Using Multivariate Techniques to Measure the Performance of R&D Programs: A Case Example

Isabelle Bourgeois

Natural Sciences and Engineering Research Council of Canada, University of Ottawa, Ottawa, Canada

Introduction

Performance management systems implemented in science-based government organizations have traditionally focused on research inputs and activities, rather than outputs or outcomes. However, recent legislative changes in several countries now require individual programs to report on their progress towards the achievement of organizational and governmental strategic objectives. In a substantive field where peer review remains the standard evaluation method against which scientific success is judged, performance measurement activities have often been articulated around complex techniques taken from the sciences and economics that yield little useful information to key decision makers (Geisler, 2002; McDonald & Teather, 2000; Roessner, 2002).

The data required to evaluate R&D programs in this context of accountability include more than the broad economic or scientific indicators used in the past: "Methods are needed that capture more fully the noneconomic benefits from research…or at least the benefits not easily transformed into monetized form…" (Roessner, 2002, p. 8). In other words, although traditional performance indicators can provide the basic description of whether an organization is producing its expected outputs, more detailed information is now needed for program planning and resource allocation; such information is best collected under the auspices of program evaluation and monitoring (Cozzens, 1997; Geisler, 1999).

Questionnaire surveys are one mechanism by which noneconomic data can be collected. Examples of measures that can be collected with surveys include industry awareness of government research and researchers, satisfaction with past interactions, level of trust in researchers and staff, and types of new processes or products introduced to market (Roessner, 2002, Rogers, 1998). These indicators point to some of the immediate and intermediate outcomes of government R&D research and would be difficult to measure using traditional means such as peer review or economic indices.

One of the difficulties associated with survey research and use by R&D organizations is the simplicity of the analyses usually conducted on the data collected. Most often, data analysis consists in reporting descriptive data for each performance indicator with little exploration of the relationships that may exist between variables (Scheirer, 2000). In many cases, a more sophisticated analysis based on the multivariate techniques developed in the social sciences may yield information of use to decision makers at little additional cost. For example, a study conducted by Harman (2004) is typical of many studies on the outcomes of

research and research training. In this particular case, the researcher sought to identify the differences between two types of training programs for Ph.D. students in science-based departments. The first type followed the traditional model, with students completing coursework, conducting research and writing their dissertation in a university laboratory and under the guidance of one faculty member. The second type of program used Cooperative Research Centres, defined as doctoral programs that integrate industry needs with professional development, emphasizing "industry-ready" graduates with a broader educational experience linked to the needs of industry research users. The study used a survey questionnaire administered to doctoral students in each of the two groups. Findings were reported as survey frequencies, with t-tests used to identify statistically significant differences. All significant differences were reported and interpreted as such, even when the difference in frequencies was minimal (e.g., "overall experience as a Ph.D. student" was identified as statistically significant, even though the reported frequencies were of 66.7% for the first group and 64.3% for the second group). No further analyses were conducted to add to or enhance the conclusions drawn as a result of the findings.

The purpose of this paper is to illustrate the more easily accessible multivariate analysis techniques in an effort to demonstrate the value of moving beyond the commonly used economic indicators and descriptive statistics in telling a program's performance story. The study presented here describes a multivariate analysis conducted using data from a postgraduate scholarship program administered by the Natural Sciences and Engineering Research Council (NSERC), a Canadian federal government agency that supports university research and the training of Highly Qualified Personnel (HQP). The data were collected

through ongoing performance measurement activities. Although the survey instrument used did not seek to measure the direct outcomes of R&D, it was selected as an illustrative case because of the availability of the data and because of its use of both scale and categorical data.

Methodology

A web based survey was used to collect the data analyzed in this study. The survey was administered in the summer of 2005 and focused on master's and doctoral students who had recently received the final installment of their Postgraduate Scholarship (N = 901). A total of 101 invitations to participate in the survey, distributed via e-mail, were returned undelivered. Out of the 800 e-mails that reached participants, 557 surveys were completed for a response rate of 69.6%. The survey results were therefore estimated to be accurate ± 3%, 19 times out of 20.

The instrument was divided into several sections: education history, the NSERC award, experiences during the award, and future plans. The purpose of the survey was to gather attitudinal and factual data about the award recipients and their experiences in order to collect concrete data on the Postgraduate Scholarship program's performance indicators.

