An Evaluation of Metrics Used by the Performance-Based Research Fund Process in New Zealand

Robert A. Buckle and John Creedy

WORKING PAPER 16/2017
October 2017

Working Papers in Public Finance

VICTORIA
UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA
O TE ŪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

The Working Papers in Public Finance series is published by the Victoria Business School to disseminate initial research on public finance topics, from economists, accountants, finance, law and tax specialists, to a wider audience. Any opinions and views expressed in these papers are those of the author(s). They should not be attributed to Victoria University of Wellington or the sponsors of the Chair in Public Finance.

Further enquiries to:
The Administrator
Chair in Public Finance
Victoria University of Wellington
PO Box 600
Wellington 6041
New Zealand

Phone: +64-4-463-9656
Email: cpf-info@vuw.ac.nz

Papers in the series can be downloaded from the following website:
http://www.victoria.ac.nz/cpf/working-papers

An Evaluation of Metrics Used by the Performance-Based Research Fund Process in New Zealand*

Robert A. Buckle and John Creedy
Victoria University of Wellington

Abstract

The New Zealand Performance-based Research Fund (PBRF) applies a unique set of metrics to assess researchers and rank disciplines and universities. The process involves giving a 'raw score' to individual researchers and then assigning them to one of four Quality Evaluation Categories (QECs), used to derive Average Quality Scores (AQS). This paper evaluates the properties of these metrics and argues that QEC thresholds influence the final distribution of assessments. The paper also demonstrates that the derivation of AQSs depends on the weights assigned to each QEC and the distribution of portfolios. The method used to determine the raw scores also has an independent effect on the distribution of scores. The number of researchers included in evaluations also influences the rankings. The paper compares how research rankings of New Zealand universities would vary if alternative summary measures, based on the raw scores rather than QECs, were used to evaluate performance.

Key words: Research quality, Peer review, PBRF, New Zealand universities,
JEL classifications: I2; I23; I28.

*We are very grateful to Amber Flynn and Sharon Beattie for providing the TEC data used here and helpful discussions regarding the PBRF process, and John MacCormick for mentioning useful references. We have benefited from helpful comments on an earlier draft by Norman Gemmell and Gary Hawke.

1 Introduction

The Performance-based Research Fund (PBRF) process was introduced in New Zealand in 2003 as part of a method of allocating university research funding to tertiary education organisations on the basis of research performance, rather than on the number of students. This is part of a world-wide emergence of performance-based evaluation schemes to underpin funding of tertiary institutions; see OECD (2010) and de Boer *et al.* (2015).[1] In their summary of the literature regarding research evaluation and metrics, Wilsden, *et al.* (2015, p. viii) argue that:

> 'Metrics should support, not supplant, expert judgement. Peer review is not perfect, but it is the least worst form of academic governance we have, and should remain the primary basis for assessing research papers, proposals and individuals, and for national assessment exercises ... Research assessment needs to be undertaken with due regard for context and disciplinary diversity. Academic quality is highly context-specific, and it is sensible to think in terms of research qualities, rather than striving for a single definition or measure of quality'.

From the point of view of these comments, positive features of the PBRF process are that it evaluates all researchers, is based on peer review and uses discipline panels. This paper concentrates on the precise measures of research quality used by the Tertiary Education Commission (TEC) to compare the quality of researchers and disciplines, and to rank universities.[2] A close examination of the metrics used is warranted in view of the considerable importance attached to the ranks, even by those universities which attract substantial non-public funding.[3] Furthermore, the absence of a clearly-stated rationale for the assessment method, along with the high compliance costs involved, makes it important to consider a range of implications of the approach.

Three measures are used to allocate Government funding to support research at universities and other Tertiary Education Organisations (TEOs). The first component

[1] For background and detailed discussion of the PBRF, see New Zealand Tertiary Education Commission (2002, 2013).

[2] A broader discussion of the process and an analysis of the evolution of research quality in NZ universities between 2003 and 2012 is in Buckle and Creedy (2017).

[3] The importance of research rankings for the reputation of universities is stressed in OECD (2010).

is Quality Evaluation, which comprises 60 per cent of the Fund allocated on the basis of an assessment of the research quality of eligible staff. The second component is Research Degree Completions, which comprises 25 per cent. The third component is External Research Income, which comprises 15 per cent and is based on the amount of external research revenue generated. This paper is concerned only with the quality evaluation component.

The research quality of each staff member is measured using a 'Quality Evaluation Category' (QEC). Each person is required to submit an Evidence Portfolio (EP), which is assessed by a relevant discipline panel. In turn, the QECs are used to produce an overall Average Quality Score (AQS) for each university. The evaluation actually has several steps. First, 'raw scores', ranging from 0 to 700 are obtained for each individual. Second, the QEC is determined using a set of four ranges within which the raw scores fall. Third, a numerical value, or weight, is attached to each QEC, and finally the AQS is obtained as a weighted average of the QECs.[4]

Faced with considerable heterogeneity of individuals within each university, the aim of the process is essentially to provide some measure of average quality. Having determined a method of obtaining a distribution of raw scores, the problem of comparing universities could be expressed in terms of determining the statistical measure of 'location' that is deemed to be most appropriate. For example, should a median, arithmetic mean or geometric mean be used? A chosen measure of location could be used to rank universities, and further information about the nature of each distribution could be provided by reporting, say, a number of quantiles or percentiles.[5] However, this direct path has been eschewed in favour of the method described above, involving the transformation of the raw scores, depending on the range into which each individual's score falls. Nevertheless, no official rationale has been given for this method or the particular weights used, and no changes have been made to the basic method. It is obvious that the process need not necessarily rank universities in the same order as

[4] In 2012 the TEC developed four alternative AQSs which varied according to the denominator. For example, $AQS(E)$ uses equivalent full time students; $AQS\,(P)$ is a subset using equivalent full time postgraduate degree students; $AQS\,(S)$ uses all academic teaching and research staff. The measures used in this paper is in fact $AQS\,(N)$, which uses the full time equivalent staff for which evidence portfolios were submitted and which TEC (2012, p. 7) considers 'reflects the concentration of research excellence in the university sector'.

[5] Of course it is recognised that the value of a measure of location depends to some extent on the nature of the distribution, for example, whether it is single or multimodal.

a conventional location measure based on raw individual scores.

Section 2 briefly explains how the research quality of each university is calculated. One view is that, having used the distribution of raw scores only to allocate people to categories, the method essentially discards a considerable amount of information compiled in the first stage of the measurement process. An alternative view is that the raw scores contain considerable 'noise' and that discrete categories are most appropriate. This issue is also discussed in Section 2, along with the numerical properties of the scoring system. Section 3 examines the construction of AQSs and QECs from the basic scores given to each research portfolio. This highlights the way the AQS varies as a result of systematic variations in the weights. The sensitivity of ranks to the weights is examined in Section 4, using the QEC distributions for each university from the 2012 'round'. Section 5 considers a range of alternative measures of location using the original raw scores assigned to each researcher portfolio in the 2003 and 2012 rounds and compares these measures and the consequential university ranks with those derived by the QEC-based AQSs. Section 5 also evaluates the distributional properties of the raw scores, and how the properties of those distributions compare with the distributional properties commonly found in the research literature on research productivity. It also considers how the distributions may have been affected by the assessment process and QEC system. Brief conclusions are in Section 6.

2 The PBRF Scoring System

As mentioned above, the Tertiary Education Commission publishes an Average Quality Score (AQS) for each university at the conclusion of each PBRF round. To evaluate each researcher's score, each evidence portfolio (EP) is assessed by a panel assigned to a discipline or group of disciplines. The process is described by Ministry of Education (2012, p. 21) as follows:

> EPs are evaluated through a rigorous, collaborative process. EPs are assigned to a primary and secondary panellist who independently assess the EP and then agree an initial score together. This score is then discussed at the panel meeting and a final score is decided. Then all the scores are moderated by panel and between the panels.

Only the previous six years are considered, and the evaluation does not necessarily consider all relevant information since the portfolio can mention only a limited number of publications and other contributions to the research environment.[6]

Subsection 2.1 explains the definition of QECs and their relationship to raw PBRF scores. Subsection 2.2 considers the general question of whether research quality is considered to be a discrete or continuous variable. Some unusual numerical features of the metric used to obtain raw scores are discussed in Subsection 2.3.

