scores alone are taken into account, it seems that students with all levels of math and science knowledge are equally likely to participate in this technology-related survey.

Figure 3a: Responses to the survey and Scatterplot matrix of ACT scores

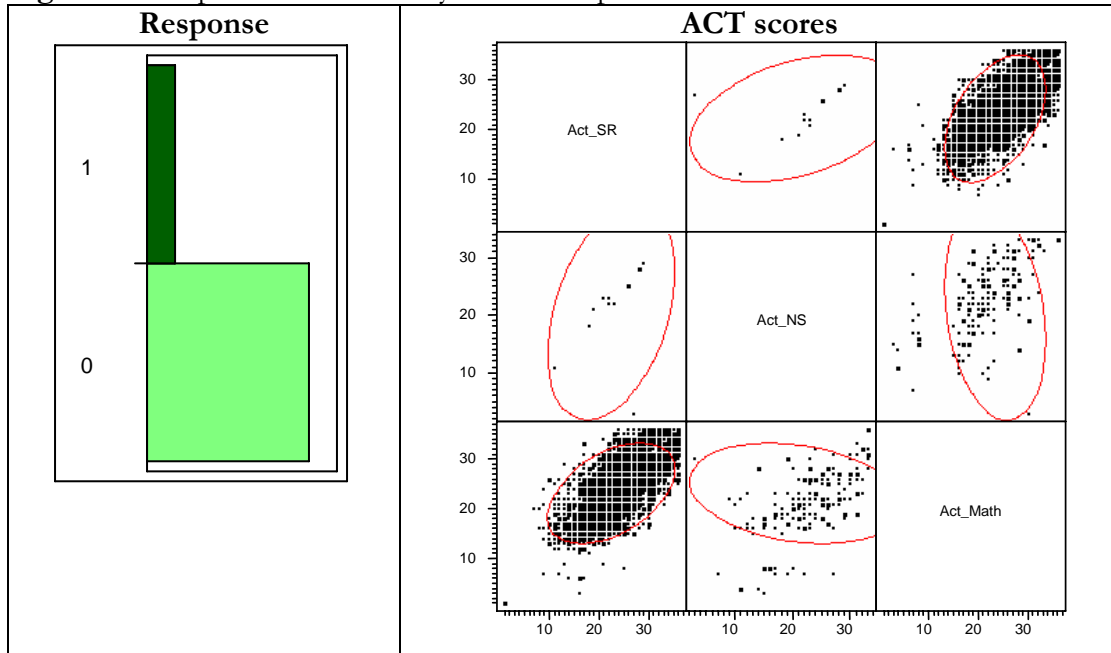
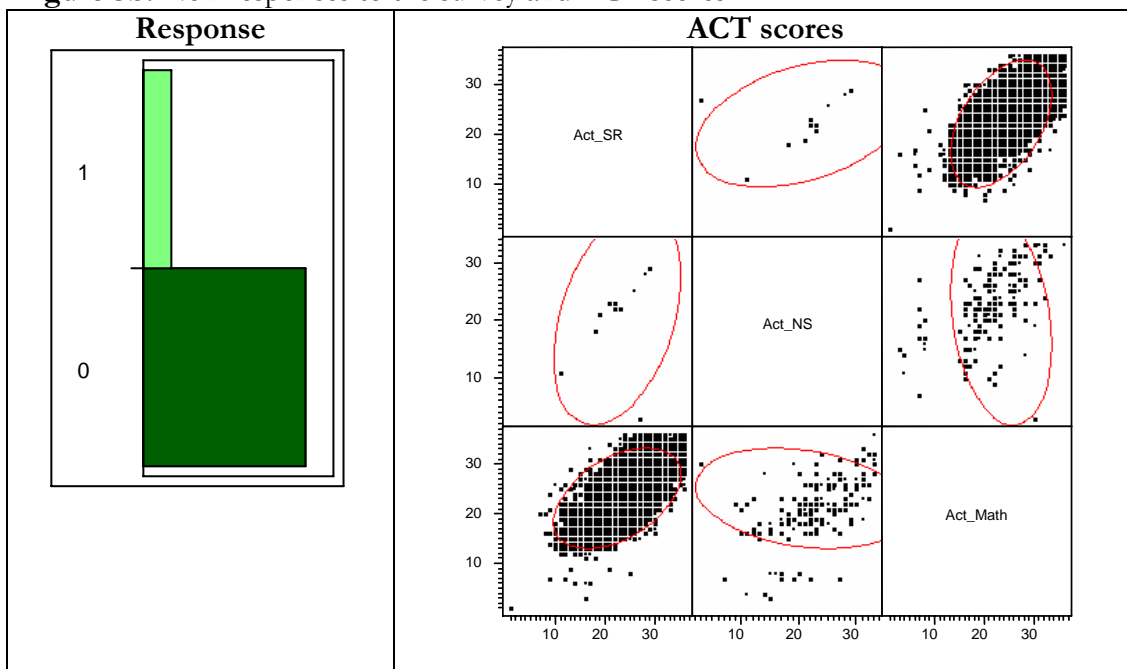


