Scientometrics (2010) 84:717–734
DOI 10.1007/s11192-010-0210-z

Assessing publication performance of research units: extensions through operational research and economic techniques

Gregory John Lee

Received: 17 September 2009 / Published online: 20 March 2010

Abstract Many quantitative measures exist to assess the publishing outputs of research units such as university departments or institutes. In addition to well-known issues with such measures, further shortcomings include inadequate adjustments for relative entity sizes and researcher intensity, the extent to which research is concentrated among a few rather than all researchers and lags between staffing and publication. This article presents a further array of possible measurement indices, based on operational research and economic ratios, which are capable of adjusting for each of these shortcomings, and which analysts can combine with relatively little effort into existing measures.

Keywords Research measures · Research units · Bibliometrics · Concentration · Lag in research

Higher education research and management has long been interested in assessments of research unit (RU) publishing (e.g. Alewell 1990; Baird 1991; Colman et al. 1995; Crewe 1988; Krampen 2008; Nederhof et al. 1993; Okrasa 1987; Schloegl et al. 2003; Tan 1986). By RUs, this article means those constituent parts of higher education institutions such as university departments, institutes and the like that are seen as discrete research-producing bodies by some utility of aggregating and measuring their output as a group. This article expands common quantitative indices for measuring RU publishing, taking into account various shortcomings in extant measures.

In research-focused institutions, many reasons exist to assess output at the RU level. Publishing is a core activity of RUs. It is generally at the level of RUs that common processes, structures and systems such as teaching loads and leadership, organization culture and disciplinary commonalities exist, all of which can have marked effects on output (Baird 1991; Dundar and Lewis 1998; Ramsden 1994). For this reason, institutions

G. J. Lee
University of the Witwatersrand, Johannesburg, South Africa

G. J. Lee (✉)
New Commerce Building, Private Bag x3, Wits 2050, South Africa
e-mail: gregory.lee@wits.ac.za

generally place accountability for research on the heads and staff bodies of RUs, and funding bodies often award grants or subsidies to RUs. Researchers also frequently find RU size to be a strong predictor of research, among other variables (e.g. Dundar and Lewis 1998).

Research assessments at the RU level have particular use in the management of institutions (Baird 1991), such as (a) Tracking trends over time in research improvement or decline for the purposes of remedial or bolstering action; (b) Enabling diagnosis of the enabling or inhibiting factors influencing research trends; (c) Guiding RUs in possible reinforcement or remedial actions in perpetuating stronger research outputs; (d) Guiding the allocation of budgets, remedial interventions such as training budgets, and the like, (e) Adherence to overriding requirements, preferences and policies of governmental higher education bodies (Donovan 2005), and (f) Benchmarking to similar RUs in other institutions. It is also possible to compare between RUs in institutions, however this is potentially problematic for many reasons ranging from disciplinary differences in publishing norms to difficulties in comparing inputs (for example, comparing minimalist investment in research resources in a humanities RU versus multi-million Euro investments in a science lab). However, within bounds, comparisons may help in identifying lessons from successful RUs, which managers may feasibly transplant to the research process in other RUs—comparisons need not be competitive and negative but can have developmental uses.

The aforementioned reasons are normative, to do with management of institutions or of policy. There is also the proliferation of positivist research seeking to explain research output (e.g. Baird 1991; Dundar and Lewis 1998; Ramsden 1994). This article, which creates more refined and complex measures and ratios, might be of some use to such studies in providing alternative predictors and control variables.

This article proceeds as follows: "Basic quantitative individual and RU-level measures" section summarizes types of common quantitative research measures, and briefly discusses known issues and common remedies. "Further shortcomings of basic quantitative measures" section discusses further problems and complications with such assessments, which this article addresses. "The basic material of research indices" section explicates simple constituent parts of and inputs to RU-level research measures. "Extensions of research assessments" section describes a variety of RU-level research measures that may serve to ameliorate the shortcomings defined previously.

Basic quantitative individual and RU-level measures

Researchers have presented many basic measures, applied at the different levels of the individual researcher, the RU or program or the overall institution (e.g. Patrick and Stanley 1998; Ramsden 1994; Toutkoushian et al. 2003). This article focuses specifically on quantitative measures, specifically those that present some form of at least quasi-continuous measure such as publication counts, reputational scoring, citation counts, or the like. Ordinal rankings of RUs are also common, especially within disciples, but I will not focus on ranks because they not lend themselves to the extensions discussed here. However, continuous-type data do often underlay ranks.

Quantitative measures generally manifest in one of several types. The most common is probably research *quantity*, notably publication counts (Martin 1996; Tan 1986; Toutkoushian et al. 2003). Citation counts are a common measure of *impact* (Verbeek et al. 2002).

Popular combinations exist, such as the h-index, which melds productivity with impact (Hirsch 2005). *Quality* or *importance* generally requires peer judgments such as more subjective 'reputational' assessments regarding the perceived overall quality of RUs (Patrick and Stanley 1998; Tan 1986). *Input* measures also exist, such as the value of research grants and other measurable funding granted to an RU (e.g. Scherer and Ross 1990). Funding measures differ as to whether they relate to other measurable research outputs. Ex ante grants or subsidies are given prior to any research outputs and may not require such outputs, ex post funding is given for prior published research outputs, as is often the case with government subsidies given to institutions (e.g. Donovan 2005; Tijssen 2007). Researchers have also constructed broader indices incorporating a variety of 'research activities', such as the Index of Research Activity (Ramsden 1994) which awards points for activities such as receiving grants, supervising higher degree students, participating in research groups, and conference papers.

Notwithstanding the variety, this article does not presuppose any particular approach to research assessment at the basic level, other than the quantitative nature of the final data. Basic research assessment will refer to any quantitative index that researchers or institutional policy makers see as an acceptable and common basic measure of research, and accordingly adopt.

A plethora of well-discussed problems exists with any of these quantitative measures (Martin 1996; Research Evaluation & Policy Project 2005; Tan 1986; Verbeek et al. 2002), which this article cannot fully explicate for reasons of space. For instance, publication counts may exclude sources (e.g. books or conference publications), and problems arise in how to weight publication types. Counts also do not always adjust for multiple authors, and do not necessarily reflect publication quality. Citations are prone to skew, and depend on the discipline at hand, making cross-field comparisons difficult without some sort of normalization. Reputational measures obviously suffer from subjectivity, and can both contain, create and embed institutional bias (Tan 1986). Input measures such as funding or composite activity indices do not guarantee output, and in the case of composite indices, the analyst must make somewhat arbitrary choices about what is included and how each activity is weighted.