2.1 Raw Scores and Definition of QECs

The relevant discipline panel assesses the quality of each portfolio and assigns an integer score from 0 to 7 for each of three categories: in the rounds to date (2003, 2006 and 2012) these are 'research output'; 'peer esteem'; and 'contribution to research environment'. These three scores, s_j, are given weights, q_j, of 0.70, 0.15 and 0.15 respectively. The total 'raw score', x, for an individual is obtained by multiplying the weighted sum of the s_j values by 100. Hence:

$$x = 100 \sum_{j=1}^{3} q_j s_j \tag{1}$$

Thus the maximum individual score is 700 and multiplication by 100 ensures that the values of x are integers. Importantly, the s_j scores are considered as cardinal measures of quality in each category.

A letter grade, referred to as the 'Quality Evaluation Category', QEC, is then assigned depending on the final assessed total raw score. These are as follows: R for scores 0 to 199; C for scores between 200 and 399; B for scores from 400 to 599; and A for scores from 600 to 700. An A is intended to reflect 'international standing'; B reflects 'national standing' and C reflects 'local standing'.[7] An R denotes that an individual is not considered to be research active.[8] The analyses in this paper makes

[6] The Tertiary Education Commission undertakes a moderation process at the commencement of each PBRF assessment round to encourage consistent assessment over time and across discipline panels. This process includes training expert panel members on assessment methods and referring to a sample of evaluations from previous PBRF rounds.

[7] For further discussion of the categories, see Ministry of Education (2012, p. 20-21).

[8] The recognition that new researchers may take time to establish their research, publications, and academic reputations led to the introduction in 2006 of the new categories, C(NE) and R(NE). These

use of the distribution of anonymous QECs as well as the distribution of raw scores for each researcher portfolio.

2.2 Discrete Versus Continuous Metrics

A fundamental initial judgement required in any research quality appraisal involves the question of whether quality is perceived to be a continuous (or quasi-continuous) variable, subject to calibration using a particular metric or set of metrics, or whether it is discrete, whereby individuals are placed into a small number of well-defined categories. Within the second view of quality, such categories may be regarded as purely ordinal, producing merely a rank-order among individuals (say from 'low' to 'high' quality) or they may be given a cardinal value, allowing addition over individuals in an institution and therefore averaging.

The basic view of 'quality' has not been articulated by the TEC or the Ministry of Education and it has been seen that the PBRF process involves what might be described as a hybrid method.[9] The PBRF's ultimate objective is to place individuals into four discrete categories, R, C, B, and A, defined above. Instead of using a scoring method in a nonlinear process, it may be expected that the classification of individuals into these discrete categories would proceed by, for example, setting out a list of performance characteristics or criteria expected within each category. A comparison might be made with the academic promotion process, where a range of characteristics describe performance expectations at each level.[10]

Here a comparison may perhaps be made with the classification of students into 'third class', 'lower second', 'upper second' and 'first class'. The process begins with a series of examination marks for a number of subjects, and a wide range of conventions are used in order to arrive at a single 'measure of location' to determine the final grade. It might be suggested that the initial stage of determining examination marks for different subjects contains a substantial amount of 'noise', when considered as

apply to new and emerging researchers who did not have the benefit of a full six-year period. For present purposes, these can be ignored and all Cs and Rs are grouped together.

[9] The TEC did not actually have the responsibility for designing the metrics used, but has only the task of administering it.

[10] Of course universities do not (unlike the PBRF process) repeatedly reconsider each individual's position to check if performance is being maintained after promotion, and adjust the assigned level where necessary.

signals of quality.[11] Yet it is known that, where large numbers of students are involved, the examination marks produce frequency distributions that conform to regular forms, in particular that of the Normal distribution. Examiners have a sufficient degree of confidence in the marking system to make substantial use of those assigned marks in awarding the final degree grade.

However, the PBRF two-stage process does not have such characteristics. Indeed, the 'first stage' consists of several stages, where the raw scores are subject to substantial adjustments. And even in the 'second' stage, of assigning the Quality Evaluation Category, adherence to the raw scores is not strict and many individuals are given a higher QEC than would be warranted by the raw score: that is, their raw score is effectively moved up to the next threshold. As shown in Section 5, the resulting distributions of raw scores display substantial 'spikes' at the thresholds used for QEC determination. However, there is considerable evidence that other continuous metrics used to measure research output or quality follow regular positively skewed distributions.

One interpretation is that the adjustments would seem to reflect a lack of confidence in the initial metrics used. The further use of QECs may reflect the view that the initial metrics contain considerable 'noise'.[12] But then the question arises of why such a process continues to be used. If the metrics are considered to provide a poor 'signal' of quality, they could be dispensed with and it would be much simpler to proceed directly to assigning discrete QEC categories, rather than retaining the ambivalent view that research quality can be measured by a quasi-continuous metric while reporting (and using) only the discrete QEC.

In the absence of a clearly articulated official view, the present paper is limited in scope. Its aim is to examine properties of the PBRF scoring method which do not seem to have been appreciated, and to consider the sensitivity of university ranks to various cardinal weights used in the production of the AQSs. In this way it can contribute

[11] Indeed for many years neither students nor potential employers were given information about the separate examinations marks. Furthermore, in many subjects the judgement that a particular examination paper deserved a 'first class mark' did not lead to excessive concern over the precise numerical score, so long as it was sufficiently above the threshold required (particularly where it was known that the final grade was not based on a simple averaging process).

[12] It is possible that the metrics used to obtain the initial raw scores are considered to be useful and accurate, but partial, measures, to be supplemented by a range of (unspecified) qualitative considerations. However, marginal adjustments would then be expected, rather than the less transparent large ones. The specified and publicised metric provides the incentive structure facing individuals, and it is therefore desirable that it gives clear and appropriate signals to researchers.

to the broader debate about the appropriate measurement of individual university research quality and the resulting set of incentives facing individuals and institutions.

2.3 Numerical Properties of Raw Scores

The method used to derive raw scores, involving equation (1), has some properties which may not be immediately apparent. With three components, s_j, for $j = 1, 2, 3$, where each is given an integer score from zero to seven, the values of the integers, x, range from 0 to 700, but many values in the range are not possible. The number of possible values actually depends on the weights, q_j, used to form the aggregate, x. The number of permutations, with three components, each ranging from zero to seven, is 512. With PBRF weights of 0.7, 0.15 and 0.15, the possible values of x start from 0 and increase in increments of 15 until reaching 60. The next two numbers in the sequence are 70 and 75, after which they increase in increments of 5 up to 570. The next number is 580 and then the x values increase in increments of 5 up to the maximum of 700. Furthermore, the number of ways of achieving a given value of x varies with x itself. To give a few examples, there are two ways of obtaining a score of 15, but 5 ways of getting a score of 60, six ways of obtaining 345, 7 ways of getting a score of 90, and 8 ways of getting 315.

This means that, if everyone has an equal independent probability of obtaining a score from 0 to 7 for each of the quality types, there is not an equal probability of falling into specified equal ranges of x. The proportions of each of the QECs, under the assumption of equal probabilities, are shown in the first column of Table 1, using the thresholds of 200, 400, and 600 by which individuals are judged to be C, B and A respectively. If the thresholds are unchanged but the weights, q, are adjusted, the resulting distributions are shown in the second and third columns of the table. With each of the two alternative sets of weights, the possible integer values of x simply move from 0 in increments of 10 up to 700, but of course the number of times each possible integer can appear varies over the range.

These columns show that, with a population having uniform probabilities of achieving from 0 to 7, and uniform intervals of 200 for the thresholds governing QECs of R, C and B, the distribution of QECs is not itself uniform. This is purely a numerical artefact of the use of integers and three measures between 0 and 7, along with the

Table 1: QEC Distribution for Uniform Probabilities and Alternative Weights

	Using scores 0 to 7			Scores 0 to 10
	q: 0.7; 0.15; 0.15	q: 0.8; 0.1; 0.1	q: 0.6; 0.3; 0.1	q: 0.7; 0.15; 0.15
R	0.227	0.258	0.186	0.217
C	0.357	0.313	0.408	0.370
B	0.350	0.313	0.348	0.357
A	0.066	0.117	0.059	0.057

various weights and thresholds. The range of x values for which a QEC of A applies is half the range for the other quality categories, yet the equal-probability case results in a far smaller probability of obtaining an A grade than might be anticipated, given simply the information that scores range from 0 to 700.