Variables

The independent variables used include the respondent's gender, the type of award received (master's or doctoral), and the main field of study. The dependent variables were divided into 4 seven-point scales, each one relating to a different aspect of the student's experience with the PGS award. It should be noted that most

of the multivariate analyses presented in this paper lend themselves better to interval rather than scale variables; nevertheless, the information obtained through these analyses can inform decision makers on the merit and worth of the program, and in this sense, the results of the analysis can be considered valid. However, although this is a legitimate use of analytical techniques for program monitoring purposes, these techniques would be under more stringent requirements if they were used in a social research context.

Other variables were also collected in the survey on dissemination, satisfaction with the service provided by NSERC during the award, gender of principal supervisor, and future plans. These variables were not included at this stage of analysis, but could be integrated in the study at a later date.

Preliminary Examination of Data

In order to determine whether the data collected through the survey could be analyzed using multivariate techniques, descriptive statistics were compiled and examined. This preliminary examination reveals little missing data, with sample sizes for each question varying between 541 and 551. In addition to this, all of the dependent variables were correlated to get a better sense of the relationships that may exist between them. Although the correlation matrix is too large to be reproduced here, it appeared upon examination that many of the dependent variables shared a certain amount of common variance, which indicates that further multivariate analysis may be helpful to better understand the results of the survey.

The preliminary examination of data also included a verification of the statistical assumptions most critical to multivariate analysis, that of *normality* and

homoscedasticity (equality of the variance-covariance matrices). Normality assumes that the frequencies of a given dependent variable are distributed normally across the range of possible values of the independent variable. Multivariate normality therefore assumes that the joint effects of multiple variables are normally distributed. Univariate normality was verified on all scale variables in the present study using both graphical examinations and statistical tests (e.g., Kolgomorov-Smirnov and Shapiro-Wilks) and the conclusions reached in this analysis were assumed to hold for multivariate normality. In this particular study, many independent variables were found to violate the assumption of normality, especially in the attitudinal scales. The assumption of homoscedasticity requires that there are no substantial differences in the amount of variance of one group versus another for the same variables. Box's M Test is typically used for establishing the equality of variance-covariance matrices. The results of the test in this case were significant, which means that the null hypothesis of equality does not hold. Data transformations were therefore made according to Osborne (2002), but further normality and homoscedasticity assumptions still did not hold. Given these results, it was determined that this was probably due to the fact that PGS recipients are selected amongst the best students in the country; for this reason, it can be expected that they would share some common traits. However, the effect of these violations on the validity of the multivariate analysis is weakened because of the large sample size. The *F* statistic, which was used to identify significant differences, is known to be quite robust against violations of these two assumptions (Lindman, 1974).

Findings: Multivariate Respondent Profile

Data analysis techniques used within the context of performance measurement and evaluation are often limited to calculating the frequency of responses to certain items or conducting t-tests or ANOVAs on individual variables. Some of the advantages of using these techniques are efficiency, speed, and ease of interpretation. They also constitute an excellent starting point for more in-depth analysis of the interaction that may occur between dependent variables. The variability in the frequencies noted in the items included in question A2, for example, hints at the fact that other factors may be at the source of the difference between the respondents' answers. The third item, "I accumulated a lot of debt during my undergraduate degree", for instance, shows a wide range in the number of respondents who strongly disagreed, strongly agreed, or neither agreed nor disagreed. Although this is an interesting finding in and of itself, it also creates further questions, such as: "Is there a difference between men and women in terms of undergraduate debt load? Do students from different research fields end up with similar amounts of debt? Are doctoral students more likely than master's students to accumulate debt as undergraduates? More importantly, is the variability in undergraduate debt dependent upon a combination of gender, field, and type of graduate award?" These questions could be asked for any of the items included in the four different scales used in the survey.

A Multivariate Analysis of Variance (MANOVA) may yield some information to answer questions about the combined effect of the dependent variables. The MANOVA method tests the null hypothesis of equality of vectors of means on multiple dependent variables across groups. In the present study, a MANOVA was

conducted on each of the four 7-point scales of the survey, using the gender, award type and field variables. Tables 1 through 4 provide a summary of the MANOVA similar to those produced by most statistical analysis software packages, using the Wilk's Lambda statistic. This statistic is typically the one used in most MANOVAs, because "it examines whether groups are somehow different without being concerned with whether they differ on at least one linear combination of the dependent variables (Hair, Anderson, Tatham & Black, 1998, p. 351).