Notwithstanding these and other issues with traditional measures, various partially effective solutions exist (Research Evaluation & Policy Project 2005; Verbeek et al. 2002). For instance, researchers urge the use of multiple criteria, balancing peer review with multiple bibliometric measures (Martin 1996). Analysts can consider multiple authorships by pro-rating for authorship, and it is increasingly possible to weight publications for impact or quality, especially journal articles (which analysts can increasingly weight or normalize by direct citations, length, impact factor of the journal and the like; or by use of the OR Data Envelopment Analysis technique (Johnes and Johnes 1992)). Less impact or quality data exist for books and the like, but requiring peer review of such outputs as books (e.g. Lewinson 2001) allows at least for dichotomous quality control. Researchers can normalize measures such as citations to account for disciplinary differences (Research Evaluation & Policy Project 2005; Verbeek et al. 2002). Such systems are certainly not perfect, but they satisfy perhaps the worst criticisms. Objectively, quantitative measures are widely utilized despite their potential shortcomings, even without the use of mitigating measures.

This article is not concerned with these well-known problems or solutions, but rather with less discussed, albeit prevalent, issues, as discussed in the following section.

Further shortcomings of basic quantitative measures

As a base-level measure, quantitative research assessments have several further shortcomings with which researchers rarely deal:

1. First, indices often fail to take into account the relative sizes of different RUs. However, all else equal, bigger RUs produce more research (e.g. Dundar and Lewis 1998). Although many researchers normalize for size e.g. (Crewe 1988; Qurashi 1993; Schloegl et al. 2003; Tan 1986), by no means do all do so.
2. Second, in addition to gross size, some RUs possess staff bodies with greater levels of research capacity than others. This might occur naturally (e.g. where a discipline is inherently more research oriented, such as an institute, or the RU does no teaching), or be a result of staffing differences. Research intensity takes several forms: (a) Proportion of research staff (as opposed to pure administrative or teaching staff); (b) Seniority of research staff (institutions generally expect more senior research staff to produce more); (c) Greater experience of research staff, as indicated by lifetime experience, tenure, and the like; (d) Interaction with teaching load, as when RUs with similar research staffing in all other respects place unequal teaching requirements on those staff, leading to variegated abilities to produce research. Very few publication assessments include any such adjustments.
3. Third, basic indices generally do not account for distribution/dispersion of research among researchers (i.e. whether a narrow core of productive members produces research, with many other unproductive colleagues, or whether everyone is contributing according to expectation). Although it has long been acknowledged that research is heavily skewed (Lotka 1926), only a small literature has investigated how to deal with dispersion e.g. (Allison 1980; Crewe 1988; Okrasa 1987), and this theory has focused on distributional qualities rather than practical integration with indices. Very few empirical investigations have integrated concentration explicitly. The concluding sections further discuss the prior work in this area in juxtaposition to the current paper.
4. Fourth, basic indices do not account for the time lag between hiring talent and their ability to produce research in published format (Carpenter et al. 1988).

Accordingly, the remainder of this article suggests alternative indices reflecting the issues of size, research intensity, concentration, and time lag. Underlying this research is the proposition that creating additional adjustments allows for a wider and broader canon of measures, fulfilling the call for a multi-faceted view of research assessment (Alewell 1990; Krampen 2008; Martin 1996).

The basic material of research indices

Research indices designed to assess the research output of RUs are most commonly and simply based on (a) the research outputs of the RUs, generated by (b) a staff complement composed of multiple staff members of differing ranks and contract types, and (c) generated in a given time period. As regards time, I stipulate no particular period; the particular need of the relevant basic measure will dictate. Research on 'citation windows' suggests timeframes of some years (Research Evaluation & Policy Project 2005), however institutional managers more often focus on shorter periods such as one year.

The following sections define and denote concepts, following the convention of using uppercase notation for larger units and lower case for smaller, e.g. RUs versus individuals.

Staff complement measures

There are two ways to enumerate staffing complement. The first is by numerical quantities (and possibly qualities) of personnel. The second is by staffing budget.

Basic numerical staff complement measures

Let $k = 1\ldots K$ represent all RUs in an institution in a stipulated time period, and the total number of staff (of all employee types) in RU k be n_k. Let i_k represent individual i in RU k so that $i_k = 1\ldots n_k$. For simplicity, I drop the RU subscript k when referring to a single RU and I will denote time only when working across multiple periods.

Since complex research measures may wish to distinguish between characteristics of RU members (specifically research intensity), the article distinguishes each staff member i on at least two characteristics. The first is *contract intensity or* "*Full Time Equivalent*" *(FTE) status*, where a full-time contract receives a score of 1 and a contract of lesser intensity is pro-rated downwards (e.g. a 50% or half-day contract receives an FTE score of .5). Let the total FTE score for the RU be F, and that for person i be denoted as f_i where $(0 < f_i \leq 1)$. Note at this stage that a full-time administrator or teacher with no research expectations still receives a score of $f_i = 1$.

The second distinguishing feature is *research position weighting*, based on whether the person is in a research post, and if so, seniority. The simplest version is simply to count each research-related post (i.e. posts that include publication as an expectation) as a score of 1 regardless of seniority, and each non-researcher as a 0. However, since many institutions are likely to expect more research output from higher-ranked staff, a more finely graded system is preferred. I will refer to this as the "Base Researcher Equivalent" (BRE, or R) points system because a given 'base' position is graded with a score of 1, and other posts are graded with a score greater or less than 1 in accordance with the base post. For example, at the author's institution, a senior lecturer has a score of 1 and other posts are scored in balance as follows: Non-researcher = 0, Associate lecturer = .60, Lecturer = .80, Associate Professor = 1.1, Full Professor = 1.3. Other institutions may prefer a different weighting. Let a unit's BRE be R, and for person i the BRE be r_i where ($r_i = 0$, 1, other scores relating to 1 as per institutional choice).