The question arises of whether a wider range for each s, say from 0 to 10, would make a difference. Such a choice results in 1,573 permutations, for integer values of x ranging from 0 to 1,000. Again, with the current weights of 0.7, 0.15, and 0.15, the intervals between integers vary, along with the number of possible ways of getting any given value. If the total range is divided in a similar way to the 0 to 700 range, the thresholds above which C, B and A apply are 285, 570 and 855 respectively. The resulting distribution of QECs, again supposing equal independent probabilities over s values, is shown in the final (far-right) column of Table 1. This distribution turns out to be similar to that obtained using the current metric, shown in the first column of the table.

The assumption of equal probabilities is obviously not realistic, but it is a useful case to take in order to illustrate the basic numerical implications of the scoring system used. It is important to recognise that the scoring system can itself have a substantial effect on the measured quality distribution of researchers. This raises questions as to whether the presence of threshold levels, created by the use of QECs and the weights attached to each category, influence the decisions of peer reviewers and discipline expert panels in assigning raw scores and moving people to the next threshold. It also raises the possibility that the final distribution of raw scores could have properties that differ from those that would emerge in a system using the only raw scores to derive average scores for disciplines and universities rather than the AQSs based on the weighted QECs. These issues are examined in Section 5 below.

3 The QEC Weights and AQS Ranks

This section examines the properties of the Average Quality Score, AQS, as computed for each university, using the QECs. Subsection 3.1 explains how the AQS is defined, using a set of weights assigned to the QECs. Subsection 3.2 clarifies the relationship between the AQS, the values or weights, w, used for each QEC, and the distribution of the score, x, discussed in the previous section and defined in equation (1). The relationship between theAQS and the degree of compression or dispersion of weights is explored in Subsection 3.3. Changes in the rank order of two universities are considered in Subsection 3.4.

3.1 Definition of the AQS

Having determined a QEC value for each individual, a numerical value, say w, is then assigned to each letter grade.[13] The values used are as follows: $w_A = 10$ for an A; $w_B = 6$ for a B; $w_C = 2$ for a C; and $w_R = 0$ for an R category. A university's AQS is the employment-weighted mean of these values. Define the employment proportion of person i as $e_i \leq 1$, and let N denote the relevant number of employees in a university. Let G_i denote the numerical quality score for person i, obtained from the appropriate weight: hence if i is judged to be an A-type researcher, $G_i = w_A$, and so on. The average quality score is defined as:

$$AQS = \frac{\sum_{i=1}^{N} e_i G_i}{\sum_{i=1}^{N} e_i} \tag{2}$$

Since the value, w_R, assigned to R-type staff is equal to zero, their number affects only the denominator in (2). For convenience, the following discussion ignores the part-time distinction, effectively setting all $e_i = 1$.

3.2 The AQS and the Distribution of x

Suppose the distribution function of 'raw scores', x, is denoted $F(x)$. The density function is positively skewed, as shown in Figure 1: for convenience this is drawn for a continuous distribution of x, but the following analysis applies for any distribution.

[13] As discussed above, and again in Section 5.1, the process is actually nonlinear.

Given the distribution, the thresholds, x_i, determine the proportions falling into the different QECs.

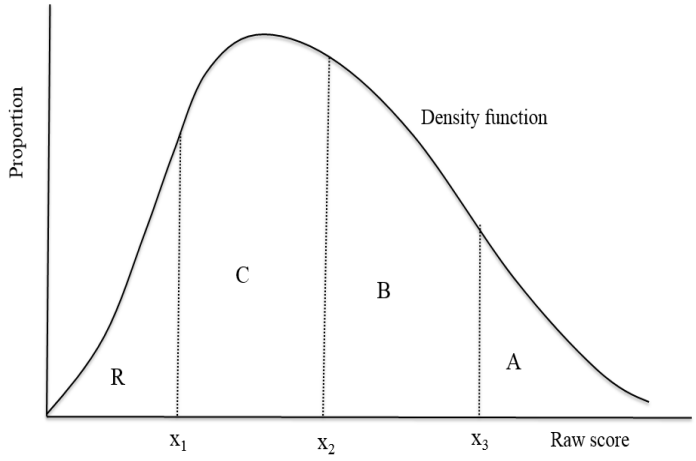

Figure 1: Density Function of Raw Scores

The numbers of R, C, B and A category people are denoted n_R, and so on. If N is the population size, these are given by:

$$n_R = NF(x_1) \tag{3}$$
$$n_C = N[F(x_2 - x_1)] \tag{4}$$
$$n_B = N[F(x_3 - x_2)] \tag{5}$$
$$n_A = N[1 - F(x_3)] \tag{6}$$

As defined above, the weights attributed to each category are w_R, w_C, and so on. With $w_R = 0$, the AQS is given by:

$$AQS = \frac{1}{N}(w_C n_C + w_B n_B + w_A n_A) \tag{7}$$

Alternatively, AQS can be expressed in terms of the distribution function $F(x)$, remembering that $w_R = 0$, using:

$$AQS = w_A - (w_A - w_B)F(x_3) - (w_B - w_C)F(x_2) - w_C F(x_1) \tag{8}$$

3.3 Variations in Weights Attached to QECs

Consider the effect on the AQS of varying the weights, for fixed thresholds. Using 8, the total differential is:

$$d\left(AQS\right) = \frac{1}{N}\left(n_C dw_C + n_B dw_B + n_A dw_A\right) \tag{9}$$

In order to investigate the effect of varying the weights systematically, while considering changes along only one dimension, it is convenient to consider varying the weight, w_A, while at the same time keeping the total, $k = w_C + w_B + w_A$, constant. Starting from an initial position of extreme compression, all weights are equal and the minimum value of w_A occurs where $w_C = w_B = w_A = k/3$. Then w_A is increased while w_C is reduced by the same amount, keeping w_B fixed. When $w_C = 0$, further increases in w_A are achieved by reducing w_B until the position of extreme dispersion of the weights is achieved: this occurs when $w_A = k$ and $w_B = w_C = 0$.

Thus for $dw_A = -dw_C$, and $dw_B = 0$, (9) gives:

$$\frac{d\left(AQS\right)}{dw_A} = -\frac{1}{N}\left(n_C - n_A\right) \tag{10}$$

That is, the relationship between AQS and w_A (as w_A increases, for fixed w_B and k) is a downward sloping straight line with gradient given by $\frac{1}{N}\left(n_C - n_A\right)$. As w_A increases initially from $k/3$, AQS falls linearly while w_B remains fixed. When w_A reaches $w_A = k - k/3 = 2k/3$, then $w_C = 0$ and further increases in w_A are achieved by reducing w_B, which falls from its initial value of $k/3$. The slope of the relationship between AQS and w_A then becomes:

$$\frac{d\left(AQS\right)}{dw_A} = -\frac{1}{N}\left(n_B - n_A\right) \tag{11}$$

The relationship is thus piecewise linear with two segments and a kink at $w_A = 2k/3$. An example is given in Figure 2, showing the piecewise linear schedule ABC.

The relative slopes depend on the relative sizes of n_C and n_B. In the case where $n_C > n_B$, the second segment is flatter than the first segment. The maximum value of w_A is $w_A = k$, with $w_B = w_C = 0$, which is referred to as maximum dispersion. If, instead, $n_C < n_B$, the segment, BC, is steeper than the range, AB. Both alternatives arise for New Zealand universities, as shown in Section 4 below. The implications for the sensitivity of university AQS-based ranks to the weights used is considered in the following subsection.

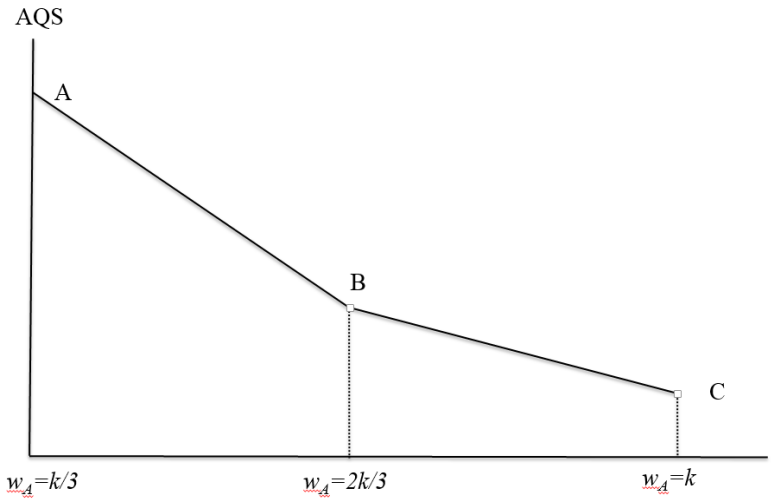

Figure 2: Relationship Between AQS and Weight Attached to QEC of A, w_A

3.4 Changes in Ranks

Consider two universities, where the proportions, $p_A = n_A/N$, $p_B = n_B/N$, and so on, in each category are distinguished using subscripts 1 and 2. As w_A is increased from its lowest value (extreme compression) to its maximum value (extreme dispersion of weights), the values of AQS can change rank at most twice, given the piecewise linear nature of the relationship derived above. In the case where $n_C > n_B$, the necessary condition for this to arise is that $AQS_1 > AQS_2$ and:

$$p_{B,2} - p_{B,1} > p_{A,2} - p_{A,1} > p_{C,2} - p_{C,1} \tag{12}$$

That is, where 1's profile begins above that of 2, it can intersect the first segment from above if $p_{A,2} - p_{A,1} > p_{C,2} - p_{C,1}$ and the second segment from below if $p_{B,2} - p_{B,1} > p_{A,2} - p_{A,1}$. However, this is not a sufficient condition, because the two universities would need to be sufficiently close for the double crossing to occur in the feasible region.

