A Bayesian Measure of Research Productivity

Lin Qin and Steven T. Buccola

Abstract We use Bayesian probability theory to develop a new way of measuring research productivity. The metric accommodates a wide variety of project types and productivity sources and accounts for the contributions of "failed" as well as "successful" investigations. Employing a mean-absolute-deviation loss functional form with this new metric allows decomposition of knowledge gain into an outcome probability shift (mean surprise) and outcome variance reduction (statistical precision), a useful distinction, because projects scoring well on one often score poorly on the other. In an international aquacultural research program, we find laboratory size to moderately boost mean surprise but have no effect on precision, while scientist education improves precision but has no effect on mean surprise. Returns to research scale are decreasing in the size dimension but increasing when size and education are taken together, suggesting the importance of measuring human capital at both the quantitative and qualitative margin.

Introduction

The centrality of research to economic growth begs for rigorous, practical methods of assessing scientific knowledge. Now occupying 2.8% of US gross domestic product, R&D is nearly universally regarded to be essential to continued economic health (Industrial Research Institute 2016). At country, institution, program, and scientist levels, however, research administrators are continually asked to demonstrate R&D's benefits over costs. That would require understanding not only the net

L. Qin
comScore, Inc., Reston, VA, USA

S.T. Buccola (✉)
Oregon State University, Corvallis, OR, USA
e-mail: sbuccola@oregonstate.edu

Fig. 1 Stages of the research and dissemination process

benefits themselves but how they are influenced by research goals, analytical strategies, scientist recruitment, and management.

The cost side of R&D assessment follows much the same protocol as in any other enterprise. The main problem is with the knowledge outputs, which are resistive of statistical simplification, difficult to track once released, and only indirectly observable. Research effort and productivity response, furthermore, begin well before and continue well after the easily observable research activities and include problem inception and funding; observations and testing; writing and presenting; and the scientific, administrative, and industrial uses to which the research will be put. Because each phase represents a certain "production," each can be the object of productivity evaluation. Figure 1 depicts a simplified schema of these production stages, emphasizing applied research.[1] The first is the research project itself, the second its communication, and the third its economic impacts.

Although distinctions among the three are ambiguous in actual situations, they are valuable for understanding the differences among research evaluation methods. Economic surplus methods focus on the economic impact stage, examining the net relationships between research expenditures and subsequent industry productivity. A virtue of the surplus approach is that because expenditures are dual to and therefore reflect research design, management, and communication, they afford an efficient, statistically unbiased focus on the bottom line – economic gains (White and Havlicek 1982; Alston et al. 1995; Huffman and Evenson 2006; Alston et al. 2011; Hurley et al. 2014). A largely separate bibliometric literature concentrates on the communication phase, especially on relationships between research funding and directly observable outputs – patent counts and citations if intellectual property markets are present (Hall et al. 2005; Fontana et al. 2013), publication counts and citations they are if not (Pardey 1989; Adams and Griliches 1996; Oettl 2012). The bibliometric approach offers a more refined view of the transmission of scientific ideas than economic surplus can.

Despite their analytical power in the purposes for which they were designed, neither of these two approaches is positioned to look very far into the laboratory itself. Neither therefore is very suitable for assessing individual projects nor the factors like research topic, laboratory resources, and management policies affecting the research programs under which they are organized. It would be useful then to

---

[1] Basic research cannot as easily be divided into Fig. 1's steps or into any regular steps at all. We note below the important differences between basic research and the applied research that motivates our present approach.

examine the possibility of a research output metric helpful for that purpose. Like a case study, the metric would require insights and data from the principal investigators (Polanyi 1974; Schimmelpfennig and Norton 2003; Shapira et al. 2006). Unlike a case study, it would be expressible in a manner that can be readily compared across projects and programs. The metric must, in particular, be flexible enough to accommodate a variety of project topics, methods, and settings, while suitable for pooling into an econometric model of research outputs and inputs. It must, for example, overcome the problems of distinguishing program from nonprogram influences on research success.

We develop such a metric here by deriving a scientific knowledge measure reflective of individual laboratory, treatment, and control conditions but useful for the administration of heterogeneous applied research projects in a variety of technical and institutional settings. We use the approach to investigate the knowledge returns to an international aquacultural research program involving, over a 4-year span, 55 studies in 16 nations, showing how returns vary by research team characteristics, scale, topic area, analytical approach, and outcome dimension. To be useful for these purposes, the metric must be capable of comparison with the factors hypothesized to influence it. Much of our effort therefore is devoted to accommodating and exploiting cross-study heterogeneity in a research program.

A Direct Approach to Scientific Knowledge Measurement

A knowledge metric accounting adequately for research program discovery would satisfy at least three requirements. It should be (a) ratio-scale comparable across the studies investigated (i.e., contain a meaningful zero point); (b) ex-ante in the sense of conditional on the anticipation of future significant R&D events; and (c) reflective of all new knowledge a study provides, regardless of whether it achieved its most ambitious goals or outperformed an earlier study in some positive respect. The ratio-scale cardinality in criterion (a) assures that statements such as "Project A provided twice as much knowledge as Project B" will be valid. Criterion (b) assures that findings will be evaluated in a way conformable to their usefulness in specific future applications. Criterion (c) assures that research "failures" be counted with potentially the same weight as "successes." A treatment's failure to outperform an old one does not imply efforts have been wasted: the disappointment was potentially valuable in pointing to more fruitful research directions (CGIAR Science Council 2009). Sufficient for satisfying these criteria is that research success should be evaluated in terms of the information it offers about the probabilities of treatment outcomes, expressed as a shifting or narrowing of the outcome probability distribution in the face of alternative settings.