To distinguish further between researchers and non-researchers in the total staff complement, let us always rank total staff complement over a given time period in terms of research assessment. The top researcher is accordingly $i = 1$, and the n staff are therefore divided into a group of researchers $i = 1\ldots m$ and a further group of non-researchers $i = (m + 1)\ldots n$ and where $1 \leq m \leq n$. Given this, the total staff complement in terms of all staff (FTEs, or F) and researchers (BREs, or R) is as follows:

$$\text{Total FTE complement of RU } = F = \sum_{i=1}^{n} f_i \tag{1}$$

$$\text{Total BRE complement of RU } = R = \sum_{i=1}^{n} f_i r_i = \sum_{i=1}^{m} f_i r_i \tag{2}$$

Budgetary staff complement measures

In addition, the staffing budget of a RU is an alternate measure of staff complement. There are two ways of conceiving of staffing budget, first aggregating from individual staff member salaries, secondly a standardized salary per unit. If staff member i is paid salary s_i then:

$$\text{Total salary bill of RU } = S_T = \sum_{i=1}^{n} s_i \tag{3}$$

$$\text{Salary bill of RU for research staff only } = S_R = \sum_{i=1}^{m} s_i \tag{4}$$

$$\text{Standardized salary bill per-FTE } = \frac{S_T}{F} = \frac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} f_i} \tag{5}$$

$$\text{Standardized salary bill per-BRE } = \frac{S_R}{R} = \frac{\sum_{i=1}^{m} s_i}{\sum_{i=1}^{m} f_i r_i} \tag{6}$$

Often RUs will know the *overall* wage bill as well as FTE and BRE staffing complements (quantities 1, 2 and 3) without necessarily knowing breakdowns by individual. This allows them at least to calculate Quantity 5, if they know Input 4 overall they can easily calculate Input 6.

Research assessment of staff members and associated RUs

As discussed in the first sections of this article, the indispensible inputs to quantitative research measures is some basic research assessment that policy-makers agreed to be relevant at and across the RU level. As discussed there, this article does not presume to dictate any specific type of quantitative index, nor specific considerations that should enter into their calculation. Naturally, such a task lies at the level of the institution concerned. The key thing is that analysts construct and apply a consistent measure.

Quantitative research indices

The following are basic quantitative inputs used in the article. First is the total RU-level 'raw' research assessment. During given period t, let RU k produce total research assessment score P_k under whatever measurement system the analysts have chosen (publication counts, citations, etc.). I will refer to this base-assessment hereafter as the 'P'-score.

Second, some of the extended indices in this paper require individual level measures. In given period t, let each individual i produce a body of publications or activities, of the total value of p_i points (e.g. individual publication counts). If researchers have individual data then:

$$\text{Total raw p-score (research assessment)} = P_k = \sum_{i=1}^{m} p_i \tag{7}$$

Given these inputs, the following section presents specific measures and ratios designed to account for issues such as research intensity, concentration and lag.

Extensions of research assessments

This section is the heart of this article: the generation and discussion of a range of possible research assessments in and between RUs, with attempts to account for all the shortcomings discussed earlier of basic research assessment P_k (the "P-score").

Indices adjusted for size of RU only

Adjustments for the size of an RU are improvements on gross publication measures, and are indeed not infrequently used (e.g. Crewe 1988; Qurashi 1993; Schloegl et al. 2003; Tan 1986). Since the FTE score of the RU is the most basic and gross indicator of total staff size, a 'P per FTE' (PPF) ratio accounts for size of the RU:

$$\mathrm{PPF} = \frac{\mathrm{P}_{\text{all staff}}}{\mathrm{FTEs}_{\text{all staff}}} = \frac{P}{F} = \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} f_i} \tag{8}$$

Another option is to calculate a 'P per Total Staffing Budget' (PPTB):

$$\mathrm{PPTB} = \frac{\mathrm{P}_{\text{all staff}}}{\text{Wage bill}_{\text{all staff}}} = \frac{P}{S_T} = \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} s_i} \tag{9}$$

However, the PPF and PPTB ratios are not particularly useful in themselves, since they only count publications per total staff unit with little accounting for research status or intensity of the staff members. The PPTB is better in that it accounts for level/seniority and even experience of staff, because institutions are likely to pay higher level and more experienced staff more. However, it does not entirely distinguish between research and non-research staff.

Indices adjusted for size and research intensity

Because RUs have different proportions of researcher staff versus non-research staff, as well as differing levels of seniority among research staff, measures should incorporate these elements. The most basic and accessible measure of this is 'P per BRE' (PPR):

$$\mathrm{PPR} = \frac{\mathrm{P}_{\text{all staff}}}{\mathrm{BREs}_{\text{all staff}}} = \frac{P}{R} = \frac{\sum_{i=1}^{m} p_i}{\sum_{i=1}^{m} f_i r_i} \tag{10}$$

Since BREs account for research staff only, and contain weights for seniority, they account for the majority of research intensity. The PPR should probably form the basis of most research measurement within institutions—it covers perhaps the most pervasive issues of relative research size and intensity of staffing, researchers can easily calculate it from readily available data, and it is amenable to adaption for further elaborations as discussed in the following sections.

As stated in "The basic material of research indices" section , however, BREs might not account for experience of those staff. The following ratios may help to account for staff experience as well.

Indices adjusted for size, research intensity, and experience

If the data are available, the analyst can weight ratios such as the PPR for the relative experience of each researcher. I will refer to this experience weighting generally as e_i, and it can take any sensible form. Options for e_i might include length of time in the position, length of time since first entering an academic career or the institution, some weighting for amount of education, a weight for number of prior publications or any other scaled experience factor. In this case, the re-weighted ‘Experience-Weighted P Per BRE’ (EW-PPR) is

$$\text{EW-PPR} = \frac{\mathrm{P}_{\text{all staff}}}{\text{Experience weighted BREs}_{\text{all staff}}} = \frac{\sum_{i=1}^{m} p_i}{\sum_{i=1}^{m} f_i r_i e_i} \tag{11}$$

However, one might reasonably argue that some experience-related information might be hard to gather (although institutions do have easy access via HR databases to ages and tenures). Another possible option is a simple variant of the staff budget score, here dubbed the ‘P per currency unit of BRE Staffing Budget’ (PPRB):

$$\text{PPRB} = \frac{\mathrm{P}_{\text{all staff}}}{\text{Wage bill}_{\text{Research staff only}}} = \frac{\sum_{i=1}^{m} p_i}{\sum_{i=1}^{m} s_i} \tag{12}$$

The PPRB assumes that, *ceteris paribus*, pay is higher for more experienced researchers, which is reasonable in a grade and seniority compensation system, through initial negotiation, or via incentives. In even a rough case in which this is true, this ratio reflects experience.