Suppose the middle term in the above set of inequalities is positive, so that $p_{A,2} > p_{A,1}$ and distribution 2 has relatively more type-As than does distribution 1. This can be combined with distribution 2 having a lower value of AQS in the extreme compression case. If, in addition, either $p_{C,2} < p_{C,1}$, or $p_{C,2} > p_{C,1}$ but the difference

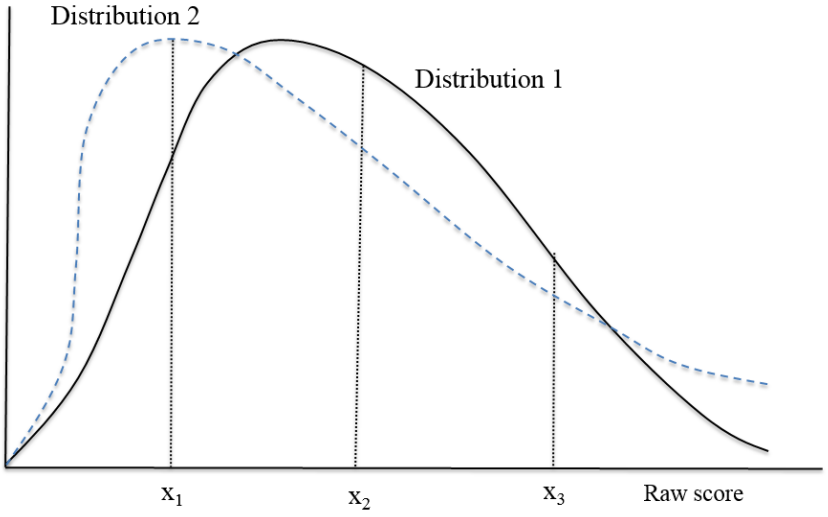

Figure 3: Two Distributions

$p_{C,2} - p_{C,1}$ is sufficiently small for the second inequality to hold, then an intersection can occur (in the relationship between AQS and w_A) between the first pair of linear segments. However, the fact that distribution 2 has proportionately more As than distribution 1, and either relatively fewer Cs (or only a slightly larger proportion of Cs) means that it necessarily has relatively fewer Bs (and relatively more Rs). This means that the first of the inequalities in (12) cannot be satisfied. Hence the second linear segments cannot intersect: they must diverge, with distribution 1 having a lower AQS at the kink point.

The two distributions and respective areas are illustrated in Figure 3. Hence, only one change in ranks is possible as w_A is varied over its feasible range. Similarly, it can be shown that if $n_C < n_B$, there can also be a maximum of only two rank changes between any two universities. The precise sensitivity depends in practice on the actual profiles: these are considered in the following section.

4 The 2012 PBRF Round

This section considers the sensitivity of university rank changes to variations in the weights, using a special dataset provided by the TEC at the request of the authors,

following a confidentiality agreement. This dataset is not publicly available. It includes anonymous raw PBRF data including, for each researcher, an anonymous identifier, age, research discipline, university of employment, and PBRF quality evaluation category for each of the three PBRF rounds in which a researcher's evidence portfolio was submitted.

Table 2: Distribution of QECs by University: 2012

	n_A	p_A	n_B	p_B	n_C	p_C	n_R	p_R	Total
AUT	18	*0.04*	112	*0.24*	280	*0.60*	57	*0.12*	467
Lincoln	19	*0.09*	58	*0.27*	110	*0.51*	27	*0.13*	214
Massey	102	*0.10*	381	*0.36*	510	*0.49*	56	*0.05*	1049
Auckland	310	*0.17*	694	*0.39*	692	*0.39*	83	*0.05*	1779
Canterbury	80	*0.12*	282	*0.42*	279	*0.42*	27	*0.04*	668
Otago	193	*0.15*	541	*0.41*	553	*0.42*	30	*0.02*	1317
Waikato	47	*0.09*	218	*0.42*	214	*0.42*	35	*0.07*	514
VUW	108	*0.16*	336	*0.51*	196	*0.30*	16	*0.02*	656
Total	877	*0.13*	2622	*0.39*	2834	*0.43*	331	*0.05*	6664

Table 2 reports, for each university, the distribution of QEC scores for the 2012 round.[14] The proportions can be used to compute the relevant cumulative proportions, $F(x_1)$, $F(x_2)$ and $F(x_3)$, used in equation (8). The resulting AQS profiles, for variations in w_A (systematically moving from complete compression to extreme dispersion of weights) are shown in Figure 4. Each line is labelled, to avoid excessive clutter, using just the first letter of the university's name (so that A is Auckland, O is Otago, and so on), with the obvious exceptions of Victoria University of Wellington (VUW) and Auckland University of Technology (AUT). In calculating the quality scores, the denominator of (2) includes the number of R-type researchers, following the approach used in 2003 and 2006; the use of alternative denominators is considered in detail by Buckle and Creedy (2017).

Differences in the $AQSs$ for the case of extreme compression (where all weights are equal to 6) arise purely because of the differences in the proportions judged to be type-R researchers. The slopes of the initial linear segments depend on the differences between proportions, p_C and p_A. These differences are largest for the lower-scoring

[14] The final column in the table corresponds to the final column of Table 2 of Buckle and Creedy (2017, p. 10).

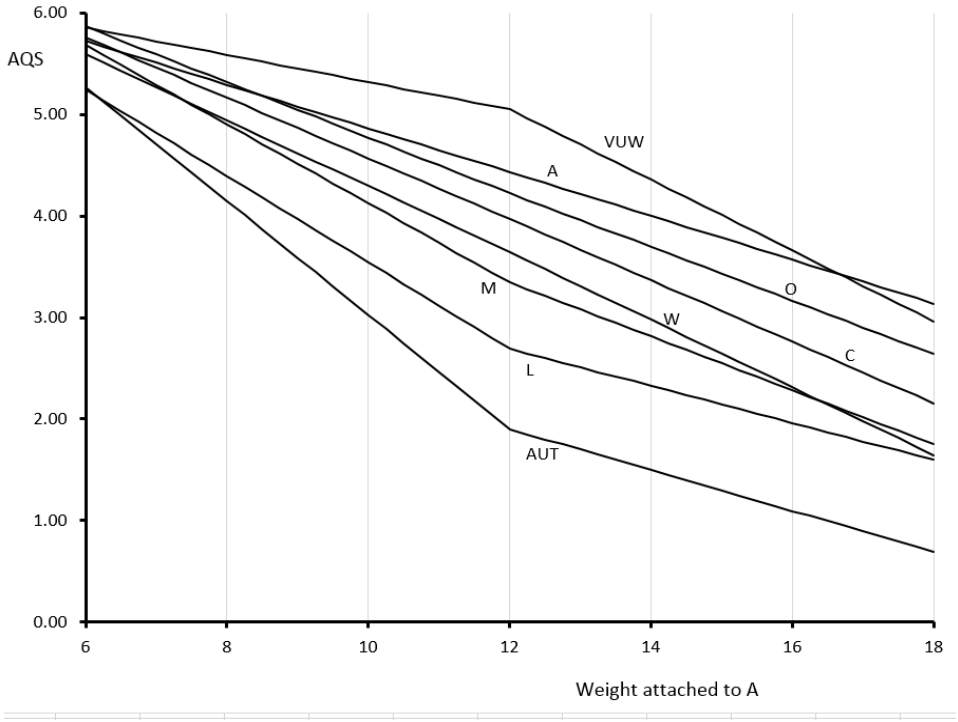

Figure 4: Variation in AQS with w_A: PBRF 2012

universities (measured according to the actual weights used by the TEC). For this reason, the profiles tend to diverge as w_A increases.