Table 1

MANOVA Undergraduate Experiences and Reasons for Continuing Studies

Effect	Λ	*F*	*df*	Error *df*	η^2	*p*
Gender	.966	1.635	11	510.000	.034	.086
Award	.984	0.730	11	510.000	.016	.710
Field	.793	2.208	55	2364.261	.045	.000*
Gender * Award	.971	1.363	11	510.000	.029	.187
Gender * Field	.882	1.180	55	2364.261	.025	.173
Award * Field	.867	1.343	55	2364.261	.028	.048*
Gender * Award * Field	.858	1.440	55	2364.261	.030	.019*

Table 2

MANOVA Research Capability of Department in Which Award Was Held

Effect	Λ	*F*	*df*	Error *df*	η^2	*p*
Gender	.984	2.122	4	508.000	.016	.077
Award	.988	1.603	4	508.000	.012	.172
Field	.925	2.018	20	1685.795	.020	.005*
Gender * Award	.986	1.864	4	508.000	.014	.115
Gender * Field	.957	1.137	20	1685.795	.011	.303
Award * Field	.961	1.019	20	1685.795	.010	.435
Gender * Award * Field	.957	1.120	20	1685.795	.011	.321

Table 3

MANOVA Experiences During NSERC Award, Including Quality of Supervision

Effect	Λ	*F*	*df*	Error *df*	η^2	*p*
Gender	.980	1.163	9	501.000	.020	.317
Award	.962	2.178	9	501.000	.038	.022*
Field	.897	1.230	45	2244.197	.022	.142
Gender * Award	.990	0.556	9	501.000	.010	.833
Gender * Field	.922	0.913	45	2244.197	.016	.639
Award * Field	.915	1.000	45	2244.197	.018	.472
Gender * Award * Field	.937	0.726	45	2244.197	.013	.913

Table 4

MANOVA Skills Improvement During Award

Effect	Λ	*F*	*df*	Error *df*	η^2	*p*
Gender	.985	.836	9	499.000	.015	.583
Award	.988	.680	9	499.000	.012	.727
Field	.871	1.562	45	2235.251	.010	.027
Gender * Award	.974	1.503	9	499.000	.026	.144
Gender * Field	.907	1.094	45	2235.251	.019	.311
Award * Field	.906	1.103	45	2235.251	.019	.295
Gender * Award * Field	.916	.983	45	2235.251	.017	.505

The tables reveal that the field variable seems to have an impact by itself and in combination with one or both other dependent variables in the question on undergraduate experiences. The field variable also had an impact on its own on the question on the research capability of the department in which the respondent studied. The award variable was also found to have an impact on the items of the question on quality of supervision. This difference between master's and doctoral students' responses to this scale can be well understood, since the relationship between student and thesis supervisor is likely to be experienced in a different way at each level.

A close relative of the MANOVA technique is the Discriminant Function Analysis (DFA). Rather than looking at the influence of the independent variables on the responses to the scale items, it assesses the extent to which the scale items are useful in classifying respondents in groups within the independent variables. In other words, it allows the analyst to predict group membership by using the responses to the scale items. Discriminant function analysis always produces (Number of groups within independent variable - 1) functions, and identifies the percentage of variance accounted for by each function. Most interestingly, however, it provides loadings similar to those found in factor analysis (described in the following section) for each of the dependent variables, or scale items. These loadings can then be used to assess the extent to which the items contribute to the difference between respondents in each level of the independent variable. Although a more in-depth discussion of DFA is warranted, it is not possible to provide one in this paper due to space limitations.

Eight discriminant function analyses were conducted in this study, one for each of the independent variables "award received" and "field of study" in each of the three scales included in the survey instrument (2 variables x 4 scales = 8 DFA).