Adjustments for teaching and administrative loads

In addition, an issue highlighted in “The basic material of research indices” section is teaching load. Related to this is administration, which I assume is directly proportional to the student size of the RU, although this might not always be true. Treatment of teaching/administration loads is a thorny issue, given both difficulties of measurement as discussed in the following paragraphs, and also because of the uncertain relationship between teaching and research, i.e. do teaching and research reinforce each other, work in opposing directions, or some more complex middle territory (Patrick and Stanley 1998; Research Evaluation & Policy Project 2005)?

This article will not attempt any concrete answer. Instead, I argue that in this article’s context of normative measurement possibilities, the question is moot. The use of teaching adjustments is subject to the user’s beliefs regarding this question, and can be ignored or used. Second, an analyst can use these adjustments in a positivist manner to ascertain whether research productivities equalize, given controls such as teaching. In either case, measures have utility.

The most basic measure of teaching/administration load is probably total number of students divided by some staff complement—in the case of research, the latter could be total full-time equivalents (FTEs) or base researcher equivalents (BREs). However, such a measure ignores the fact that teaching loads are neither quantitatively not qualitatively equal, in either words, different staff members teach different numbers and types of students, with each class requiring potentially different requirements. For this reason, a full-information teaching load adjustment would reduce each staff member’s individual research intensity by some proportion relevant to that particular staff member’s teaching/administration intensity. However, the information gathering requirements of such a measure are inhibitive and unpractical in the extreme even if the assessments are internal (think of gathering course sizes and levels for each individual staff member, disaggregating students across different courses taken, and accounting for ad hoc changes in registrations and teaching allocations). Instead, because I am ultimately interested in RU-level research rather than individual, it is more practical to adjust overall research intensity by some measure accounting for overall average teaching/administrative loads.

Three major considerations exist. First, there is the overall number of students taught in the RU (which I assume is proportional to total administration), generally weighted for cross-RU course listings—that is, students taking courses from multiple RUs - and possibly weighted for level of student (first year, second year, etc.) Let this overall index be denoted C_k (class numbers for RU k). Without any weighting, the index is simply the total count of registered students doing any course in the RU. However, this does not allow for the fact that students take courses across multiple RUs, which instills double accounting of numbers. Many higher education institutions possess systems for this, such as counting student numbers not by person but rather as numbers of 'course-attendees' in a given RU (e.g. an Economics 1 attendee might be counted as one quarter of a student as, on average, this course comprises a quarter of a year's study credits). The analyst can easily construct an index of student numbers using these data. In addition, weights for level of teaching are possible (e.g., a Masters course might receive three times the weight of an undergraduate). Analysts might even choose to weight different fields of study differently—such considerations would be up to the institution.

Second, there is the *relative* numbers of research staff (R) versus non-research staff (F-R) allocated to teach and administrate this overall student body. An index that allocates researcher complement to teaching might simply take the ratio R_k/F_k, suggesting that if researchers comprise $q\%$ of the staffing body of RU k then they also do $q\%$ of the teaching/administration load. $C_k(R_k/F_k)$ is then an expression of the total numbers of students taught on average by research staff. However, many institutions would not operate like this: staff members allocated partially to research will generally do less teaching and administration. Therefore, we might incorporate an allocation constant (say λ) that is a standard average proportion of the extent to which the average non-research member deals with teaching more than an average researcher (i.e. λ indicates how many students are dealt with by a non-researcher for every 1 student dealt with by a researcher). This would not be difficult to estimate, the analyst can draw direct data from specific RU examples or a survey of heads of RUs may serve as a guide. $C_k(R_k/\lambda F_k)$ is then an adjusted expectation of the numbers of students taught by the research body of RU k.

Finally, because these figures mean relatively little on their own (they are relative to absolute student numbers), a cross-RU comparison of the teaching/administration intensity attached to researchers is helpful. Therefore, analysts might express the adjustment as a ratio of RU k's researcher teaching intensity compared to the average for the whole institution. This creates what I refer to as the 'Index of Researcher Teaching Intensity' (IRTI):

$$\mathrm{IRTI} = \frac{C_{k^*}}{\lambda \sum_{i_{k^*}=1}^{n_{k^*}} f_{i_{k^*}}} \sum_{i_{k^*}=1}^{m_{k^*}} f_{i_{k^*}} \Bigg/ \sum_{k=1}^{K} \left(\frac{C_k}{\lambda \sum_{i_k=1}^{n_k} f_{i_k}} \sum_{i_k=1}^{m_k} f_{i_k} \right) \tag{13}$$

where RU k* refers to any specific RU.

Consider an RU with weighted teaching numbers $C_k = 2000$, 50 staff members in total of whom 40 are (standard BRE-scored) researchers, and where in the institution it is found that on average a research and teaching staff member does 50% of the teaching and administration of an average non-researcher ($\lambda = 1.5$). The average RU on the other hand has 35 staff members, of whom 30 are researchers, and class numbers of 1000. Here the focal RU has raw researcher teaching/administration numbers of $(2000 \times 40)/(50 \times 1.5) = 1066.67$. The institutional average is $(1000 \times 30)/(35 \times 1.5) = 571.43$. The focal RU's IRTI is therefore $1066.67/571.43 = 1.87$. This infers that researchers in the focal RU teach and administrate roughly 87% more than average.