From Table 2, it can be seen that $p_C > p_B$ for AUT, Lincoln and Massey, so that the relevant profiles are convex. The proportions of C and B researchers are similar for Auckland, Canterbury, Otago and Waikato, so their profiles are close to being linear over the whole range. For VUW, $p_C < p_B$, so that its profile is concave. At the 'midpoint', where $w_A = 12$, VUW has the highest AQS, while AUT, Lincoln and Massey have the lowest AQSs. For this reason the profiles tend to converge as w_A increases beyond the kink where $w_A = 12$. As the weights are varied, unsurprisingly there is considerable variation in the value and relative size of university AQSs as the weights are varied. Nevertheless, within a wide range of weights there are no rank changes other than at the extreme values.

The lack of sensitivity is also illustrated in Figure 5,which plots the Spearman rank correlation coefficient against the weight assigned to A, with other weights adjusted as

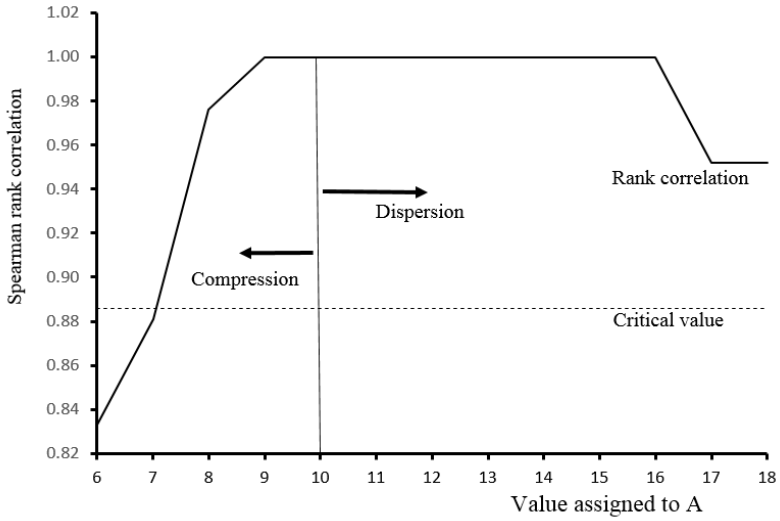

Figure 5: Rank Correlation Between TEC Ranks and Alternative QEC Weighting Systems

described above: only integer values of weights are considered. The ranks derived using alternative weights are perfectly correlated with those derived using the TEC weights, over a large range of dispersed weights and a smaller range of compressed weights. The range of weights assigned to QEC grade A between 9 and 16 gives exactly the same ranking as the TEC method: this range corresponds to the distance between the intersections shown in Figure 4.

In view of the considerable reputational value attached by universities to their rank position, any alteration in the rank order is regarded as very important. It is therefore perhaps not appropriate in this case to apply conventional significance levels to tests. Nevertheless, the horizontal dashed line in the figure gives the critical value for 2.5 per cent level of significance, of 0.886, for 6 degrees of freedom.[15]

5 Research Quality and Raw Scores: 2003 and 2012

Previous sections have concentrated on the production of AQS measures and the role of Quality Evaluation Categories, which in turn are based on the distribution of raw

[15]The critical value for a significance level of 5 per cent is 0.829. Hence at this level, the correlation coefficients for all of the weights in Figure 1 exceed the critical value.

scores, x, as defined by equation (1), and obtained as a weighted aggregate of three separate scores. This section turns to an analysis of the distribution of the raw scores, using the anonymous data provided by the Tertiary Education Commission. In fact, as explained earlier, three stages are used in the determination of the QEC. First, 'indicative scores' are produced by the appropriate discipline peer reviewers. Second, these initial scores are reviewed by the discipline panels, and adjustments are made to produce what is called a 'final score'. However, the category assigned to each individual is not simply obtained using a mechanical allocation of individuals to ranges of the distribution of final scores. At this third stage, individuals can be assigned to a different category so that, for example, some people with a score of $x < 600$ are nevertheless judged to be A-type researchers. The dataset provides details of the indicative and final scores, along with the assigned QEC. This section reports results for 2003 and 2012 only; the information for 2006 is affected by the fact that this was a 'partial round'.[16]

Subsection 5.1 examines the distribution of alternative score measures, and the adjustments made at each stage. The use of QECs, by neglecting information about the distribution of x values within ranges, limits the range of summary measures of research quality that could be used to compare disciplines and universities. Subsection 5.2 uses the raw scores to compare a number of alternative summary measures and the impact on the research ranking of New Zealand universities.

5.1 The Distribution of Research Quality Measures

Figure 6 shows the frequency distribution of indicative scores for all individuals who submitted EPs in 2003. However, the very large number of 1,817 non-administrative staff submitting an evidence portfolio who were given a zero score are not shown, for the obvious reason that it would dominate the diagram. Indeed, as shown by Buckle and Creedy (2017), the major improvement in AQSs from 2003 to 2012 was associated

[16] Ministry of Education (2012, p. 11, n. 2) describe this as follows. 'Researchers did not have to participate in the 2006 round. The Working Group had recommended that a QE three years after the first would help to ensure a managed transition, to develop good practices in performance evaluation, and to acknowledge the need to learn from experience after the first QE. In reviewing the 2003 QE, it was determined that a full round was not necessary. Instead, the partial round assessed staff who had not been assessed in 2003, staff who wished to be reported under a different subject area, and staff who wished to be reassessed in the hopes of achieving a higher quality rating.'

with the exit of a large number of R-type staff. Furthermore, among remaining Rs submitted by universities in 2012, there is a complete absence of zero scores.

The first point that is evident from Figure 6 is that the distribution looks nothing like the kind of distribution reported in many measures of research performance. It is typically found that research metrics have a positively skewed distribution, the most common form being well approximated by the lognormal distribution. The rationale for such a distribution, stemming from the seminal work of Shockley (1957), is that measured performance depends multiplicatively on a range of qualities.[17]

It has been shown above that the distribution of x is likely to have unusual characteristics, given the way it is constructed, and of course there is a maximum attainable score of 700, which affects the right-hand tail. Secondly, it is clear from the spikes at the values, 200, 400 and 600, that the indicative scores were already, at the first stage, influenced by the existence of the thresholds. A different distribution would arise if panel members were not consciously thinking of judging x in relation to threshold values for the eventual determination of a QEC.

Figure 7 shows the differences in 2003 between frequencies at each indicative and final score, for given indicative scores. This second stage clearly involves a substantial number of revisions to the scores. The frequencies just below the crucial threshold values fall substantially, while the numbers of researchers placed at each of the thresholds, 200 and 400, increase substantially.[18] However, within the range $x \geq 600$, the largest increases are of frequencies in the middle range.

In the third stage, some individuals were moved across the thresholds: as explained above, the QEC was not based rigidly on the 'final raw score'. However, no explicit change in the score was reported, so in considering a truly final score that corresponds with the category awarded, those people who are found to be moved into a higher category have simply been given the relevant threshold value. The resulting differences

[17] Models of the genesis of lognormal distributions are examined by Aitchison and Brown (1957). For a flavour of the literature on productivity and research metric distributions see, for example, Allison and Stewart (1974), Cortes *et al.* (2016), Heckman and Sattinger (1991), Moreira *et al.* (2015), Perc (2010), Roy (1950), Ruocco *et al.* (2017), and Swan *et al.* (1999). However, some authors report evidence of a fatter upper tail than the lognormal distribution.

[18] The sum of the differences between frequencies, when adding over the whole range of indicative scores, do not add to zero. The sum is positive, because there are some missing values in the distribution of indicative scores. These arise where agreement could not be reached by the appropriate subject panel members at the initial stage.

are illustrated in Figure 8. Clearly the major concern of the panels at this stage is to move people upwards into higher categories, so that there are no frequency increases at non-threshold values.[19] In some cases this involves a substantial effective increase in the score. Not all upward shifts come from people who are adjacent to the threshold, though this is the most common source of movements.

