Appendix D. Analysis of Citation Counts

D.1 Background Statistics on Citation Counts

For citations measured by fractional counts in an expanding journal set, 95% of the observations by institution-year are between 0 and 45,075, the mean number per institution-year is 8,498, the median is 3,238, and the standard deviation is 13,129. Average citation counts vary by field. The average number of citations per institution-year by field group is as follows: biology-life sciences-agricultural sciences (3,958), medical sciences (2,325), engineering-math-physical sciences (1,826), social sciences-psychology (369), and computer science (21).

D.2 Analyses of Fractional Citation Counts in an Expanding Journal Set

A series of regressions were performed to determine, for each set of personnel and financial variables, which were most highly correlated with fractional citation counts in an expanding journal set. The data set was aggregated to the institution-year level. Financial variables were deflated using the GDP deflator; personnel variables were lagged by one year and financial variables by two years. Analyses were conducted using stepwise linear regression. We typically retained any variable that, when entered into the regression, increased r-squared by at least 0.01. All such variables were highly statistically significant when entered.

The set of faculty counts variables (all lagged by one year) in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) Full Profs (0.143), 2) Assoc Profs (0.298), and 3) Total Fac (0.325). These are the same set of variables, and order of entry, as for fractional publication counts, although the amount of variance explained is reduced. Instructors and assistant professors did not meet the 0.01 threshold. We retained the count of full professors, associate professors and total faculty for inclusion in later regressions. When only these three independent variables were included, the coefficients were 27.4 for full professors, -95.5 for associate professors, and 21.2 for total faculty. Since addition of an associate professor also increases total faculty, the net effect of adding an associate professor would be -74.3.

The set of Carnegie classifications and type of control variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) R-1 (0.247), 2) Private control (0.269) and 2) Medical (0.278). Types R-1 and Medical appeared in the fractional publication count analysis; private control did not show up in the publication analysis but makes a small (0.02) positive contribution here to r-squared. The other Carnegie classifications did not enter the regression equation. We retained R-1, private control, and Medical.

The set of academic R&D expenditures variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) Total academic R&D expenditures (0.640), 2) federally financed academic R&D expenditures (0.755), and 2) other academic R&D expenditures (0.806). These are the same variables as were found in the fractional publication counts analysis. The remaining academic R&D expenditure variables did not meet a 0.01 threshold. The remaining variables included various academic R&D expenditure types (industry, the institution, or state/local government financed), total academic basic research expenditures (total, federally financed, and non-federally financed), research equipment expenditures (total, federally financed, and non-federally financed), institute financed organized research, and unreimbursed indirect costs.

The set of enrollment variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) number of S&E grad students (0.192) and 2) fall enrollment of undergraduates (0.297), and fall enrollment of graduate students (0.328). The first two of these variables were present in the analysis of fractional publication counts; the last one is additional. Total fall enrollment did not enter the model. All fall enrollment variables were institution-wide rather than S&E specific. The coefficient for undergraduates was negative, perhaps reflecting less publication activity in some institutions with more undergraduate students.

Among the set of postdoctoral count variables, only total postdoctorates entered the regression with an incremental r-squared (0.842) in excess of 0.01. This variable was also the most important in the regression for fractional publication counts, and indeed, had a larger r-squared here than the total provided by all four variables that entered into the regression for fractional publication counts. The remaining variables that did not meet a 0.01 threshold included postdoctorates with M.D.s in total and by support type (federal research grants, federal fellowships, federal traineeships, and non-federal sources), postdoctorates without M.D.s in total and by support type, and postdoctorates by support type.

The set of non-faculty research staff count variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) non-faculty research staff (0.414), and 2) non-faculty research staff without M.D.s (0.441). The first of these, but not the second, appeared in the regression for fractional publication counts.

The set of degrees awarded variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) S&E doctorates from the NSF Survey of Earned Doctorates (0.401), 2) S&E BA/BS degrees (0.467), and 3) S&E master's degrees (0.480). The first two of these variables entered into the regression for fractional publication counts. S&E doctorates from the IPEDS Completions Survey did not meet the 0.01 incremental threshold. The coefficient for BA/BS degrees was negative, consistent with the finding for number of undergraduate students. The two doctoral degree variables had a correlation of 0.973.

The complete set of retained variables identified above was entered into a stepwise regression. The variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) Total academic R&D expenditures (0.639), 2) postdoctorates (0.940), and federally financed academic R&D expenditures (0.951). All of the other variables increased r-squared to 0.961. Finally, we ran a stepwise regression including all independent variables, including a variable not previously included in regressions (number of patents issued in 1988 to 2001 for the institution). No change was found in the variables entering the regression equation. Because total and federally-financed R&D expenditures were highly correlated (0.945), and the incremental r-squared obtained by adding federally-financed R&D expenditures was barely more than 0.01, we further simplified the model by restricting it to total academic R&D expenditures and postdoctorates.

D.3 Analyses of Whole Citation Counts in an Expanding Journal Set

A series of regressions were performed to determine, for each set of personnel and financial variables, which were most highly correlated with whole citation counts in an expanding journal set. The data set was aggregated to the institution-year level. Financial variables were deflated using the GDP deflator; personnel variables were lagged by one year and financial variables by

two years. Analyses were conducted using stepwise linear regression. We typically retained any variable that, when entered into the regression, increased r-squared by at least 0.01. All such variables were highly statistically significant when entered.

The set of faculty counts variables (all lagged by 1 year) in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) Full Profs (0.136), 2) Assoc Profs (0.285), and 3) Total Fac (0.313). These are the same set of variables, and order of entry, as for fractional citation counts. When only the these three independent variables were included in the regression, the coefficients were 82.4 for full professors, -131.4 for associate professors, and 56.2 for total faculty. Since addition of an associate professor also increases total faculty, the net effect of adding an associate professor would be -74.2.