Results of the DFA, Undergraduate Experiences

The items included in this question deal with undergraduate experiences and the reasons that brought respondents to pursue graduate studies. The two items loading highest on the Award Received variable were "I would have gone on or stayed in graduate school even without NSERC support" and "I was exposed to research during my undergraduate years". The latter also loaded highly on the Field variable, along with "It is difficult to find a job in my field without a graduate

degree". The importance of exposure to research at the undergraduate level appears to be particularly important in the decision to attend graduate school, and has an impact on whether students choose to pursue doctoral studies. In addition to this, it appears to be more important in certain fields than others.

Results of the DFA, Research Capability of the Department

The items in this scale focused on the quality of the learning environment provided to the respondents. The items with the highest loading on the Award Received variable included "Technical support" and "Faculty in the department", while those with high loadings on the Field variable included "Laboratory equipment and instruments" and "Buildings, laboratory space, office space". These loadings are logical, when considering that students who obtain the necessary support and have access to faculty members in their department may be more likely to pursue doctoral studies, while the differences in laboratory equipment and space are likely to vary according to the field of study of the respondent.

Results of the DFA, Quality of Supervision

This scale focused on the support provided to the student by his or her advisor, as well as other experiences related to the PGS award. The items loading most highly on the Award Received variable included "The experience gained during my NSERC award increased my desire to pursue a career in research" and "Funding from NSERC will help me to complete my degree faster". The same two items also loaded highest on the Field variable, which suggests that these items allow a particularly good discrimination between both types of awards, as well as between fields of study.

Results of the DFA, Skills Improvement

This scale required respondents to assess the extent to which various research skills had improved over the course of their NSERC award. The items loading most highly on the Award Received variable included "Theoretical knowledge of the discipline", "Analytical techniques/experimental methods", and "Communication/ presentation". Those loading highest on the Field variable included "Project management" and "Interdisciplinary research". Once again, doctoral students are more likely to report an improvement in their knowledge and research skills because of their years of experience compared to master's students, while project management and interdisciplinary research skills are likely to vary according to the field of study.

Findings: Exploration of Scale Properties

Aside from providing a profile of the respondents to the survey, it is also possible to use multivariate analysis to verify the reliability of the scales used in the instrument. Two analysis techniques were used for this purpose: Exploratory Factor Analysis (EFA) and internal consistency verification.

Exploratory Factor Analysis

Exploratory Factor Analysis is a data summarization technique that identifies underlying, or latent, dimensions in a dataset that, when interpreted, describe the data in a much smaller number of concepts than the original individual variables. This is done by decomposing the correlations or covariances between variables to identify the structure of relationships among variables. Latent variable models such

as EFA are based on the hypothesis that variation in a latent variable will induce variation in the observed variables to which it is linked. In other words, the responses to the items on the PGS survey scales are hypothesized to vary as a function of one or several latent variables.

An exploratory factor analysis was conducted on each of the four scales included in the survey, as well as on the combined scales in order to identify the underlying dimensions present in the dataset. Bartlett's Test of Sphericity was first conducted for each analysis to determine whether or not a factor analysis was appropriate on the data collected. This test provides the statistical probability that the correlation matrix has significant correlations among at least some of the variables (Hair et al., 1998). All four tests revealed a significant difference; therefore, a factor analysis was deemed appropriate in all cases.

An unweighted least squares (ULS) extraction was selected for the EFA, since this extraction method makes no assumptions of normality, and normality was not clearly demonstrated graphically or statistically for the dataset, even when transformations were applied to the data. The results of the EFA conducted for the combined scale items are presented below for illustrative purposes.

Ten factors with eigenvalues higher than 1 were extracted for the first scale analyzed. These factors account for 66.3% of the total variance and were retained for interpretation. The loadings therefore indicate the degree of correspondence between the item and the factor, with higher loadings indicating greater representation of the factor. Because the unrotated factor matrix yielded inconclusive results (i.e., some items loaded highly on more than one factor, or poorly on all five factors), the factors were rotated orthogonally using the

VARIMAX rotation to obtain a clearer picture of the loadings of each item. Factor rotation, as the name implies, involves rotating the reference axes of the factors about the origin until a different position has been reached.

Even though the analysis yielded 10 factors, the scale items loaded highly only on 9 of them. Many of the factors represent items that are grouped together on the survey, such as Factor 2, with items such as "I felt that I received adequate supervision from my advisor", and "I met regularly with my supervisor to discuss my work". Other factors only had one item, such as Factor 5, with "I do not want to go into debt for graduate education". The latent dimensions identified for each factor as well as the items loading highly on each one are presented below.