If desired, an analyst could multiply the IRTI ratio into any research output measure to produce a teaching-adjusted research index, for example a 'Teaching-Adjusted PPR' (TA-PPR):

$$\text{TA-PPR} = \text{IRTI} * \text{PPR} = \frac{\sum_{i=1}^{m} p_i}{\sum_{i=1}^{m} f_i r_i} \times \frac{C_{k^*}}{\lambda \sum_{i_{k*}=1}^{n_{k^*}} f_{i_{k^*}}} \sum_{i_{k^*}=1}^{m_{k^*}} f_{i_{k^*}} \Bigg/ \sum_{k=1}^{K} \left[\frac{C_k}{\lambda \sum_{i_k=1}^{n_k} f_{i_k}} \sum_{i_k=1}^{m_k} f_{i_k} \right] \tag{14}$$

Indices adjusted for concentration of research

The prior indices did not reflect the extent to which few hyper-productive staff members are responsible for the contribution to research or all staff hired to do research actually produce. A body of literature has considered this issue, suggesting the assessment of dispersal using measures such as Theil's measure, Gini coefficients, coefficient of variation, and Lorentz curves (Allison 1980; Okrasa 1987; Rousseau 2000). However, these measures have three weaknesses: they are more useful for diagnosis of concentration than integration into indices (which is the usage required here), they have failed to penetrate common RU-level measurement, and they generally fail to consider that perfect dispersion is not the natural benchmark. I address the latter point later in this section.

I present two alternate indices here that adjust for concentration, namely the *concentration ratio* and *HHI* ratios, based on economics measures of concentration (Scherer and Ross 1990). An analyst can add these extra ratio adjustments to any of the basic research assessments; here I use the *PPR* again.

The concentration ratio

The concentration ratio is the percentage of total research score generated by the top j out of m research staff members (e.g. the top 5 of 25 research staff). We have already stipulated that out of n total staff members we rank all members from highest $i = 1$ down so that person $i = m$ is the lowest-producing member, members $i = (m + 1 \ldots n)$ are non-research staff, and now we have added the stipulation that members $i = 1 \ldots j$ are the top j out of m total researchers $(0 < j < m)$. Therefore:

$$\text{Concentration ratio} = \frac{\text{P}_{\text{top j staff}}}{\text{P}_{\text{all researchers}}} = \frac{\sum_{i=1}^{j} p_i}{\sum_{i=1}^{m} p_i} \tag{15}$$

Say that $j = 5$ so we are examining the top 5 researchers. If the top 5 members produce 80% of research then the concentration ratio is .80. Since a higher ratio is presumably bad (as it indicates an increasing percentage of unproductive staff members), it is introduced as a penalizing factor in the research ratios, so that the research ratios will decrease by the extent of concentration. A 'Raw Concentration-Adjusted PPR' (CA-PPR) is therefore:

$$\text{CA-PPR} = \frac{\text{P per BRE}}{1 + \text{Concentration}} = \frac{\text{PPR}}{1 + \left(\text{P}_{\text{top j staff}}/\text{P}_{\text{all staff}}\right)} = \frac{\sum_{i=1}^{m} p_i / \sum_{i=1}^{m} f_i r_i}{1 + \left(\sum_{i=1}^{j} p_i / \sum_{i=1}^{m} p_i\right)} \quad (16)$$

This CA-PPR ratio unfortunately introduces a bias, because in penalizing all concentration it implicitly assumes that any concentration is bad. However, the optimal concentration ratio is *not* zero. One would expect the top j researchers to occupy a certain percentage of the RU's publishing output according to their position. The optimal concentration ratio is therefore, in fact, a denominator incorporating the BRE points of the top j researchers in the denominator divided by the total BRE points of the RU. I reflect this in an 'Optimality Ratio' for concentration:

$$\text{Optimality ratio} = \frac{\text{BREs}_{\text{top p staff}}}{\text{BREs}_{\text{all staff}}} = \frac{\sum_{i=1}^{j} f_i r_i}{\sum_{i=1}^{m} f_i r_i} \quad (17)$$

Therefore if the top 5 researchers are being used and they constitute 10% of the BRE points of the school, then one would *want* the concentration ratio to be .10 (indicating that these researchers are 'punching exactly their weight' in publishing). Therefore, the analyst can reduce the concentration ratio further by adjusting it for optimality. For instance, say concentration is .65 for $j = 4$ (4 people are producing 65% of the RU's P-score). However, those 4 comprise 50% of the RU's BREs (they are senior enough in comparison to others to comprise 50% of the weighted research-related staff complement). An analyst would expect them to produce 50% of the RU's research output, so the 65% concentration is only 15% over the expectation. The optimality-adjusted concentration ratio is therefore only $65\% - 50\% = 15\%$ or 0.15.

Given this, a 'Concentration and Optimality-Adjusted PPR' (COA-PPR) ratio is:

$$\frac{\text{P}_{\text{all staff}}/\text{BREs}_{\text{all staff}}}{1 + \left(\text{P}_{\text{top j staff}}/\text{P}_{\text{all staff}}\right) - \left(\text{BREs}_{\text{top j staff}}/\text{BREs}_{\text{all staff}}\right)} = \frac{\sum_{i=1}^{m} p_i / \sum_{i=1}^{m} f_i r_i}{1 + \left(\sum_{i=1}^{j} p_i / \sum_{i=1}^{m} p_i\right) - \left(\sum_{i=1}^{j} f_i r_i / \sum_{i=1}^{m} f_i r_i\right)} \quad (18)$$

Sum-of-squares adjustment through the HHI (penalizing concentration at an increasing rate)

The second measure is again adapted from industry concentration measures, the Herfindahl–Hirschman Index or HHI (Scherer and Ross 1990). The HHI ensures that the *more* that one or a few researchers dominate the output, the higher the penalty, and unlike the concentration ratio differentiates total inequality by group size. One can juxtapose the HHI and raw concentration ratios as follows. The raw concentration ratio attaches the same weight to an extra percentage of concentration no matter what base the increase is from (for example, an increase in concentration from 12% to 13% is penalized as much as an increase from 80% to 81%). The HHI attaches a higher penalty at higher levels of concentration than at lower levels. The HHI is:

$$\text{HHI} = \sum_{i=1}^{m} \left(100 \times \frac{\text{p}_{\text{individual}}}{\text{p}_{\text{all researchers}}}\right)^2 = \sum_{i=1}^{m} \left(100 \times \frac{\text{p}_\text{i}}{\sum_{i=1}^{\text{m}} \text{p}_\text{i}}\right)^2 \quad (19)$$