Figure 3b: Non-responses to the survey and ACT scores



However, when we look at the ASU Math Placement (ASU-MP) test scores, new information emerges. In Figure 4, it is clear that survey respondents are concentrated on the higher quartile of the ASU-MP score distribution. In other words, *the ASU internal math test scores show a conclusion opposite to that yielded by standardized test scores. It is suggested that students who are good at math were more likely to respond to this computing-related survey.* The same response pattern could be found among students who have higher GPAs. Figure 5 shows that respondents are concentrated on the highest quartile of the GPA distribution. To diagnose the preceding issue, one viable approach is to examine the relationship among SAT-Q, ASU-MP, and GPA. SAT-Q and ASU-MP seem

to be fairly correlated ($r = 0.59$), and students who responded to the survey scatter all over the graph with a slight concentration on the upper quadrant (Figure 6). The relationship between GPA and SAT-Q is weaker ($r = 0.36$), and again, there is a slight concentration of respondents on the upper quadrant (Figure 7). One of the plausible conjectures is timing. Standardized test scores are collected when the students were younger, and thus the ASU internal indicators may be more temporally relevant. However, without qualitative data (interviews and focus groups), numbers and figures alone cannot provide a sufficient explanation of why the use of standardized test scores and ASU's internal academic performance indicators lead to different conclusions.