Bayesian reasoning is well-suited to criterion (b) as well as (c) because it considers knowledge in terms of improvements in the predictability of unknown future outcomes (Lindley 1956; Winkler 1986, 1994). A valid experiment or survey can never reduce our forecast ability and generally will improve it. To see this, consider

the prospective user, a fish farmer say, of a clinical study to predict the efficacy of a fish vaccine. We want to know how much the farmer would gain from vaccination decisions or regime based on research forecasts of their effect on fish mortality and, therefore, eventual harvest, as opposed to those that use no research information. To do so we use the Bayesian notion of a loss function: the expected utility of the vaccination regime unsupported by a study from the research program in question, less the expected utility of a regime that does take advantage of the program information.

The negative of that loss is the value K of the research information itself. If we let d be the vaccination decision or regime, Y the mortality or harvest outcome, and Z the study forecast of that outcome, our measure of research knowledge gain K (viz., the expected value of sample information EVSI) therefore is:

$$K \overset{\text{def}}{=} \text{EVSI} = EU\left[d|;Z|;p(Y|Z)\right] - EU\left[d;p(Y|Z)\right] \tag{1}$$

(Winkler 1972, p. 311; Berger 1985, p. 60). Here $p(Y|Z)$ is the probability that outcome Y will occur given that we know its forecast Z. Its presence in both the first and second right-hand term of K indicates only that the utility of research-informed decision $d|Z$ and the research-uninformed decision d is each being evaluated with the use of the research forecast model from which forecasts Z have been drawn. This is valid because the research studies we are investigating have already been conducted, so the forecasts are already known. Including these forecasts in both right-hand terms of the knowledge equation therefore is necessary as well as possible, because it is the only way to compute the expected utility the farmer foregoes by declining to use the research information that the program has made available.

Our principal interest is in determining how the research program's purview, resources, and management policies – including the topics it takes up, the research disciplines involved, its budgets, and the training of its scientists – affect the knowledge it produces. Many of these factors, which we indicate by the vector **X**, are observed at the research project level, such as the research discipline, human and physical research capital, study methods and materials, research treatments, and the type and difficulty of the topic addressed. Others are observed at the program rather than project level, such as the physical environment spanning a number of projects.[2]

We model an applied research program's economic value in terms of its improvement to decision makers' forecast accuracy. Thus, we must adopt a functional form for utility U or equivalently for the loss incurred when outcome Y (such as harvest volume) diverges from its forecast Z. For that purpose, we use the mean absolute deviation (MAD) functional form $U(d, Y) = -|d - Y|$, so that loss is proportionate to the absolute difference between decision, and hence prediction, and outcome

---

[2] Basic research in contrast could be argued to lack any concrete outcomes or probabilities, being a matter more of discrete realizations than incremental steps. Probit models might in future be useful in representing that kind of discrete space.

(Robert 2001). Research utility and knowledge *K* improve, in other words, to the extent forecasts come closer to outcomes, whether overshooting or undershooting them.

In this case, it is not too difficult to show that Eq. (1) should depend to a high degree of accuracy on two terms. The first is the difference the research has made in the scientist's outcome prediction *Z*. That is, it is the difference between the prior forecast M_{prior} and the posterior forecast M_{post}, defined as the shift in the location of the outcome's probability distribution, which we call the study *mean surprise*. The greater the mean surprise, the more knowledge the research has provided. The second is the sample variation of the research outcomes, or in a regression context the standard deviation of the model error σ_{post}, which we will call the study's *imprecision*. The lower the imprecision, the greater the study's knowledge contribution. In sum, our own regression model of a research program's knowledge production is

$$K = g\left(\left|M_{\text{prior}} - M_{\text{post}}\right|, -\sigma_{\text{post}}\right) = f\left(\mathbf{X}, \varepsilon\right). \tag{2}$$

Knowledge contribution is greater to the extent of its mean surprise and lower to the extent of its precision, and this contribution depends on the research program characteristics **X**.

Application to Research Assessment

As an illustration, we apply this framework to a pond fisheries research program funded by the US Agency for International Development (USAID), which during the 2007–2011 period comprised 55 studies managed under seven subprograms in 16 nations.[3] It combined the resources of 17 US universities and 31 foreign universities and institutes. Twenty-five of the studies were completed during the program's 2007–2009 phase, examined here. Data are drawn from a research input and output questionnaire administered to the 25 2007–2009 investigators, plus associated interviews. In each controlled experiment study, an output observation consisted of a pair of prior and posterior probability distributions for each major treatment and for each dimension if the treatment involved multiple outcome dimensions. In each survey study, an output observation consisted of such a probability distribution pair for each major survey question posed. Study expenditure data were obtained from the study proposals and their subsequent quarterly, annual, and final reports. Sample size from the 25 studies was 415.

---

[3] Feed-the-Future Innovation Lab for Collaborative Research on Aquaculture and Fisheries, Oregon State University, sponsored by US Agency for International Development. The countries are Bangladesh, Cambodia, China, Ghana, Guyana, Indonesia, Kenya, Mexico, Nepal, Nicaragua, the Philippines, South Africa, Tanzania, Thailand, Uganda, and Vietnam.