Similar information is provided for the 2012 round in Figures 9, 10 and 11. In comparing the figures for different years, it should be recognised that the vertical axes have different scales. A comparison of Figures 6 and 9 shows that the spikes at all the thresholds are much more prominent in 2012, particularly at the bottom and top thresholds. This suggests perhaps that the awareness of the importance of these thresholds was much greater in 2012. Figure 10 shows too that in 2012 many more people were moved to a higher threshold value in the second round, compared with that carried out in 2003. The extent of such revisions determined by the thresholds at the very last stage of assigning the QECs was also greater in 2012, particularly at the higher ranges. The 2012 distribution also reflects the rise in the quality of researchers over the period, particularly the higher number at the top end of the scale, the reduction in the number of R-researchers and, as mentioned above, there are no $x = 0$ values in 2012.[20] However, at this final stage, the patterns of *adjustment* in 2003 and 2012 are very similar.[21]

5.2 Alternative Measures of Quality using Raw Scores

The information about QECs and associated scores, x, can be used to examine the implications of using alternative summary measures of research quality for each university. Figure 12 provides information about research quality in 2012 based on four measures. The universities are ordered from left to right in terms of their 2012 AQS published

[19] For 2003 there is an increase by one portfolio in the range below 600, but this is associated with a portfolio without a final raw score. This arose because the panel could not agree at that stage, so no score was recorded.

[20] The 2012 round also provided greater opportunities to avoid submitting portfolios by those expected to be judged as R-researchers.

[21] The correlation coefficient between the differences in frequencies between QEC and final score in 2003 and 2012 is high, at 0.93. However, the correlation coefficient between the differences in frequency between final and indicative scores in these years is much lower, at 0.37.

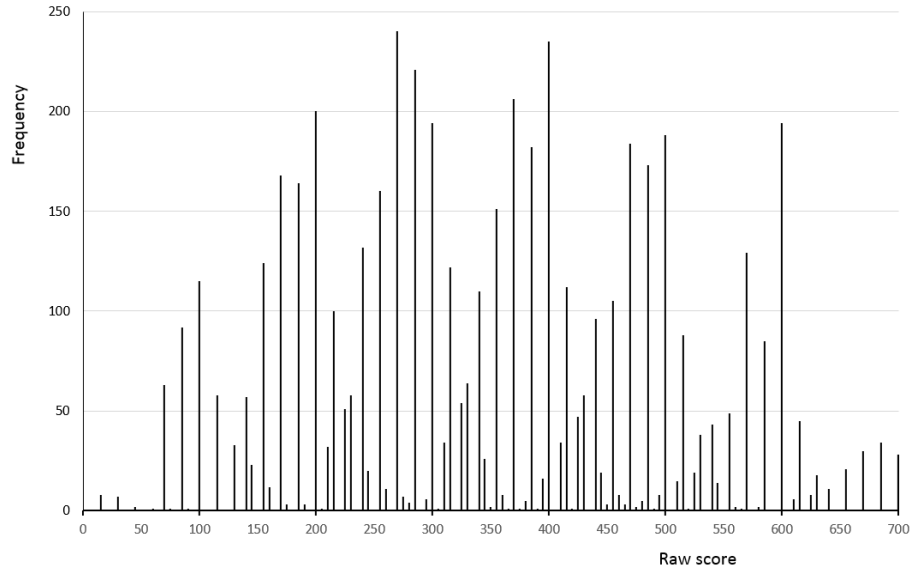

Figure 6: Distribution of Indicative Scores 2003: All Universities Combined

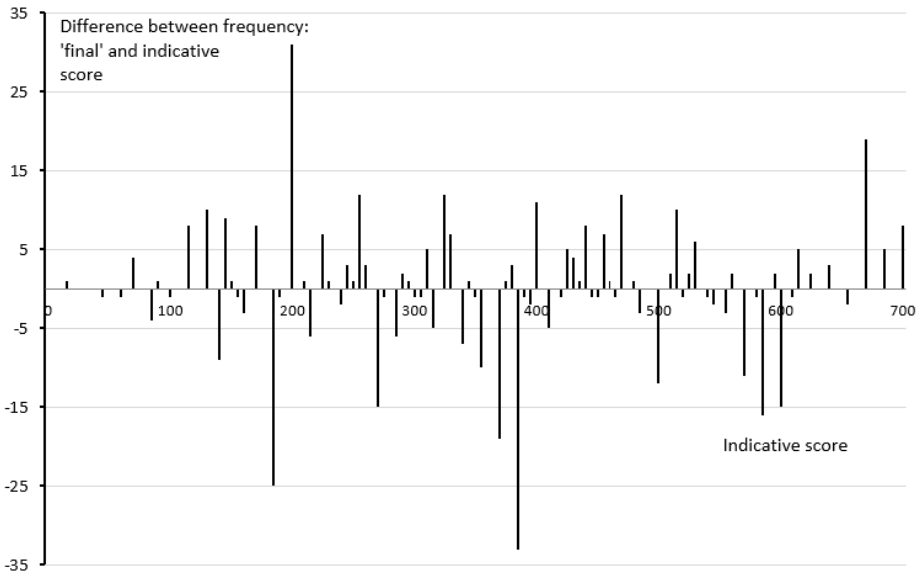

Figure 7: Difference Between Frequencies of 'Final' and Indicative Scores 2003: All Universities Combined

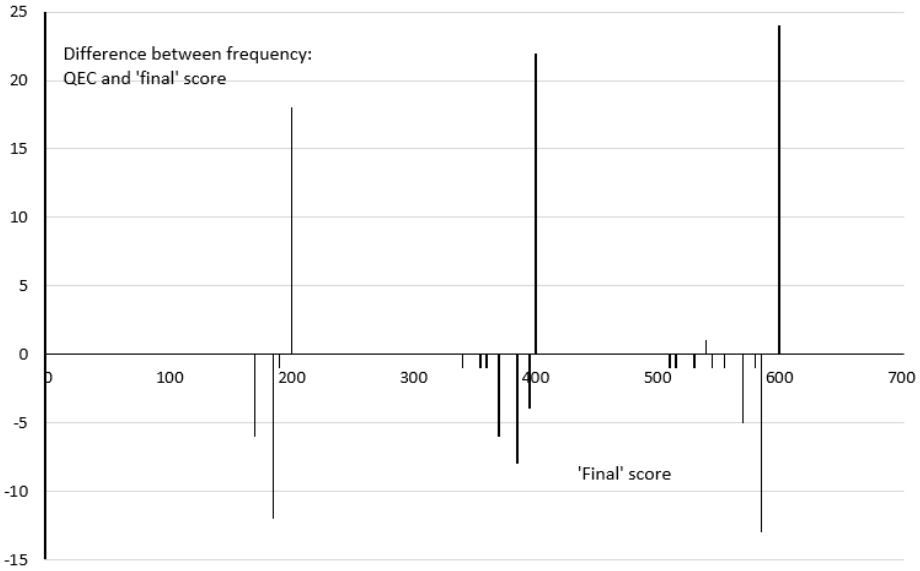

Figure 8: Difference Between Frequencies of QEC-adjusted and 'Final' Scores 2003: All Universities Combined