Figure 4: Responses to the survey and ASU math placement scores

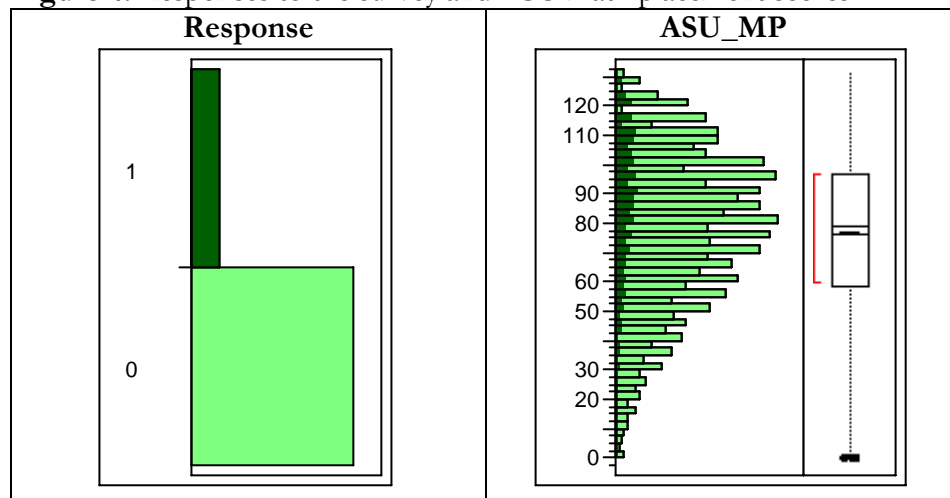


Figure 5: Responses to the survey and ASU's GPA

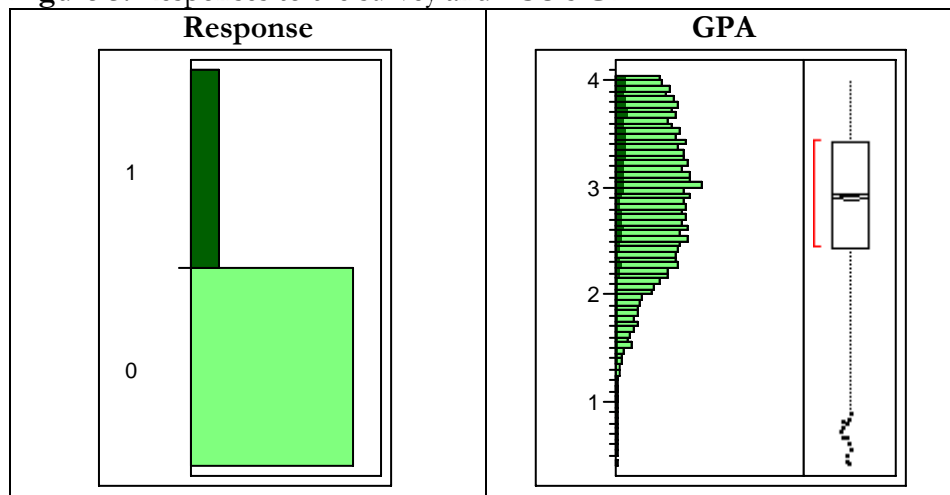


Figure 6: Bivariate Fit of ASU_MP By SAT-Q

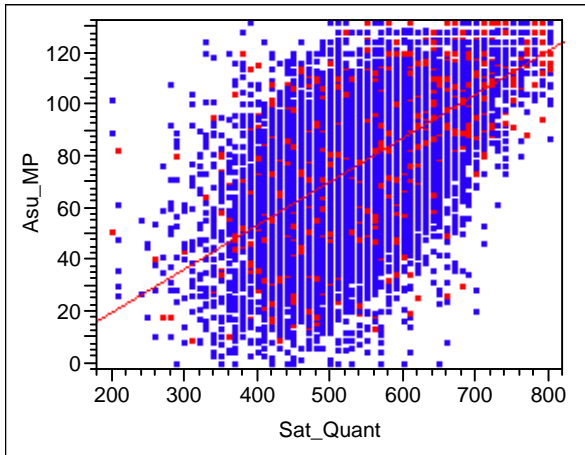
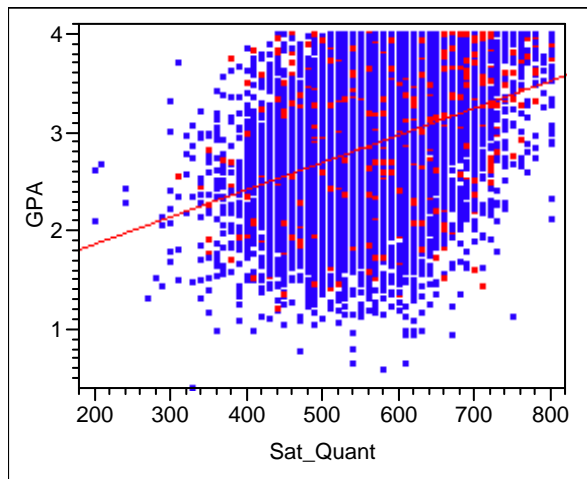


Figure 7: Bivariate Fit of GPA By SAT-Q



To examine the relationship between the response pattern and the categorical variables concerning student demographic information (college, race, gender, age group, and academic level) and the two indicators of academic performance (GPA and SAT-combined), the data mining technique referred to as a classification tree is implemented using JMP (Figure 8). It is important to note that, like data visualization in EDA, data mining also attends to pattern recognition rather than probabilistic inferences. During the process of classification, all variables are simultaneously evaluated, and the most crucial factor will be used to split the data set. Afterward, the second crucial factor will further partition the data in a hierarchical fashion. In this analysis, the variable “college” is identified as the most crucial

factor. When the splitting is saturated, and the model is overfitted, the tree is “pruned.”