Research Problem Type

Research topic can affect knowledge output because some topics are more easily exploitable than others – requiring fewer scarce resources, of more recent interest and thus fewer scientific competitors, or benefiting from earlier discoveries that enhanced the likelihood of new ones (Alston et al. 2000). Program administrators would have intuitions about which areas will conduce to the greatest study output with a given budget. Ex post however, the productivity of a topic area is empirical and can be determined only by comparing topic research performance when costs are held constant. In a highly diversified program like this aquacultural one, categorizing problem types is difficult because "topic" can refer to a scientific discipline like biology, a subdiscipline like developmental biology, or a problem area like mutation. Topic areas in are here aggregated into four groups: development biology, human health science, economic science, and environmental science.

Research outcome dimensions, too, have implications for research performance because, for example, water microcystin problems may be less familiar and so costlier than water phosphorus problems. At the same time, relative unfamiliarity can bring greater breakthrough opportunities in the sense of shifting the expected outcome away from the literature's current one. Similarly, some outcome dimensions are more easily measured than others and hence more amenable to predictive precision, a disease's immunity rate more predictable, for example, than its duration. A typical study in our data focused on five or six separate outcome dimensions. For parsimony, they are aggregated here into four categories: mortality and growth, demand and price, species diversity, and water quality. Outcome dimensions crosscut topic areas. Developmental biology and environmental science studies, for instance, frequently consisted of mortality/growth and water quality dimensions.

A crucial element of problem difficulty is the analytical approach required to address it. Experiments are usually more expensive than surveys. But the difficulty of managing a controlled experiment depends on the lead investigator's training and experience and the topic at hand. Because experimental controls are designed to reduce random noise, we expect them to bring lower unexplained sample variance, that is, greater precision, than surveys do. On the other hand, science's normal preference for a controlled setting suggests surveys are used only when a problem's conceptual frame is too poorly understood to formulate an incisive experiment. That absence of a strong a priori likely brings large and frequent distribution shifts – a substantial amount of mean surprise – in statistical surveys.

Research Cost

The potentially best indicator of a project's resources is its budget, encapsulating its physical and human capital and material costs. Research budgets in our data often included administrative support and lab space costs but rarely equipment. In either case their service flows to the given project were unreliable because they were based on local

accounting conventions, which vary by country. A reliable measure of project scale, however, was budgeted principal investigator and research assistant time: project FTE.

Researcher quality might be as important as size: education has been widely shown to lift labor productivity. This would be especially true in a knowledge-intensive activity like research (Cohen and Levinthal 1989; Rynes et al. 2001; Schulze and Hoegl 2008). A valuable proxy for unobserved infrastructure can be researcher travel cost because most program study sites were far from the home institution. Travel consumes resources – the scientist's time and energy as well as cash cost – otherwise devotable to analysis. Education and travel time therefore were included in our model in addition to laboratory assistant FTE. Research material and training expenses were unavailable. Table 1 lists the knowledge measures

Table 1 Summary statistics, aquaculture and fisheries research, 2007–2009 (*N* = 415)

Variable	Unit	Mean	Standard deviation	Coefficient of variation
Knowledge generated				
Value of sample information (*K*)	Proportion	0.123	0.212	1.723
Mean surprise (distribution shift)	"	0.224	0.278	1.241
Imprecision (error st. deviation)	"	0.380	0.653	1.718
Laboratory human capital				
Lab size (FTE)	Years	4.8	2.9	0.604
Mean lab education	Years/person	17.1	1.2	0.070
Mean lab age	Years/person	33	6.1	0.185
Research problem type				
Topic area category				
Development biology	Category	0.51	0.50	0.98
Human health science	"	0.12	0.32	2.67
Economic science	"	0.11	0.31	2.82
Environmental science	"	0.27	0.44	1.63
Research outcome dimension				
Mortality and growth	Category	0.68	0.47	0.69
Demand and price	"	0.14	0.37	2.64
Species diversity	"	0.002	0.05	25.00
Water quality	"	0.18	0.39	2.17
Analytical approach				
Experiment vs survey	Category	0.70	0.46	0.657
Public infrastructure				
Site proximity	Kilometers	843	891	1.06
Region of world				
Asia	Category	0.78	0.41	0.53
Africa	"	0.05	0.22	4.40
Latin America	"	0.17	0.37	2.18

and relevant productivity factors for which we have data: laboratory human capital, topic area category, research outcome dimension, analytical approach, and public infrastructure proxies.

Data Construction and Econometric Specification

In sum, research program analysis involves eliciting investigators' prior and posterior density functions and formulating and estimating the associated loss function.

Eliciting Prior and Posterior Densities

The director of each *controlled experiment study* was asked to identify the three most important experimental controls to be used in her research. For each control, and each major outcome dimension of that control, she then was asked to state her prior probabilities that a respectively low, medium, and high outcome level would be observed. These stated probabilities were used to compute that control's and dimension's prior mean M_{prior}. When the experimental results were later obtained for that control and dimension, she was asked to provide its mean outcome M_{post}. The corresponding research precision measure σ_{post} was obtained as the standard deviation of the ANOVA model's residual error. The director of each *survey study* was asked to name the three most important survey questions to be enumerated. For each, we asked him to identify his prior probability that the respondent would give a respectively low, medium, and high answer, giving us the expected survey question outcome M_{prior}. The corresponding precision measure was the residual standard deviation of the investigator's multiple regression analysis of the responses (Winkler 1972).