Factor 1: Learning and Skill Development

- Funding from NSERC will help me to complete my degree faster
- Theoretical knowledge of the discipline
- Analytical techniques/experimental methods
- Use of laboratory equipment or instruments
- Project management
- Communication/presentation
- Supervision of other students
- Writing reports and publications
- Interdisciplinary research

Factor 2: Supervision

- I felt that I received adequate supervision from my advisor
- I met regularly with my supervisor to discuss my work

- I was encouraged to present my work outside my group
- I received useful feedback on my research work from my supervisor
- My supervisor helped me progress through my degree requirements

Factor 3: Resources

- Faculty in the department
- Laboratory equipment and instruments
- Buildings, laboratory space, office space
- Technical support

Factor 4: Career Orientation

- I enjoyed my undergraduate student life
- Graduate studies are an important part of my career goals
- I would recommend my field of studies to others
- I would have gone on to or stayed in graduate school even without NSERC support
- The experience gained during my NSERC award increased my desire to pursue a career in research

Factor 5: No Debt

- I do not want to go into debt for graduate education

Factor 6: Encouragement

- I was exposed to research during my undergraduate years
- My friends are pursuing graduate degrees
- My family encouraged me to pursue graduate studies
- A professor I had encouraged me to pursue graduate studies

Factor 7: Collaborative Experiences

- Collaborative research with industry and/or government researchers

Factor 8: Debt Load

- I accumulated a high debt during my undergraduate degree
- I had more debt at the end of my NSERC award than at the beginning

Factor 9: Job Prospects

- It is difficult to get a job in my field without a graduate degree
- The experience gained during my NSERC award will improve my prospects of getting a permanent job in a relevant area

Although many of the factors grouped items that had been part of the same scale on the survey instrument, some of the factors clearly demonstrate linkages between items that had been part of separate scales. This is an important issue to consider in the improvement of the survey instrument for future use, and will be addressed to a greater extent in the following section on scale reliability.

A second type of factor analysis, Confirmatory Factor Analysis (CFA), could be used to verify the results of the EFA. The goal of CFA, as its name implies, is to confirm the fit of a given structure to a dataset. Fit indices are used to determine whether a given model fits the data. However, a CFA was not conducted in this study due to the specialized software required for such an analysis, since the goal of this paper is to demonstrate feasible methods that can be used in program monitoring with standard, accessible software packages.

Scale Reliability

Scales that exhibit a high degree of internal consistency are considered reliable since they minimize the contribution of random error to the item scores. The statistic used to measure the internal consistency of a scale is the alpha coefficient and is interpreted in the same way as a correlation. The alpha coefficients for each of the three scales as well as for the combined scales were calculated to get a sense of the internal consistency of the scales as presented in the survey instrument. These coefficients are shown in Table 5 below.

Table 5

Alpha Coefficients, Scales as Presented in Instrument

Scale	α
Question A2	0.59
Question C1	0.81
Question C2	0.77
Question C3	0.85
All Items Combined	0.82

The reliability for each of the scales was rather high, with the exception of a more moderate coefficient for question A2. The alpha coefficient of all combined items suggests that all of the items are well suited to the instrument. This means that no major adjustments to the instrument are necessary. An analysis of the reliability of the EFA results was also conducted in order to verify whether the 9-factor structure obtained would yield higher internal consistency if the survey items were presented in this manner in the instrument. Table 6 summarizes the EFA-based reliability estimates.

Table 6

Alpha Coefficients, Scales as Presented in Instrument

Factor	α
1	0.84
2	0.91
3	0.81
4	0.64
6	0.50
8	0.42
9	0.24

The reliability coefficients for the first three factors suggest that the items in each of these revised scales are consistent with one another and could form new scales on the instrument. No alpha coefficients were calculated for Factors 5 and 7, since these only had one item each. The four other factors, 4, 6, 8, and 9, displayed relatively low alpha coefficients and were deemed to have poor internal consistency. Therefore, although the first three factors show high internal consistency, no changes have been made to the survey instrument in light of the EFA results.