It was found that *students who belong to the College of Science and Technology, College of Engineering, and College of Law were more likely to respond to the survey* (see Table 3 for the college code description). This confirms the finding based on ASU-MP test scores that science and engineering students are more interested in this technology-related survey.

Within the college group less likely to take the survey, GPA does matter. Higher GPA students were more inclined to take the survey than were lower students (GPA \Rightarrow 3.195 is the cut off recommended by the classification tree.) Among the lower respondents, older students were more likely to take the survey. Here, age 26 is the cut off recommended by the classification tree. Surprisingly, academic level (undergraduate and graduate), gender, and race were not identified as crucial factors in determining the response rate, although the latter two were identified by other studies, as discussed in the literature review, as factors contributing to survey bias. Although SAT-combined is considered significant, it is located at the bottom of the tree, indicating it is the least significant predictor. Again, this is not in alignment with other studies concerning non-response bias.

It is noteworthy that these results do not imply that the responses collected from the survey are biased against younger and lower GPA students. As mentioned before, the second, third, and fourth level partitions are less crucial than the first level. Nonetheless, *the first level classification conveys that science and engineering students, law students, and older students may be more vocal in this survey.*

Figure 9 shows the output yielded from the preceding configuration. There are similarities and dissimilarities between the JMP’s tree and the Insightful tree. Unlike its JMP counterpart, the Insightful model suggests that GPA contributes to the most decisive split of survey responses, with SAT as the second most crucial factor. Among those students whose SAT is below 3.195, age makes the most decisive split, whereas for those lower SAT students, college affiliation is an important predictor. The tree trunk can go on and