   At two international program meetings and several other workshops, we trained principal investigators in the process of quantitatively expressing their prior probabilities, together with the kinds of information, such as the scientific literature and earlier experience with related projects, admissible in priors. Among the experimental studies, the 14 investigators each reported an average of 8 treatments and 3 outcome dimensions per treatment. In the statistical survey studies, the 11 investigators reported an average of 4 respondent subgroups and three survey questions per subgroup.

Data

Key sample statistics are shown in Table 1. In the average study, research control, and outcome dimension, the prior outcome expectation was 22.4% greater or less than the mean outcome in the subsequent experiment or survey. That is, research

outcome expectations tended to be 22.4% of what eventually happened, creating a 22.4% mean surprise. Study precision is reflected in the sample mean posterior standard deviation (0.380) in Table 1 – measuring the average spread of unexplained research outcomes around their unity-normalized experimental or survey means. Research utility lost, that is, on account of prediction or estimation noise was 38% of the typical outcome mean. The mean expected value of sample information – our knowledge metric K – was 12.3% of the typical outcome mean.

The knowledge density functions provide a broader picture. The distribution of mean surprises (Figs. 2 and 3) is skewed strongly to the right, most observations lying just above zero. Mean surprise appears to be a tournament: the greater the outcome distribution shift, the fewer that achieve it. Something of a reverse tournament is evidenced in study precision, the worst performances being the least likely. The bulk of investigators, that is, maintain a relatively low error variance. Aggregating mean surprise and precision together, the new knowledge distribution appears to be dominated by its mean-surprise component. Most projects bunch near the low end of the new knowledge range, a phenomenon often noticed in competitive outcomes (Hausman et al. 1984; Griliches 1990; Lanjouw and Schankerman 2004). Logs of mean surprise and precision (Fig. 2) are rather symmetrically distributed. But especially in knowledge K (Fig. 2c), left-tail outliers are evident, representing mean surprises near zero and hence with large negative logs.

As Table 1 shows, the average lab assistant had 17 years of education – about 1 year of postgraduate work. Seventy percent of the projects were controlled experiments and the remaining 30% surveys. Fifty-one percent of topics were in development biology, 27% in environmental science, 12% in human health science, and 11% in economics. Crosscutting these areas, 68% of treatment outcomes were on mortality and growth, 18% on water quality, and about 14% on demand and price. Seventy-eight percent of the studies were in Asia, 17% Latin America, and 5% Africa. Coefficients of variation (CV) of explanatory variables, reflecting adequacy of variation for statistical inference, varied widely. Education's relative variability (CV = 0.07) is the lowest, as expected on a research team. None of these factors were correlated to an extent creating inference problems.

Results

A way to think of basic research is that it is an effort to find a "whole new approach" to the problem in question, shifting the entire probability distribution of predicted outcomes. That is, if successful, it will generate a substantial mean surprise $|M_{\text{prior}} - M_{\text{post}}|$.