The HHI is harder than the concentration ratio to understand and incorporate, as it has no particular upper limit. However, researchers can calculate an 'optimal' HHI, because if there is no concentration then one expects the per-person research assessment to equal to the per-person contribution to the total unit's BRE (i.e. with no concentration, a member's share of the research assessment should equal their contribution to staff complement, weighted for seniority (identical thinking to the optimality ratio introduced earlier). Therefore, the 'Optimal HHI' for an RU is:

$$\text{Optimal HHI} = \sum_{i=1}^{m}\left(100 \times \frac{\text{BRE}_{\text{individual}}}{\text{BRE}_{\text{all researchers}}}\right)^2 = \sum_{i=1}^{m}\left(100 \times \frac{f_i r_i}{\sum_{i=1}^{m} f_i r_i}\right)^2 \tag{20}$$

Therefore, instead of using the HHI directly, we can use it as a proportion of the optimal in the denominator of the research ratio, for instance an HHI-Adjusted PPR (HHI-PPR) would be:

$$\text{HHI-PPR} = \frac{\text{PPR}}{\text{HHI}_{\text{actual}} / \text{HHI}_{\text{optimal}}} \tag{21}$$

It would probably be unwise to compare the HHI-adjusted PPR to the raw PPR, as the adjustment for HHI is meaningful in input but it's scale effect is not necessarily so. For example, if raw PPR is .76 (on average each staff researcher produces .76 points), but the HHI-adjusted ratio is .50, this does not necessarily mean we have 'lost' .26 points in value. Rather, we compare the adjusted ratio with the same index in other years or between justifiably comparable RUs.

Adjustments for lag

Research might not occur contemporaneously with staffing complement. In fact, quite the opposite: one might feasibly expect staffing to lag research output by some time, especially in the case of citations. This is especially true when research staff members are new, and institutions cannot yet expect them to have written, submitted and published material in the name of the RU. However, a search of literature reveals almost no empirical studies that lag staffing complement on research. Operations research adjustments (Stevenson and Ozgur 2007) allow for various lagging options.

Simple single-period lag-adjustment

Probably the simplest lag method is to link a given period's publication scores to a prior period's staffing complement (e.g. a one-year lag). To achieve this let time be denoted as $t \geq 0$ where $t = 0$ is the most recent period and $t > 1$ indicates number of time periods prior to the most recent period. Then P_t and R_t represent an RU's P and BRE points respectively at time t. The analyst simply estimates staffing for a specific $t > 0$ (e.g. for R_1) whereas (s)he estimates research measures for time $t = 0$ (P_0). Therefore, a 'Single-period Lagged PPR' ratio (SL-PPR) would be:

$$\text{SL-PPR} = \frac{P_{\text{time0}}}{BREs_{\text{timet}>0}} = \frac{P_0}{R_{t>0}} \tag{22}$$

Simple moving average multi-period lag adjustments

Prior-period lagging obviously assumes that all current research output is entirely due to the prior period's staffing complement alone. However, it is likely that a research assessment at $t = 0$ (P_0) is a cumulative effect of not only of the current staffing complement at $t = 0$ (R_0, as assumed by the non-time-related ratios above), or a single lagged $t > 0$ period back ($R_{t>0}$, as assumed by the SL-PPR), but of a combination of both current staffing and potentially also of *several* prior periods rather than one. In other words, research might be a combined effect of several years of interactions between staff members, planning, acquisition of skills, loss of skills, phases of grants, and so-forth. Researchers generally assume in such calculations, however, that recent inputs affect outcomes such as P_0 more than inputs farther away, although the analyst might place a higher weight on, say, $t = 1$ than $t = 0$ if desired.

Generally, the strategy in multiple-period lagging is to take some weighted average of several periods, where periods closer to the present are often, although not always, given a higher weight. There two methods of doing this, weighted moving average and exponential smoothing.

Research ratios including a weighted average lag for prior staffing complements use a specific limited number of time periods prior to and, if desired, including, the current period. For example, one might decide that current research assessment is dependent on the current year's staff complement as well as the previous two years' staff complements. The analyst would then weight staffing complement of each of these periods by any chosen set of weights.

Total weighted staff complement at time t will be denoted WR_t where weighting is done over a time periods prior to t, (where $a > 0$, so for example going back two years prior to the current period requires $a = 2$). The total staff complement is then:

$$WR_t = \sum_{t}^{t+a} w_t R_t \quad \text{where} \quad \sum_{t=0}^{a} w_t = 1 \tag{23}$$

Given this, a 'Lagged Moving-Average PPR' (LMA-PPR) would be:

$$\text{LMA-PPR} = \frac{P_t}{BREs_{\text{multiple weighted periods}}} = \frac{P_t}{WR_t} = \frac{P_t}{\sum_{t=0}^{t+a} w_t R_t} \tag{24}$$

Two decisions determine weighted averages. First is how many and which periods to incorporate, including whether the analyst include the last period's staffing complement (i.e. the same period in which the research output is generated) and the number of prior periods to include. Second, the analyst must choose the weights to assign to each period. Such a weighted variable can then be used in lieu of R in any given research ratio, not just the PPR.

Similarly, in ratios calling for staff budgets such as S_R (Eq. 6), the researcher may again use weighted averages for several years of budgets. However, the analyst should adjust for inflation (because previous years' budgets were in values related to those periods). If you can calculate an average/overall inflation rate for each period, then if the inflation (discount) rate is d the analyst can estimate the weighted salary budgets (WS_R) over several prior periods as follows:

$$WS_{R_t} = \sum_{t=0}^{a} w_t S_{R_t} (1 + d)^t \tag{25}$$

If the inflation rates are realistically not amenable to a single average/overall value (e.g. if there is a far larger increase in one of the years compared to others), then a researcher may need to inflate salaries to current values one at a time using different values of inflation.

Exponential smoothing

Exponential smoothing (Brown et al. 1961) is a specific weighted average method with no particular cut-off for number of prior periods—the measure continually updates every new period, down-weighting periods further back and adding a portion of the new period's value. Specifically, if a given time period's actual staff complement is R_t, the total weighted average staffing (WR_t) is based on two things: (a) the previous period's weighted average staff complement (WR_{t+1}) and (b) an update for the extent to which actual staffing for the current period (R_t) differs from that for the previous period's weighted average (again WR_{t+1}) according to the following formula:

$$WR_t = WR_{t+1} + \beta(R_t - WR_{t+1}) \tag{26}$$

where $0 \leq \beta \leq 1$ is an "update" proportion determined by the analyst.