Conclusion

The purpose of this paper was to illustrate some of the commonly accessible multivariate data analysis methods for monitoring the performance of science-related programs. The data collected in the context of performance management often do not yield to the stringent requirements of the research methods developed in the social sciences; however, the use of some of these methods may provide important information to decision-makers charged with monitoring program progress towards outcomes. In the present case, the multivariate analysis allowed

evaluators to move beyond the descriptive statistics normally used in survey-based studies and to make claims related to the interaction of factors such as the type of award held, the recipients' gender, and the field of study on the survey results. It also provided information on the quality of the instrument used and suggested potential changes that could be made in an effort to further improve the instrument. The findings of the analysis confirmed that the program under study is achieving its outcomes as stated in the program logic model and identified areas in which different variables have a combined impact on program outcomes. For example, the MANOVA analysis revealed that the field of study of a program participant may have an impact on his or her undergraduate experiences as well as on the research capability of the department in which the award is held. This certainly warrants further investigation, as the PGS program is assumed to have similar outcomes across all disciplines. The award variable was found to have an effect on the quality of supervision scale items, which suggests that Master's students and Ph.D. candidates have different relationships with their advisors. A more in-depth investigation of this finding may reveal different training modes for students at each level, and may provide further clues as to how NSERC can best ensure quality training for all PGS recipients.

The exploration of scale properties also provided important information on the instrument currently used to monitor program outcomes. The findings of this segment of the analysis revealed that the scales as they are designed have an acceptable level of internal consistency and that although the survey as a whole is multidimensional, each scale seems to focus on one particular latent variable, or factor. The alternate distribution of items obtained through the Exploratory Factor

Analysis did not provide a better model for survey design, and so the original survey structure was maintained for future iterations of the study.

Taken together, it is hoped that these results will lead to increased use of survey findings and better decision-making on the program's design and delivery in the future. The use of multivariate analysis methods has provided evaluators and program managers with additional tools in the monitoring of program outcomes, and has also given them more confidence in the survey instrument designed for this purpose.

About the Author

Isabelle Bourgeois, M.Ed., is a Program Evaluation Officer at the Natural Sciences and Engineering Research Council of Canada. She is also pursuing a doctoral degree in measurement and evaluation at the Faculty of Education, University of Ottawa. Her research interests focus on evaluation capacity building in Canadian federal government departments and agencies.

Acknowledgements

A sincere thank you to Professor André Rupp for his insightful comments on an earlier draft of this paper.

Many thanks to the Natural Sciences and Engineering Research Council of Canada for the data used in the examples provided in the paper.

References

Cozzens, S. E. (1997). The knowledge pool: Measurement challenges in evaluating fundamental research programs. *Evaluation and Program Planning, 20* (1), 77-89.

Geisler, E. (2002). *What Do We Know About: R&D Metrics in Technology-Driven Organizations.* Paper prepared by invitation for the Center for Innovation Management Studies at North Carolina University. Retrieved September 14, 2005, from http://cims.ncsu.edu/documents/rdmetrics.pdf

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis (5th Edition).* Upper Saddle River, NJ: Prentice Hall.

Harman, K. M. (2004). Producing 'industry-ready' doctorates: Australian Cooperative Research Centre approaches to doctoral education. *Studies in Continuing Education, 26* (3), 387-404.

Lindman, H. R. (1974). *Analysis of Variance in Complex Experimental Designs.* San Francisco, CA: W.H. Freeman.

McDonald, R., & Teather, G. (2000). Measurement of S&T performance in the Government of Canada: From outputs to outcomes. *Journal of Technology Transfer, 25*, 223-236.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research and Evaluation, 8* (6). Retrieved December 4, 2004, from http://PAREonline.net/getvn.asp?v=88n6.

Roessner, D. (2002). *Outcome Measurement in the United States: State of the Art.* Paper presented at the Annual Meeting of the American Association for the Advancement of Science, Boston, MA.

Rogers, M. (1998). *The Definition and Measurement of Innovation.* Melbourne Institute Working Paper, No. 10/98. Retrieved September 14, 2005 from, http://melbourneinstitute.com/publications/working/1997-1999wp.html

Scheirer, M.A. (2000). Getting more "bang" for your performance measures "buck". *American Journal of Evaluation, 21*, 139-149.