One expects to see statistical surveys used most frequently in these situations, when the problem's stochastic structure is poorly understood. Survey approaches, that is, might be expected to bring greater mean surprise than experiments do. Conversely, experimental controls would be used when the structure is better known and so have greater success in achieving research precision.
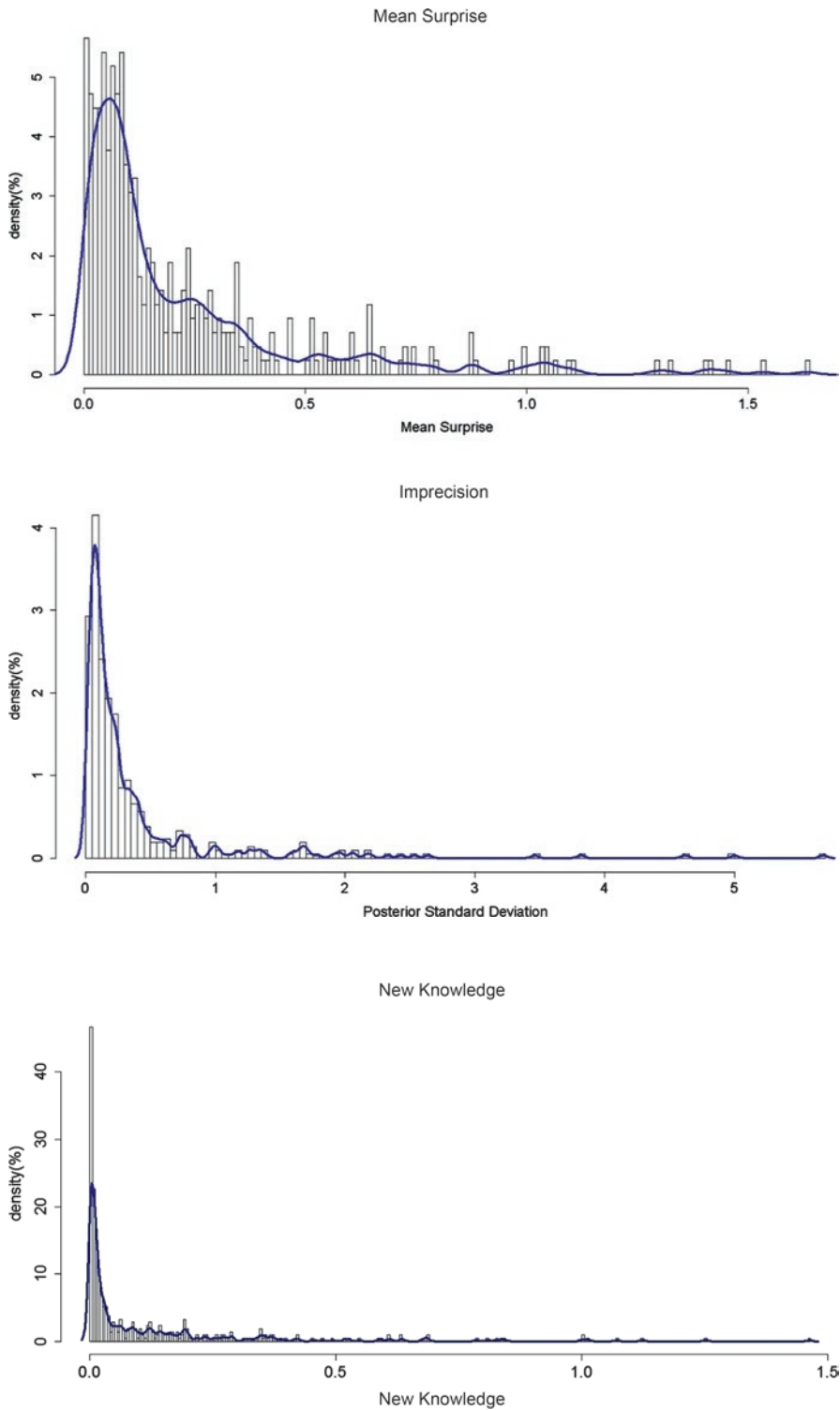
Fig. 2 Histograms of (a) Mean Surprise $|M_{\mathrm{prior}} - M_{\mathrm{post}}|$, Imprecision (posterior standard deviation σ_{post}), and (c) New Knowledge K
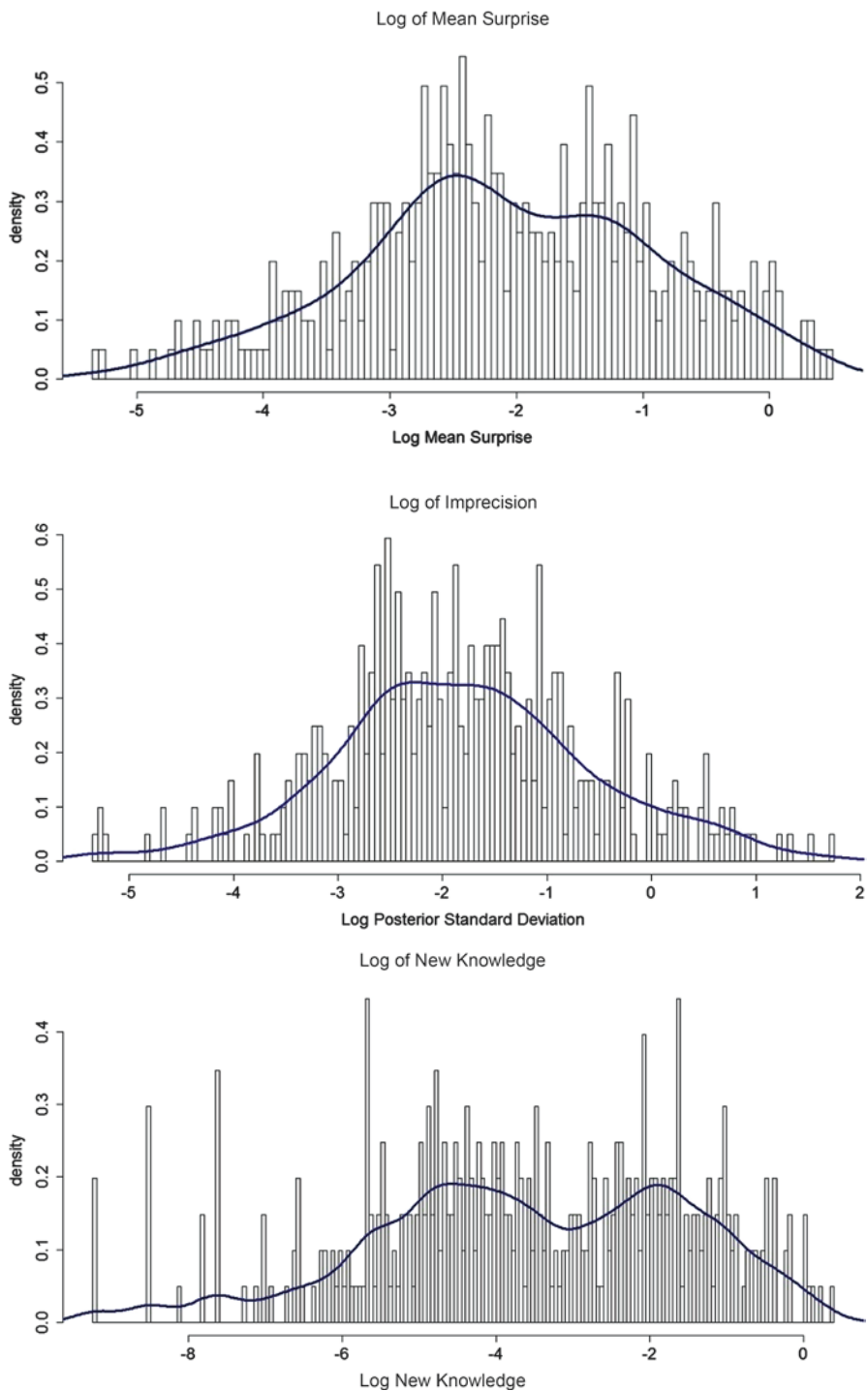
Fig. 3 Histograms of (a) Log Mean Surprise ($\ln |M_{\text{prior}} - M_{\text{post}}|$), (b) Log Imprecision (log posterior standard deviation $\ln \sigma_{\text{post}}$), and (c) Log New Knowledge ($\ln K$)