In other words, weighted average staffing here is a proportion of last period's weighted average (the most recent indicator of past staffing WR_{t+1}) updated for the extent to which the current period's staffing differs from the past ($R_t - WR_{t+1}$). The update proportion β is the indication of how much the analyst wishes to build recent changes in staffing into the measure rather than previous periods' staff complement. The base research measure does not change (e.g. Eq. 24 would still stand, but it would use the WR_t measure from Eq. 26).

Exponential smoothing has several advantages. First, it has no particular limit on which prior periods the researcher builds into the current estimate. It is more efficient than weighted moving averages as it only requires a decision on the update factor.

Two technical issues arise in the calculation of an exponentially smoothed staffing factor. First is with what initial values to start the calculations? From the estimation of Eq. 26, one can see that for any given current weighted staffing complement WR_t, one needs a previous value on which to base it (WR_{t+1}). However, when starting off, what previous value do you use? There are several possibilities, but the usual suggestion is simply to use the same weighted average as the current value (i.e. in the first period $WR_t = R_t$). Second, how high does one set the value of β? The higher the update factor (towards 1) the more that the index includes recent changes in staffing at the expense of past staffing levels, and vice versa. An analyst can employ trial and error, but it is possible to use statistical modeling to find the level of β that will maximize the correspondence between staffing and research levels. This assumes, of course, that staffing complement 'causes' or is statistically associated with research levels, but this is not an unreasonable assumption and is the assumption that underlies this paper (e.g. Dundar and Lewis 1998). Panel regressions are the likely option for finding a maximizing level of β.

Possible problems with combining too many adjustments

Problems may arise if an analyst attempts to integrate too many of these sorts of adjustments into a basic index (P) at the same time. An example is combining principles of concentration and lag. This is because of the variable selection and attrition of staff, and the fact that the top j out of m researchers often differ between periods. For example, at $t = 0$ we may pick four top producers. One of these might only have joined the RU at the beginning of the year. If we are using single-period lagging, this top researcher was not even part of the RU in the lagged period. With multi-period weighted lagging, the weighted staffing complement is by definition an amalgamation of several periods, such that concentration measures such as the concentration ratio are not quite exact. Is this a problem? Perhaps not, but if there is cause for concern, then various remedial measures might be possible. For instance, one could lag overall staffing complement (e.g. BREs) in both the basic research measure and the optimality portion of the concentration ratio, probably using multi-period not single-period lagging. Take for example the Concentration and Optimality-Adjusted PPR (COA-PPR) Eq. 18, in which BREs for all staff feature in both basic PPR and the denominator of the optimality ratio. Lagging both for multiple times changes interpretation only slightly, assessment is now current research per head of *ongoing* staff complement, adjusted for current concentration yet excused for the fact that the top j current researchers are optimally a certain proportion of *ongoing* staffing.

However, the overall point applies: excessive combination of adjustments potentially obfuscates a core productivity index. A better approach is to present a range of control indices and adjusted measures (e.g. the original P-score, stand-alone concentration measures, the P-score adjusted for lag and for concentration separately, and a complete combination). This allows for more careful consideration of the importance and impact of adjustments and their combinations.

Practical considerations in implementation

This paper has presented many advanced ratios and indices for potential inclusion into research measures. These indices are not intended to be used alone, but following prior suggestions should be included as complementary measures in a range of indices (Martin 1996). In addition, I note that some of the indices reflect contextual or input factors (such as the concentration ratios). These can be included and analyzed on their own or incorporated into net research assessments (Research Evaluation & Policy Project 2005). On their own, such ratios nonetheless give much information.

Many of these ratios require individual-level staffing and research data. This is feasible for institutions, which after all control the human resource databases of RUs and other sources or information, although some finesse might be required. It would be harder, although not completely impossible, for external researchers to gather this data (for example, while RUs and individual commonly publish publication lists allowing for analysis of counts and citations, staffing and other budgets are generally proprietary to institutions).

Contribution

The proposed indices have advantages and complementarities compared to extant measures. First, very few research indices explicitly take into account the characteristics of RUs mooted here. For instance, I identified relevant historical contributions assessing research at the level of RUs by a) doing a complete literature search for all titles including words such as department(s), unit(s), entitie(s) and the like, and furthermore, b) assessing post-1999 *Scientometrics* abstracts in detail to get a sense of whether these keywords might tend to omit relevant articles. The search uncovers a representative and relevant set of RU-level contributions (Alewell 1990; Baird 1991; Carpenter et al. 1988; Colman et al. 1995; Crewe 1988; Donovan 2005; Dundar and Lewis 1998; Krampen 2008; Nederhof et al. 1993; Nederhof and Noyons 1992; Okrasa 1987; Ramsden 1994; Research Evaluation &

Policy Project 2005; Schloegl et al. 2003; Sternberg and Litzenberger 2005; Tan 1986). Of these, the following is noted:

1. With regard to RU size, only about half (Baird 1991; Crewe 1988; Krampen 2008; Qurashi 1993; Research Evaluation & Policy Project 2005; Schloegl et al. 2003; Sternberg and Litzenberger 2005; Tan 1986) adjust or suggest adjusting for gross (FTE) staffing size;
2. Only Krampen (2008), Nederhof et al. (1993), Research Evaluation & Policy Project (2005), Sternberg and Litzenberger (2005), and Tan (1986) account barely for research intensity (broadly-defined position or PhD percentages, none account properly for experience, level, or teaching load);
3. Only Crewe (1988) and Okrasa (1987) measure concentration and only Crewe (1988) actually adjusts research measures for it (using a measure similar to the raw concentration ratio adjustment). No contribution adjusts for concentration optimality. The concentration adjustments posited here have an advantage over traditional statistical measures such as Gini coefficients, Theil's measure and coefficient of variation (Allison 1980; Okrasa 1987) as they are optimality adjusted, allowing for a natural dispersion of staff ranks and experience with the expectation that some inequality of dispersion is expected and even desired. In addition, they are easier to compare between different RUs. Although it is possible to include a comparable 'cumulative advantage' model that increments expected output for prior productivity (Allison 1980), researchers rarely use such models as they require lifetime publishing records, unlike those mooted here which only require current-period data;
4. Only Schloegl et al. (2003) includes a crude form of lag in staffing (i.e. average of staff numbers over a number of years prior to citation).