Figure 8: Classification tree using Entropy in JMP

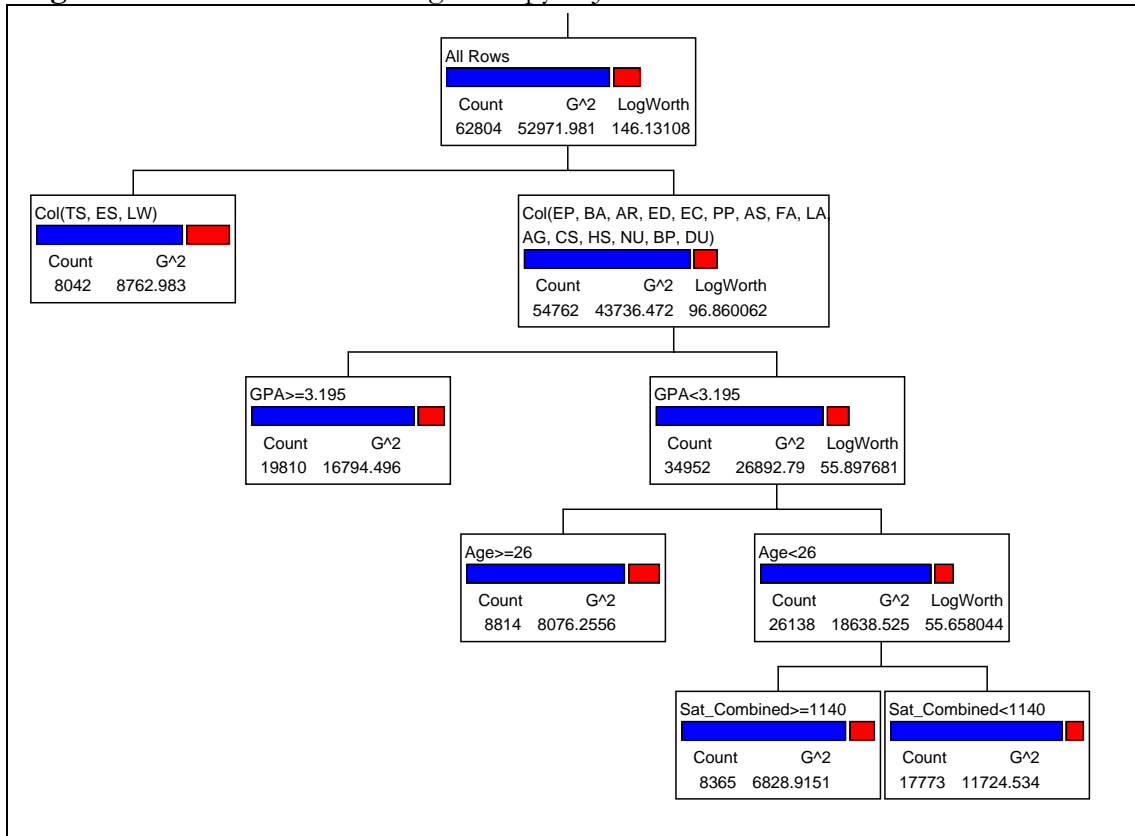


Figure 9: Classification tree using GINI in Splus Insightful Miner

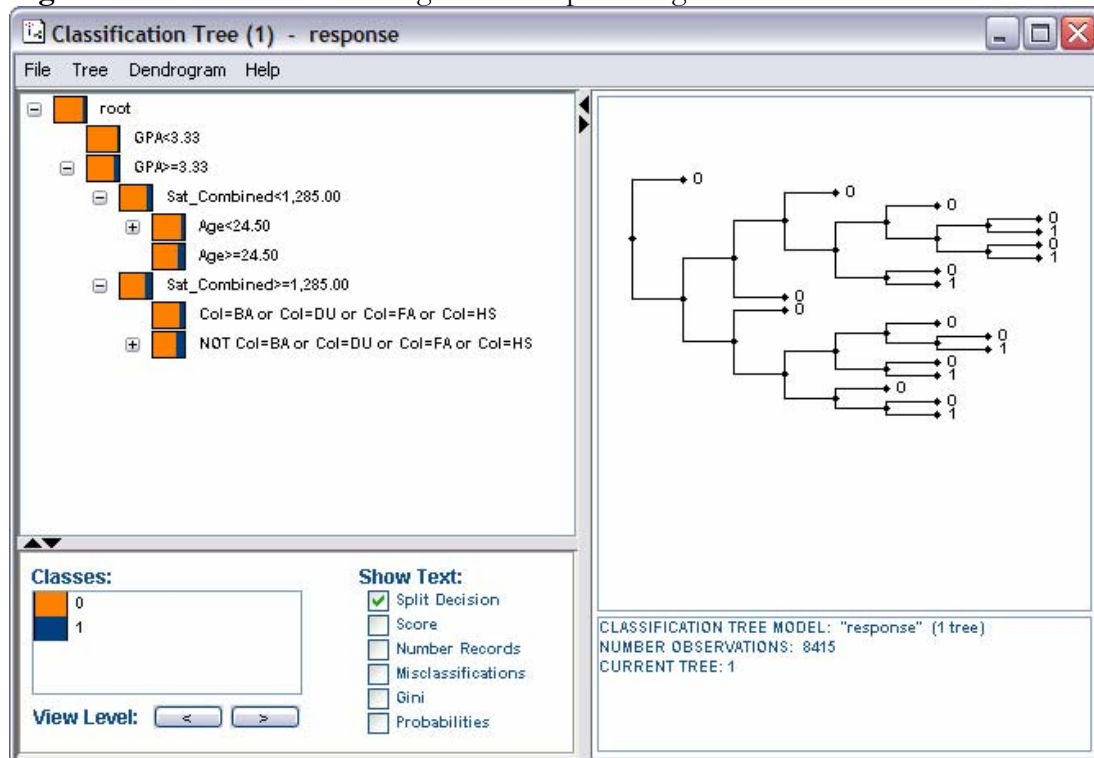


Table 3: College code and description

College Code	Description
AG	MORRISON SCHOOL OF MANAGEMENT
AR	DESIGN
AS	INTERDISCIPLINARY ARTS/SCIENCE
BA	BUSINESS
BP	SCHOOL OF GLOBAL MANAGEMENT AND LEADERSHIP
CJ	CRIMINAL JUSTICE
CS	CRONKITE SCHOOL OF JOURNALISM AND MASS COMMUNICATION
DU	UNIVERSITY COLLEGE
EC	EAST COLLEGE
ED	EDUCATION
EE	EXTENDED EDUCATION
EN	APPLIED SCIENCE, ENGINEERING & TECHNOLOGY
EP	TEACHER EDUCATION & LEADERSHIP
ES	ENGINEERING
FA	FINE ARTS
GC	DIVISION OF GRAD STUDIES
HS	HUMAN SERVICES
LA	LIBERAL ARTS & SCIENCES
LW	LAW
NU	NURSING
PP	PUBLIC PROGRAMS
SW	SCHOOL OF SOCIAL WORK
TS	COLLEGE OF SCIENCE & TECHNOLOGY

on, but like the JMP tree, gender and ethnicity are not considered significant predictors to survey responses. To obtain a usable model for suggesting courses of action, a less complex tree consisting of a smaller set of predictors was retained by the research team.

CONCLUSION

Although different remedies, such as use of Web-based survey and incentives, have been employed to countermeasure nonresponse in surveys, nonresponse bias continues to be a challenge to survey research because there is no fool-proof method to reach a desirable response rate. Stepwise logistic regression is widely applied to analyze the predictors to survey responses and nonresponses, but its weaknesses are well-documented, thus the validity of its inference should be under scrutiny. To further examine the issue, both data visualization and data mining techniques are used in this study. Unlike conventional hypothesis testing, data visualization and EDA do not depend on probabilistic inferences. The crux of data

visualization is to unveil the hidden pattern in the data. Similarly, data mining is regarded as an extension of EDA for its emphasis on pattern recognition (Luan, 2002). Unlike logistic regression, classification trees in data mining are capable of ranking independent variables in terms of their predictive power to splitting the data. More importantly, some variants of classification trees are equipped with cross validation features, thus alleviating the problem of a model capitalizing on sampling errors.

Examining the findings, the ASU team will refrain from making bold claims, such as asserting that these findings have overthrown prior research. Our data were collected in only one institution and have yet to be replicated. On the other hand, these surprising results can encourage other researchers to go beyond conventional multivariate statistical tools to explore the issue of nonresponses through alternate methodologies. As identified in this analysis, differing statistical methodologies can result in different interpretations of data. Further

research is needed to explore this issue and identify the most accurate procedure for analyzing future survey data.