At the same time, because structural shifts are more poorly observable to the econometrician and hence harder to model than are the more marginal scientific efforts to reduce prediction variance, mean surprise should be more difficult for research evaluators to assess than research precision is. Explaining surprise would require deeper attention to the nature of the subject and its institutional and intellectual constraints, modeled here in only a summary way by analytical approach, topic area, and outcome dimension. Explaining precision therefore would involve greater attention to the more numerate factors like budget, laboratory equipment, and the number, experience, and education of the research assistants.

Factor Effects on Mean Surprise and Predictive Precision

The first step in our own evaluation of the USAID program is to regress mean surprise on the associated project inputs and program features. We then separately regress the study outcome residual (i.e., unexplained) variances on these same project inputs and program features. The results are shown in Tables 2 and 3. Several factors not included in these tables – research assistant numbers and age and travel mode to study site – were consistently nonsignificant in earlier regressions and dropped. R-squares, 0.26 in the mean surprise and 0.44 in the research precision model, are reasonably high considering the sample's cross-sectional structure and the variety of research problems, methods, treatments, outcome types, institutional settings, and cultural settings it contains.

Topic area category and research outcome dimension each help explain mean surprise (Table 2). Developmental biology studies bring 0.47 [i.e., 0.266 – (−0.204)] percentage points more mean surprise than economics studies do, a difference 2.24 times as large as the average (0.21) of the associated standard errors. Human health science work similarly brings 0.35% points more mean surprise than economics work does, 1.45 times the average (0.24) standard error. Mean-surprise differences between the remaining topic-area pairs are small. Among outcome dimensions, demand and species diversity each bring significantly more mean surprise than water quality does (*t*-statistics 1.9 and 1.6, respectively).

Topic area and outcome dimension have even more striking influences on research precision (Table 3) than they do on mean surprise. Predictive precision averages 48% higher in development biology, 48% lower in human health, and 73% lower in economics than in environmental science. The low associated standard errors make clear that the precision differences among biology, human health, and economic science are also statistically significant at normal confidence levels. Among outcome dimensions, mortality/growth predictions carry 78% greater precision, and demand/price predictions 67% greater precision, than do water quality predictions. Separately, and consistent with our hypothesis, statistical surveys produce the greater probability distribution shifts, and controlled experiments produce the more precise predictions. All else equal, surveys bring an average 105% greater mean surprise than controlled experiments do (Table 2) and experiment an average 75% lower posterior outcome error than surveys do (Table 3).

Table 2 Input effects on absolute difference between prior and posterior mean finding (mean surprise)

Research input	Estimate	Standard error
Intercept	0.171	3.790
Continuous inputs		
Lab size (FTE)	0.206	0.097
Mean education	−0.353	1.333
Site proximity	0.107	0.025
Topic area category		
Development biology	0.226	0.196
Human health science	0.143	0.252
Economic science	−0.204	0.232
(Base: Environmental science)		
Research outcome dimension		
Mortality and growth	−0.087	0.200
Demand and price	0.568	0.300
Species diversity	1.836	1.140
(Base: Water quality)		
Analytical approach		
Experiment vs survey	−1.051	0.248
(Base: Statistical surveys)		
Region of world		
Asia	−0.667	0.163
Africa	0.576	0.303
(Base: Latin America)		
Notes:		
Dependent variable:	Absolute difference between prior expectation and posterior sample mean of experimental finding or survey response. Estimation linear in logs	
Residual standard error:	1.091	
Sample size:	415	
Multiple *R*-square:	0.26	

Laboratory size (proxied by employment) does turn out to significantly boost mean surprise, lifting it 0.21% for every 1.0% lab size increment (Table 2). It has no significant effect however on predictive precision (Table 3).[4] Study site proximity to the research center enhances mean surprise only moderately (elasticity 0.11 in Table 2) and statistical precision even less.

---

[4] If laboratory expansion did impair precision, we would be unlikely to observe any expansion unless the mean-surprise advantage more than compensated for the precision loss.