Therefore, this article adds to a sparse literature, which has previously dealt with these issues infrequently, incompletely or not at all.

Second, these indices provide additional direction to the practical institutional management of research. Institutions and RUs are interested not only in the quantity or quality of research production but also in the diagnosis of underlying causes and consequences. Since RU-level characteristics such as research intensity are most amenable to managerial intervention (e.g. through human resources policies designed to alter workloads or funding policies), the current indices are most useful to institutional and RU-level research management. Most articles looking at RU-level research are external and positivist, with relatively little internal, managerial view.

Conclusion

Assessing RU-level research has its uses, especially in identifying and acknowledging successes or problems, and potentially in diagnosing changes in research. This article serves as an initial set of more advanced complementary options for issues that exist but that researchers have hitherto not addressed sufficiently. The benefit of such indices and adjustments lays in their balanced and judicial use alongside more conventional bibliometric measures, with the aim of identifying areas for positive improvement. Assuming that institutional managers and researchers use these measures in a mature and constructive manner, increased knowledge of the patterns, trends and underlying factors of research can only improve the management of research RUs.

References

Alewell, K. (1990). Criteria for performance profiles of departments and universities. *Scientometrics, 19*(5–6), 337–347.

Allison, P. (1980). Inequality and scientific productivity. *Social Studies of Science, 10*, 163–179.

Baird, L. (1991). Publication productivity in doctoral research departments: Interdisciplinary and intradisciplinary factors. *Research in Higher Education, 32*, 303–318.

Brown, R. G., Meyer, R. F., & D'Esopo, D. A. (1961). The fundamental theorem of exponential smoothing. *Operations Research, 9*(5), 673–687.

Carpenter, M. P., Gibb, F., Harris, M., Irvine, J., Martin, B. R., & Narin, F. (1988). Bibliometric profiles for British academic institutions: An experiment to develop research output indicators. *Scientometrics, 14*(3–4), 213–233.

Colman, A. M., Dhillion, D., & Coulthard, B. (1995). A bibliometric evaluation of the research performance of British university politics departments: Publications in leading journals. *Scientometrics, 32*(1), 49–66.

Crewe, I. (1988). Reputation, research, reality: The publication records of UK departments of politics 1978–1984. *Scientometrics, 14*(3–4), 235–250.

Donovan C. (2005). *A review of current Australian and International practice in measuring the quality and impact of publicly funded research in the humanities, arts, and social sciences*. Discussion paper 05/3, Research School of Social Sciences, Australian National University.

Dundar, H., & Lewis, D. R. (1998). Determinants of research productivity in higher education. *Research in Higher Education, 39*(6), 607–631.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572.

Johnes, G., & Johnes, J. (1992). Apples, oranges: The aggregation problem in publications analysis. *Scientometrics, 25*(2), 353–365.

Krampen, G. (2008). The evaluation of university departments and their scientists. *Scientometrics, 76*(1), 3–21.

Lewinson, G. (2001). Evaluation of books as research outputs in history. *Research Evaluation, 10*(2), 89–95.

Lotka, G. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*, 317.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics, 36*, 343–362.

Nederhof, A. J., Meijer, R. F., Moed, H. F., & Van Raan, A. F. J. (1993). Research performance indicators for university departments: A study of an agricultural university. *Scientometrics, 27*(2), 157–178.

Nederhof, A. J., & Noyons, E. C. M. (1992). Assessment of the international standing of university departments' research: A comparison of bibliometric methods. *Scientometrics, 24*(3), 393–404.

Okrasa, W. (1987). Differences in scientific productivity. *Scientometrics, 12*(3–4), 221–239.

Patrick, W. J., & Stanley, E. C. (1998). Teaching and research quality indicators and the shaping of higher education. *Research in Higher Education, 39*(1), 19–41.

Qurashi, M. M. (1993). Dependence of publication-rate on size of some university groups and departments in UK and Greece in comparison with N.C.I., USA. *Scientometrics, 27*(1), 19–38.

Ramsden, P. (1994). Describing and explaining research productivity. *Higher Education, 28*, 207–226.

Research Evaluation & Policy Project. (2005) *Quantitative indicators for research assessment—A literature review*. Discussion paper 05/3, Research School of Social Sciences, Australian National University.

Rousseau, R. (2000, October 23–25). *Concentration and evenness measures as macro-level scientometric indicators*. Paper presented at the 2nd International Seminar on Quantitative Evaluation of Research Performance, Shanghai.

Scherer, E. M., & Ross, D. (1990). *Industrial market structure and economic performance*. Boston: Houghton Mifflin.

Schloegl, C., Gorraiz, J., Bart, C., & Bargmann, M. (2003). Evaluating two Austrian university departments: Lessons learned. *Scientometrics, 56*(3), 287–299.

Sternberg, R., & Litzenberger, T. (2005). The publication, citation output of German Faculties of Economics, Social Sciences: A comparison of faculties and disciplines based upon SSCI data. *Scientometrics, 65*(1), 29–53.

Stevenson, W. J., & Ozgur, C. (2007). *Introduction to management science*. Boston: McGraw-Hill Irwin.

Tan, D. (1986). The assessment of quality in higher education: A critical review. *Research in Higher Education, 24*, 223–265.

Tijssen, R. J. W. (2007). Africa's contribution to the worldwide research literature: New analytical perspectives,trends, and performance indicators. *Scientometrics, 71*(2), 303–327.

Toutkoushian, R. K., Porter, S. R., Danielson, C., & Hollis, P. R. (2003). Using publication counts to measure an institution's research productivity. *Research in Higher Education, 44*(2), 121–148.

Verbeek, A., DeBackere, K., Luwel, M., & Zimmerman, E. (2002). Measuring progress, evolution in science, technology—I: The multiple uses of bibliometric indicators. *International Journal of Management Reviews, 4*(2), 179–211.