REFERENCES

- Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth International Group.
- Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M., and Ouimet, J. A. (2003). College student responses to web and paper surveys: Does mode matter? *Research in Higher Education* 44(1), 1–19.
- Cassell, D. (2006, October). *Sample survey data and the process you ought to be using*. Paper presented at the Annual Conference of West Users of SAS software, Irvine, CA.
- Cunradi, C. B., Moore, R., Killoran, M., & Ames, G. (2005). Survey nonresponse bias among young adults: The role of alcohol, tobacco, and drugs. *Substance Use & Misuse*, 171–185.
- DiGangi, S., Kilic, Z., Yu C. H., Jannasch-Pennell, A., Long, L., Kim, C., Stay, V., & Kang, S. (2007). 1 to 1 computing in higher education: A survey of technology practices and needs. *ACE Journal*, 15(4). Retrieved October 29, 2007 from http://www.editlib.org/index.cfm?fuseaction=Reader.ViewAbstract&paper_id=22813
- Glymour, C. (2001). *Mind's arrows: Bayes and Graphical causal models in psychology*. Cambridge, MA: MIT Press.
- Greenland, S. (1989). Modeling and variable Selection in epidemiologic analysis. *American Journal of Public Health*, 79, 3, 340-349,
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques (2nd ed.)*. Boston, MA: Elsevier.
- Haziza, D., & Rao, N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32, 53-64.
- Insightful, Inc. (2006). *Splus Insightful Miner* [Computer software and manual]. Seattle, WA: The Author.
- Johnson, W. T., & Delamater, J. D. (1976). Response effects in sex surveys. *The Public Opinion Quarterly*, 40, 165-181.
- Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). Web and Mail survey response rates. *The Public Opinion Quarterly*, 68, 94-101
- Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.
- Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of official statistics*, 18, 257-273
- Kypri, K., & Gallagher, S. J. (2003). Incentives to increase participation in an internet survey of alcohol use: a controlled experiment. *Alcohol & Alcoholism*, 38(5), 437–441.
- Luan, J. (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), *Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). PA: Josey-Bass.
- Moore D. L., and Tarnai J. (2002). Evaluating nonresponse error in mail surveys. In: Groves, R. M., Dillman D. A., Eltinge J. L., and Little R. J. A. (eds.), *Survey nonresponse*, (pp. 197–211). New York: John Wiley & Sons.
- Penuel, W. R. (2006). Implementation and effects of one-to-one computing initiatives: A research synthesis. *Journal of Research on Technology in Education*, 38, 329-348.
- Porter, S. R., & Whitcomb, M. E. (2004). Understanding the effect of prizes on response rates. *New Directions for Institutional Research*, 121, 51-62.
- Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement and personality. *Research in Higher Education*, 46, 127-152.
- Porter, S. R. Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 121, 63-73.
- Porter, S.R. & Umbach, P.D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education*, 47, 229-247.
- Quinlan, J. R. (1993). *C4.5 programs for machine learning*. San Francisco, CA: Morgan Kaufmann.
- Rigotti, N. A., Lee, J. E. & Wechsler, H. (2000). US College students' use of tobacco products: Results of a national survey. *Journal of the American Medical Association*, 284, 699-705.
- Rosella, Inc. (2007). Decision tree classification software. Retrieved January 29, 2007 from <http://www.roselladb.com/decision-tree-classification.htm>

Yu, Jannasch-Pennell, DiGangi, Kim & Andrews, Data visualization and mining for survey responses

- SAS Institute. (2006). *JMP* [Computer software and manual]. Cary, NC: The Author.
- Sax, L. J., Gilmartin, S. K., and Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, 44, 409–432.
- Singer, E., Groves, R. M., and Corning, A. D. (1999). Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly*, 63, 251–260.
- Singer, E., van Hoewyk, J., and Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 171–188.
- Sosdian, C. P., & Sharp, L. M. (1980). Nonresponse in mail surveys: Access failure or respondent resistance. *The Public Opinion Quarterly*, 44, 396-402.
- Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55, 4, 525-534.
- Wechsler, H., Lee, J. E., & Rigotti, N. A. (2001). Cigarette use by college students in smoke-free housing: Results of a national study. *American Journal of Preventive Medicine*, 20, 202-207.
- Willimack, D. K., Schuman, H., Pennell, B. E., & Lepkowski, J. M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *The Public Opinion Quarterly*, 59, 78-92.
- Yammarino, F. J., Skinner, S. J. & Childers, T. L. (1991). Understanding mail survey response behavior: A meta-analysis. *The Public Opinion Quarterly*, 55, 613-639.

Acknowledgement

Special thanks to Charles Kaprolet for editing this article.

Citation

Yu, C.H., Jannasch-Pennell, A., DiGangi, S., Kim, C., & Andrews S. (2007). A data visualization and data mining approach to response and non-response analysis in survey research. *Practical Assessment Research & Evaluation*, 12(19). Available online: <http://pareonline.net/getvn.asp?v=12&n=19>

Author

Correspondence concerning this paper should be addressed to

Chong Ho Yu, Ph.D.
PO Box 612
Tempe AZ 85280
480-727-0670
Chongho Yu [at] gmail.com