Table 3 Input effects on negative of standard deviation of unexplained finding (precision)

Research input	Estimate	Standard error
Intercept	−11.543	3.302
Continuous inputs		
Lab size (FTE)	−0.112	0.085
Mean education	4.263	1.161
Site proximity	0.033	0.021
Topic area category		
Development biology	0.481	0.171
Human health science	−0.476	0.219
Economic science	−0.735	0.202
(Base: Environmental science)		
Research outcome dimension		
Mortality and growth	0.797	0.174
Demand and price	0.670	0.261
Species diversity	1.614	0.993
(Base: Water quality)		
Analytical approach		
Experiment vs survey	0.752	0.216
(Base: Statistical surveys)		
Region of world		
Asia	0.370	0.142
Africa	−0.516	0.264
(Base: Latin America)		
Notes:		
Dependent variable:	Negative of standard deviation of unexplained experimental finding or survey response. Estimation linear in logs	
Residual standard error:	0.9508	
Sample size:	415	
Multiple R-square:	0.44	

Implications for Knowledge Production

The second step in our evaluation is to decompose total knowledge production into its mean surprise and unexplained error variance components. Our regression fit to the 415 sample observations ($R^2 = 0.97$) is, with t-statistics in parentheses:

$$\ln K = -1.188 + 1.76 \ln \left| M_{prior} - M_{post} \right| + \left(-0.71 \ln STD_{post} \right)$$
$$\left(-30.97 \right) \left(122.08 \right) \quad \left(-50.79 \right) \tag{3}$$

Three conclusions can be drawn from this estimate. (a) The high R^2 suggests mean surprise and unexplained posterior variance together virtually exhaust the knowledge generated. In other words, EVSI in a MAD functional form is very successfully decomposed into mean surprise and statistical precision. (b) Surprise and

Table 4 Decomposition of input effects on new knowledge, continuous variables

Research input	Knowledge contribution via research mean surprise	Knowledge contribution via research precision	Total knowledge contribution
Lab size (FTE)	0.363	0	0.363
Mean education	0	3.027	3.027
Site proximity	0.189	0.023	0.212

Notes:

Contributions in the first column are Table 2 elasticities multiplied by mean surprise's marginal knowledge contribution, and in the second column, the Table 3 elasticities multiplied by precision's marginal knowledge contribution

Third column numbers are the sums of those in the first and second columns

Continuous input effects are percentage changes induced by a 1% change in the indicated input

Analytical approach effects are the percentage changes associated with switching from the reference group to the group indicated

precision are each positive contributors to new knowledge *K*. (c) Surprise does turn out, in both an elasticity and goodness-of-fit sense, to be the more powerful knowledge factor, and precision the less powerful: the mean-surprise knowledge elasticity is 1.76/0.71 = 2.5 times greater than precision.

We now use the above regression weights 1.76 and 0.71 in conjunction with the research output elasticities in Tables 2 and 3 to compute in Table 4 how each selected research input affects new knowledge output *K*. The contribution of lab size to scientific knowledge is, by way of its effect on mean surprise, (1.76) (0.206) = 0.36% and, by way of its effect on research precision, statistically nonsignificant. Scaling-up lab size 1% thus lifts knowledge output by 0.36%. Decreasing returns to lab scale are evident; the average project's ability to produce new findings with its observable physical resources is tightly constrained by, presumably, not only unaccounted for missing inputs but breakthrough opportunities in the research field.

On the other hand, research scale economies can be considered in a qualitative as well as physical or quantitative direction. In particular, we might want to know how knowledge output is affected by a simultaneous expansion of research lab size and quality. An elasticity at the combined quantitative and qualitative margins can, to the degree that quality is reflected in formal training, be obtained by adding the elasticity with respect to size together with the elasticity with respect to education. We have found that team education has essentially no mean-surprise effect (Table 2), although a precision elasticity of 4.263 (Table 3). Weighting the latter by precision's 0.71 knowledge, weight in Eq. (3) says a 1.0% education improvement boosts knowledge production by a very strong 3.027% (Table 4). Combining this with the lab size elasticity discussed above implies that expanding research capacity, 1.0% in both quantitative and qualitative dimensions lifts knowledge output by 0.363 + 3.027 = 3.39%. That is, taking input quality as well as quantity into account, increasing rather than decreasing returns to research scale is evident. Kocher et al. (2006) and Wang and Huang (2007) also find increasing returns to research scale, although with bibliometric methods and in situations much different than examined here.

Conclusions

We have outlined a method of estimating research productivity at program, project, and scientific control level, permitting, in turn, direct comparisons with the associated research inputs and costs. New knowledge is modeled as the Bayesian expected value of sample information. The mean absolute deviation (MAD) utility form used here for that metric enables decomposing knowledge into mean surprise (outcome probability density shift) and research precision (density compactness), permitting independent examination of how each moment is influenced by the research settings.

In an application to an international research program, we find that (i) mean surprise and precision explain nearly the entire variation in research productivity, surprise more so than precision; (ii) greater laboratory size brings decreasing scale returns in the mean-surprise dimension and insignificant returns in the precision dimension; and (iii) researcher education powerfully improves precision, to the extent that, if expanded along with laboratory size, it brings increasing returns to scale in aggregate scientific knowledge. Furthermore, gains at the qualitative margin are much greater than at the quantitative margin.

Despite efforts to quantify the sources of research productivity, many lie beneath the surface even in as comparatively detailed a model as the present one. We have been able to match treatment- and dimension-specific research outcome statistics, hidden to most outside viewers, to many of the factors affecting them. But, for instance, we have not controlled for a research assistant's allocation across treatments and trials, which would affect the number of trials per treatment and the quality of effort per trial and hence research productivity. Although our model suggests they might be valuable, program accountants rarely record that kind of data.