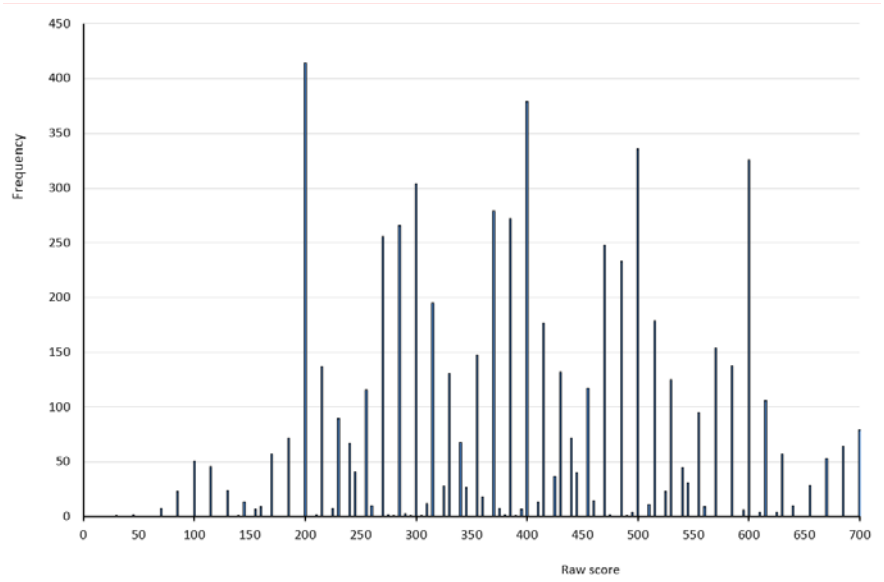

Figure 9: Distribution of Indicative Scores 2012: All Universities Combined

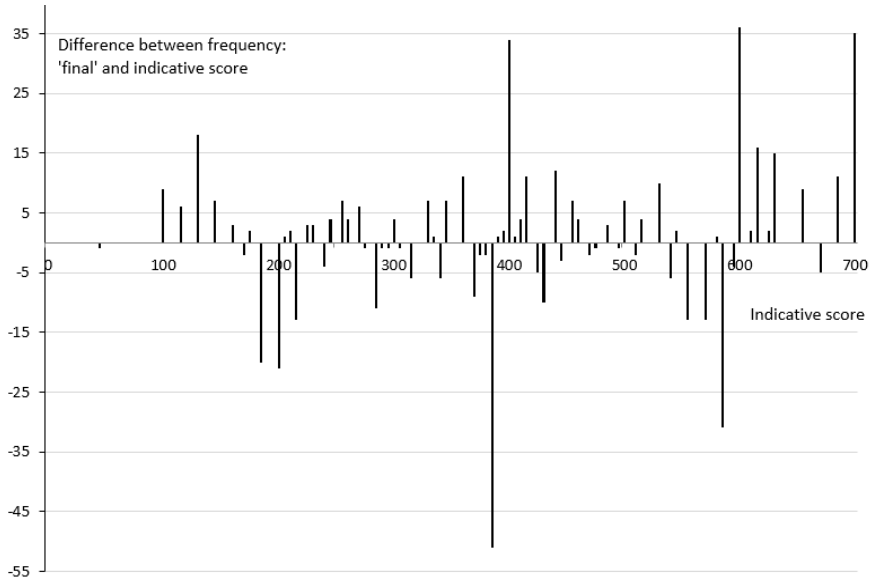

Figure 10: Difference Between Frequencies of 'Final' and Indicative Scores 2012: All Universities Combined

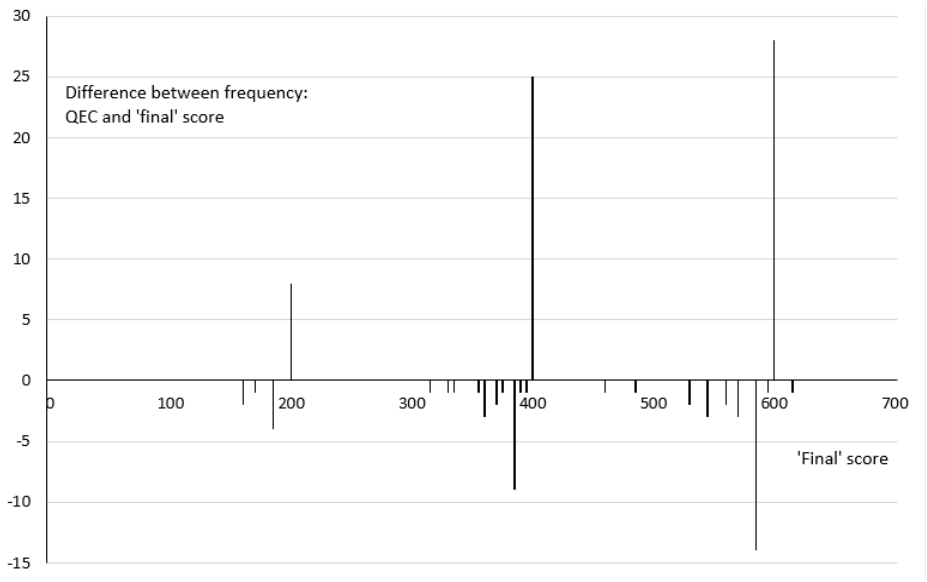

Figure 11: Difference Between Frequencies of QEC-adjusted and 'Final' Scores 2012: All Universities Combined

rank.[22] Hence a profile in the figure that is consistently downward sloping from left to right produces a ranking that is consistent with the reported outcome in New Zealand Tertiary Education Commission (2013). In order to show AQS values and arithmetic mean values of x in the same diagram, the former were arbitrarily multiplied by 100. In this figure the absolute differences between AQS and average measures are of no relevance: the only concern is with the slope of the relevant profile to see if different rankings would be produced.

The top profile in Figure 12 is for the AQS obtained using the number of portfolios submitted as the denominator in taking the average.[23] The solid line showing values of the arithmetic mean 'final score', when using the number of all submitted portfolios as denominator (that is, including R researchers), also declines monotonically, showing that it produces the same ranks as the TEC report. However, this profile is somewhat flatter than that for the AQS.

The lower two lines in Figure 12 are for arithmetic mean and AQS values obtained when the total number of non-administrative staff is used for the relevant denominator. The differences between these two profiles and the two discussed in the previous paragraph are therefore influenced by the different proportions of total staff who submitted portfolios. The fact that this proportion is very high for Canterbury University, relative to other closely-ranked universities, explains the upward movement in the profiles, so that the rank of Canterbury is higher using this approach; it becomes the second-highest ranked university when using the AQS method based on QECs, and the top-ranked university when using the arithmetic mean of the raw scores for all non-administration staff. Hence, using AQSs and all non-administration staff as the denominator, the rank order of the top four New Zealand universities is: VUW, Canterbury, Auckland, Otago. Using the arithmetic mean of the raw scores for all non-administration staff the rank order is: Canterbury, VUW, Auckland, Otago.

Further information about the distribution of 'final' scores is provided by Figure 13. The relative closeness of the median and arithmetic mean profiles reflects the fact, as seen above for all universities combined, that the distributions are not clearly skewed: this carries over to distributions for each individual university. Again, the downward

[22] The precise values are taken from Buckle and Creedy (2017), where no employment weights are used. However, the rankings are exactly the same as reported in New Zealand Tertiary Education Commission (2013).

[23] These values are taken from Buckle and Creedy (2017, Table 1, p. 9).

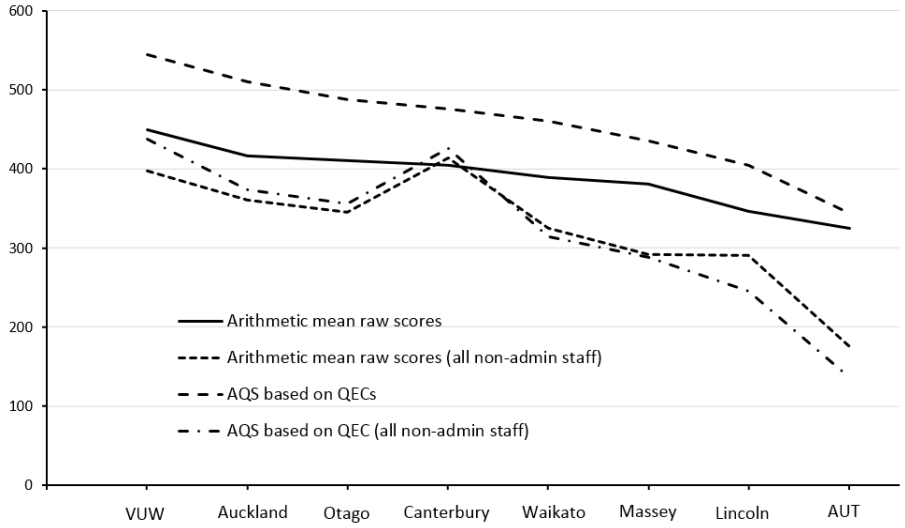

Figure 12: Alternative Quality Measures for Each University: 2012

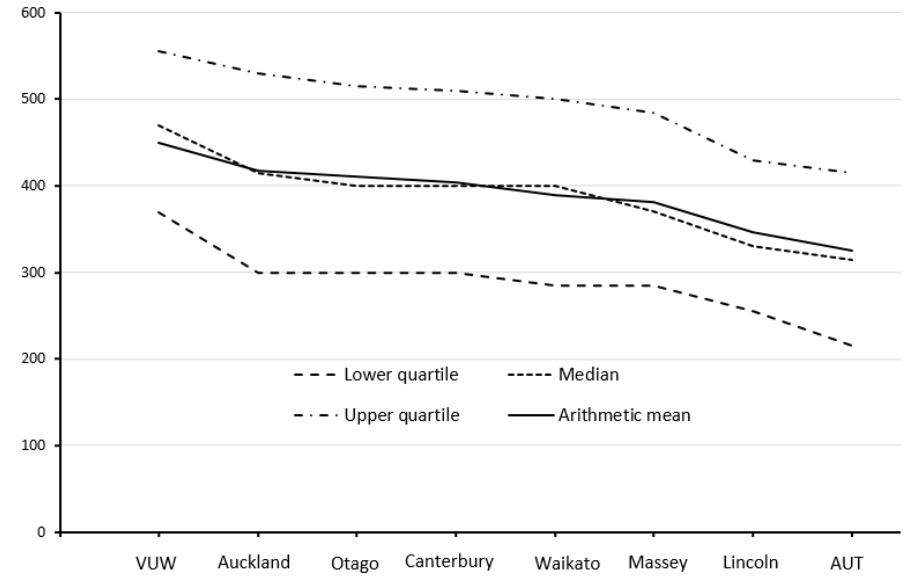

Figure 13: Summary Measures of Distribution of Raw Scores for Each University: 2012

sloping nature of these two profiles suggests that the ranking is not generally affected, compared with the published ranks. An exception is that the use of the median gives a slight improvement for Waikato. The flatness of the profiles suggests that, using these measures of location of the distribution of x values, there is little to choose between the middle-ranked universities (Auckland, Otago, Canterbury, Waikato and Massey). The lower and upper quartiles are evenly spaced around the median, again reflecting the rough symmetry in the distributions.