The set of Carnegie classifications and type of control in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) R-1 (0.231), 2) Private control (0.254) and 2) Medical (0.264). These are the same set of variables, and order of entry, as for fractional citation counts. The other Carnegie classifications did not enter the regression equation.

The set of academic R&D expenditures variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) Total academic R&D expenditures (0.600), 2) federally financed academic R&D expenditures (0.715), and 2) other academic R&D expenditures (0.772). These are the same set of variables, and order of entry, as for fractional citation counts.

The set of enrollment variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) number of S&E grad students (0.180), 2) fall enrollment of undergraduates (0.284), and fall enrollment of graduate students (0.322). These are the same set of variables, and order of entry, as for fractional citation counts. The coefficient for undergraduates was negative, perhaps reflecting less publication activity in some institutions with more undergraduate students.

Among the set of postdoctoral count variables, only total postdoctorates entered the regression with an incremental r-squared (0.842) in excess of 0.01. This is the same variable found in the regression of fractional citation counts.

The set of non-faculty research staff count variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) non-faculty research staff (0.387), and 2) non-faculty research staff without M.D.s (0.418). These are the same set of variables, and order of entry, as for fractional citation counts.

The set of degrees awarded variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) S&E doctorates from the IPEDS Completions Survey (0.379), 2) S&E BA/BS degrees (0.445), and 3) total S&E degrees (0.046). This is slightly different than obtained from the regression on fractional citation counts, where S&E doctorates from the NSF SED survey was entered first. The two doctoral degree variables have a correlation of 0.973.

The complete set of retained variables identified above was entered into a stepwise regression. The variables in their order of entry into the regression and the cumulative r-squared values after entry were as follows: 1) Total academic R&D expenditures (0.600), and 2)

postdoctorates (0.940). All of the other variables increased r-squared to 0.958. Finally, we ran a stepwise regression including all independent variables, including a variable not previously included in regressions (number of patents issued in 1988 to 2001 for the institution). No other variables increased r-squared by at least 0.01, and the total r-squared with all variables was 0.965.

D.4 Path Analytic Model for Citation Counts

We developed a path analytic model under the assumption that total R&D expenditures have a direct and indirect effect on citations, where the indirect effect is through funding of some postdoctorates. We also assumed that postdoctorates have a direct effect on number of citations. A path analytic model allows total academic R&D expenditures to directly influence the number of postdoctorates, and for postdoctorates and total academic R&D expenditures to directly influence the number of publications.

For fractional citation counts in an expanding journal set, the regression coefficient for total academic R&D expenditures was 43.6 (i.e., 43.6 citations per each $1M in total academic R&D expenditures), and the coefficient for postdoctorates was 41.1 (i.e., 41.1 citations per additional postdoctorate). However, the intercept for this model was negative (-74.5), which was counterintuitive. Consequently, we found it necessary to set the intercept to zero. In the revised regression, the regression coefficient for total academic R&D expenditures was 43.0 and the coefficient for postdoctorates was 41.1. The corresponding standard errors were 1.2 and 0.42, respectively. We then performed a regression where the dependent variable was the number of postdoctorates and the independent variable was total academic R&D expenditures. We find that each $1M additional funding is associated with an increase of 2.53 additional postdoctorates. Thus, adding together the direct and indirect effects, we find that each $1M in additional academic R&D expenditures is expected to result in 147.0 (i.e., 43.0 + 41.1 x 2.53) citation counts. We note however, that since the citation count is the number of citations to publications in any of 3 prior years, the $1M increase must be maintained for three years. The effect of a $1M increase in a single year would only be an increase of 49 (i.e., 147.0 / 3) fractional count citations.

For whole citation counts in an expanding journal set, the regression coefficient for total academic R&D expenditures was 55.7 per $1M and the coefficient for postdoctorates was 74.8. The corresponding standard errors were 2.6 and 0.73, respectively. The coefficient for total academic R&D expenditures regressed on the number of postdoctorates was 2.53 per $1M. Consequently, adding together the direct and indirect effects, we find that each $1M in additional funding for 3 consecutive years is associated with an increase of 244.9 (i.e., 55.7 + 74.8 x 2.53) whole publication counts. The effect of a $1M increase in a single year would only be an increase of 81.6 (i.e., 244.9.0/3) whole count citations.

We note that the number of citations that the model predicts will be generated is more than we would expect given the additional number of publications that are predicted. For each $1M in additional total academic R&D expenditures, the publications model predicts 7.18 fractional count and 11.0 whole count publications. On average there are 5.44 fractional count citations per fractional count publication and 6.19 whole count citations per whole count publication. Therefore, we would expect per $1 M in additional total academic R&D expenditures to generate 39.0 fractional count and 68.1 whole count citations. However, the path model for citations predicts 49.0 fractional count and 81.6 whole count citations. This appears to be a consequence

of the non-constant ratio of citations to publications. For example, whole count publications in 1988 to 1990 each generate approximately 5.48 whole count citations in 1992, whereas whole count publications in 1997 to 1999 each generate approximately 6.60 whole count citations in 2001. Some of this increase may be due to the expansion of the journal set over time; citation counts can increase in an expanding journal set even when publications do not, because the larger database of journals provides more opportunities for citation. Similarly, fractional count publications in 1988 to 1990 each generate approximately 4.95 fractional count citations in 1992, whereas fractional count publications in 1997 to 1999 each generate approximately 5.70 fractional count citations in 2001. The interaction of this change in the citation to publication ratio with the increase in academic R&D expenditures over time (so larger expenditures are associated with higher citation to publication ratios) results in a model that tends to overpredict the slope for the academic R&D expenditures variable, and consequently overpredict the increase in citations per $1M of additional funds.