Just as importantly, the present work points to the advantage of requiring scientists to specify their quantitative expectations of research outcomes. Priors have three virtues in a proposal. They encapsulate intuitions about previous work and about the scientist's own ideas, resources, and objectives that will differ from it. They require the proposer to specify precisely what the study controls or treatments are expected to be. And finally, they give managers and funders a more precise basis for judging the study's eventual success.

References

Adams, J., and Z. Griliches. 1996. Measuring Science: An Exploration. *Proceedings of the National Academy of Science* 93: 12664–12670.

Alston, J.M., G.W. Norton, and P.G. Pardey. 1995. *Science Under Scarcity: Principles and Practices for Agricultural Research Evaluation and Priority Setting*. Ithaca: Cornell University Press and ISNAR.

Alston, J.M., M.C. Marra, P.G. Pardey, and T.J. Wyatt. 2000. Research Returns Redux: A Meta-Analysis of the Returns to Agricultural R&D. *Australian Journal of Agricultural and Resource Economics* 44 (2): 185–215.

Alston, J.M., M.A. Andersen, J.S. James, and P.G. Pardey. 2011. The Economic Returns to US Public Agricultural Research. *American Journal of Agricultural Economics* 93 (5): 1257–1277.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer.

CGIAR Science Council. 2009. *Defining and Refining Good Practice in Ex-Post Impact Assessment – Synthesis Report*. Rome: CGIAR Science Council Secretariat.

Cohen, W.M., and D.A. Levinthal. 1989. Innovation and Learning: The Two Faces of R&D. *Economic Journal* 99: 569–596.

Fontana, R., A. Nuvolari, H. Shimizu, and A. Vezzulli. 2013. Reassessing Patent Propensity: Evidence from a Dataset of R&D Awards, 1977–2004. *Research Policy* 42 (10): 1780–1792.

Griliches, Z. 1990. Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature* 27: 1661–1707.

Hall, B.H., A.B. Jaffe, and M. Trajtenberg. 2005. Market Value and Patent Citations. *RAND Journal of Economics* 36: 16–38.

Hausman, J.A., B.H. Hall, and Z. Griliches. 1984. Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica* 52 (4): 909–938.

Huffman, W.E., and R.E. Evenson. 2006. Do Formula or Competitive Grant Funds Have Greater Impacts on State Agricultural Productivity? *American Journal of Agricultural Economics* 88 (4): 783–798.

Hurley, T.M., X. Rao, and P.G. Pardey. 2014. Re-Examining the Reported Rates of Return to Food and Agricultural Research and Development. *American Journal of Agricultural Economics* 96 (5): 1492–1504.

Industrial Research Institute. 2016. 2016 Global Funding Forecast. *R&D Magazine* (Winter 2016 Supplement), 3–34.

Kocher, M.G., M. Luptacik, and M. Sutter. 2006. Measuring Productivity of Research in Economics: A Cross-Country Study Using DEA. *Socio-Economic Planning Sciences* 40: 314–332.

Lanjouw, J.O., and M. Schankerman. 2004. Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators. *Economic Journal* 114 (495): 441–465.

Lindley, D.V. 1956. On a Measure of the Information Provided by an Experiment. *Annals of Mathematical Statistics* 27: 986–1005.

Oettl, A. 2012. Reconceptualizing Stars: Scientist Helpfulness and Peer Performance. *Management Science* 58 (6): 1122–1140.

Pardey, P.G. 1989. The Agricultural Knowledge Production Function: An Empirical Look. *The Review of Economics and Statistics* 71: 453–461.

Polanyi, M. 1974. *Personal Knowledge*. Chicago: University of Chicago Press.

Robert, C.P. 2001. *The Bayesian Choice*. 2nd ed. New York: Springer.

Rynes, S.L., J.M. Bartunek, and R.L. Daft. 2001. Across the Great Divide: Knowledge Creation and Transfer between Practitioners and Academics. *Academy of Management Journal* 44 (2): 340–355.

Schimmelpfennig, D.E., and G.W. Norton. 2003. What is the Value of Agricultural Economics Research? *American Journal of Agricultural Economics* 85: 81–94.

Schulze, A., and M. Hoegl. 2008. Organizational Knowledge Creation and the Generation of New Product Ideas: A Behavioral Approach. *Research Policy* 37 (10): 1742–1750.

Shapira, P., J. Youtie, K. Yogeesvaran, and Z. Jaafar. 2006. Knowledge Economy Measurement: Methods, Results, and Insights from the Malaysian Knowledge Content Study. *Research Policy* 35 (10): 1522–1537.

Wang, E.C., and W. Huang. 2007. Relative Efficiency of R&D Activities: A Cross-Country Study Accounting for Environmental Factors in the DEA Approach. *Research Policy* 36: 260–273.

White, F.C., and J. Havlicek. 1982. Optimal Expenditures for Agricultural Research and Extension: Implications of Underfunding. *American Journal of Agricultural Economics* 64 (1): 47–55.

Winkler, R.L. 1972. *An Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart and Winston.

———. 1986. Expert Resolution. *Management Science* 32 (3): 298–303.

———. 1994. Evaluating Probabilities: Asymmetric Scoring Rules. *Management Science* 40 (11): 1395–1405.