6 Conclusions

This paper has examined the metrics used by the New Zealand Tertiary Education Commission to rank the research quality of New Zealand universities following each PBRF round. It was shown that the derivation of the AQS for each university in any round critically depends on the weights assigned to each Quality Evaluation Category (QEC) and the distribution of portfolios. The scoring method used to determine the raw score for a research portfolio, and in turn the QEC assigned to it, also has an effect on the distribution of scores and hence QECs and the average quality scores. Furthermore, the scope of the PBRF assessment and hence choice of denominator used in any summary measure is likely critically to influence the summary measures and rankings.

Examination of the distribution of raw scores in the 2003 and 2012 PBRF rounds, and movements in the raw scores during the expert panel assessments stages, indicates that the use of QECs and associated threshold scores appear to have influenced the performance scores assigned to research portfolios and the final allocation of portfolios to Quality Evaluation Categories. This has consequences for the form of the distribution of research quality. In particular, the process has generated large spikes in the distribution of raw scores at the threshold values for QECs. Furthermore, the distributions of raw scores are roughly symmetric. These features contrast strongly with the distribution of metrics used by other performance evaluation methods.

Without a clear rationale for the use of QECs and the selection of threshold values, the choice of weights assigned to each quality category also appears arbitrary. Yet, these weights and the distribution of scores can influence relative university AQSs and, in principle, the research quality ranking of universities and disciplines. Moreover, this

system used by the TEC to summarise research quality discards information about the relative quality of researcher portfolios which could be used to derive alternative summary measures of researchers' quality.

Several alternative summary measures were derived in the paper, using the raw scores for researcher portfolios. These alternative measures included the arithmetic mean, median and upper and lower quartile values. In 2012, the university ranks using these alternative measures are the same as those obtained using the AQSs derived by the Tertiary Education Commission. However, there is very little difference between the four middle-ranked universities for this set of alternative measures. The choice of all non-administration staff as the denominator for the QEC-based and raw-score based arithmetic means was found to generate a change in rankings in 2012. Furthermore, it is not possible to know what distribution of raw scores would have arisen without the review panelists' prior knowledge of the use of QEC thresholds. An alternative method, using QECs, would involve the independent evaluation of raw scores, followed by the determination of thresholds. This is still subject to the criticism that the use of QECs compresses all scores within a given range to a single value.

Although there are several positive features of the PBRF process, including the use of a peer review process and discipline panels, the findings in this paper raise serious questions regarding the merits of the metrics applied in the PBRF process to assess the research quality of New Zealand universities. In addition, it has been found elsewhere that the main transformation in the average quality of New Zealand university research came through reductions in non-research-active faculty (Rs), and this rate of transformation is not likely to be sustainable over a long period: see Buckle and Creedy (2017).

When, in addition, the high administrative and compliance costs associated with the process are recognised, a debate needs to take place regarding the metrics used and scope of future PBRF processes. Although some assessment is needed to allocate public funds and to generate the appropriate incentives facing universities in the use of those funds, the research quality in New Zealand universities is different from that which prevailed in the early 2000s when the current scheme was devised and introduced. Furthermore, since the introduction of the PBRF there has been a substantial accumulation of knowledge from experiences of a wide range of measurement approaches

used in other countries.[24] This includes information about the properties of alternative metrics and and the incentive effects of different shemes, including their unintended and often unanticipated consequences. A more-informed debate can now take place regarding the design of future schemes to monitor the use and allocation of public funds by universities.

References

[1] Aitchison, J.A. and Brown, J.A.C. (1957) *The Lognormal Distribution with Special Reference to its Uses in Economics*. Cambridge: Cambridge University Press.

[2] Allison, P.D. and Stewart, J.A. (1974) Productivity differences among scientists: evidence for accumulative advantage. *American Sociological Review*, 39, pp. 596-606.

[3] Buckle, R.A. and Creedy, J. (2017) The evolution of research quality in New Zealand Universities as measured by the Performance-Based Research Fund process. *Victoria University of Wellington Chair of Public Finance Working Paper*, no. WP11/2017.

[4] Cortés, L. M., Perote, J. and Andrés, M-V. (2016) The productivity of top researchers: A semi-nonparametric approach. *Universidad EAFIT, Centro de Investigacion Economicas y Financieras Documentos de trabajo working paper*, no. 16-05.

[5] de Boer, H. Jongbloed, B. Benneworth, P. Cremonini, L. Kolster, R. Kottmann, A. Lemmens-Krug, K. and Vossenstey, H. (2015) Performance-based funding and performance agreements in fourteen higher education systems, *Report for the Ministry of Education, Culture and Science, Centre for Higher Education Policy Studies*, C15HdB014, *Universiteit Twente, Enschede*. www.utwente.nl/cheps

[6] Heckman, J.J. and Sattinger, M. (1991) Introduction to the Distribution of Earnings and of Individual Output, by A.D. Roy. *Economic Journal*, 125, pp. 378-385.

[24] Reference has already been made to de Boer *et al.* (2015), Wilsden *et al.* (2015) and OECD (2010), both of which provide extensive bibliographies.

[7] Ministry of Education (2012) A History and Overview of the PBRF. Wellington: Ministry of Education. Available at: https://education.govt.nz/assets/Documents/Further-education/Policies-and-strategies/Performance-based-research-fund/PBRFHistoryAndOverview.pdf

[8] Moreira, J.A.G., Zeng, X.H.T. and Amaral, L.A.N. (2015) The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model. PLoS ONE 10(11): e0143108. doi:10.1371/journal.pone.0143108.

[9] New Zealand Tertiary Education Commission (2002) *Investing in Excellence: The Report of the Performance-Based Research Fund Working Group*. Wellington: Ministry of Education and Transition Tertiary Education Commission.

[10] New Zealand Tertiary Education Commission (2013) *Performance-Based Research Fund: Evaluating Research Excellence – the 2012 Assessment*. Wellington: Tertiary Education Commission.

[11] OECD (2010) Performance-Based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings. OECD Publishing. Available from: http://www.oecdilibrary.org/education/performance-based-funding-for-public-research-intertiary-education-institutions_9789264094611-en

[12] Perc, M. (2010) Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics*, 4, pp. 358–364.

[13] Roy, A.D. (1950) The distribution of earnings and of individual output. *Economic Journal*, 60, pp. 489–505.

[14] Ruocco, G., Daraio, C., Folli, V. and Leonetti, M (2017) Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent. *Palgrave Communications*, July (DOI:10.1057/palcomms.2017.64).

[15] Shockley, W (1957) On the statistics of individual variations of productivity in research laboratories. *Proceedings of the Institute of Radio Engineers*, 45, pp. 279-290.

[16] Swan, J.E., Powers, T.L. and Bos, T. (1999) The Concentration of Career Research Productivity among Marketing Academicians: Implications for Faculty Evaluation. *Marketing Education Review*, 9, pp. 39-50 (DOI:10.1080/10528008.1999.11488659)

[17] Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J. and Johnson, B. (2015) *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.* (DOI: 10.13140/RG.2.1.4929.1363).

About the Authors

Robert Buckle is Professor Emeritus at Victoria Business School, Victoria University of Wellington
Email: bob.buckle@vuw.ac.nz

John Creedy is Professor of Public Finance at Victoria Business School, Victoria University of Wellington
Email: john.creedy@vuw.ac.nz

VICTORIA UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

Working Papers in Public Finance