Construct validity maps and the NIRF 2019 ranking of colleges

Gangan Prathap*

A.P.J. Abdul Kalam Technological University,
Thiruvananthapuram 695 016, India

In this study, we prepare construct validity maps from the National Institutional Ranking Framework (NIRF) 2019 data for the top 100 colleges in India. Tamil Nadu, Delhi and Kerala together have a disproportionate 82% share of the top-ranking colleges in the country that participated in the 2019 exercise. The higher education system in India comprises about 52,000 units of assessment from universities, premier institutes of technology and colleges to stand-alone institutions, and many participate in the NIRF exercise. The NIRF score is computed from five broad parameters, of which one is a peer review-based perception score for participating institutions. Using its teaching, learning and resources parameter as a proxy for teaching and learning resources input and its research and professional practices and graduation outcomes parameters as proxies for teaching and research outputs or outcomes, we also compute a quality or excellence proxy and from this compute a second-order X-score. The three scores, NIRF, perception and X are used in the context of construct validity to construct two-dimensional maps to determine how the top colleges are placed with respect to each other. A quantitative estimate is obtained using Peirce's measure of predictive success to determine if the use of one construct measure to predict another is acceptable or not. In terms of the construct validity paradigm, we are able to recognize possible biases in the peer review perception scores and also recommend that the X-score, which is based on an input–output model, may give a better representation of reality.

Keywords: Bibliometrics, construct validity, institutional ranking, research evaluation.

As reported in a survey[1] there are 864 universities, 40,026 colleges and 11,669 stand-alone institutions in India, making a total of more than 52,000 units of assessment. The higher education sector is so fragmented that few institutions have the critical mass to deliver high-end research along with high-quality education.

Prathap[2] used bibliometric and economic data from the National Institutional Ranking Framework (NIRF) 2017 rankings to show that the two top-ranking colleges if only research excellence is considered are both from one state: Loyola College, Chennai and Bishop Heber College, Tiruchirappalli in Tamil Nadu. The NIRF 2019 results have now been released. Among the top 100 colleges in India, Tamil Nadu (35), Delhi (29) and Kerala (18) now account for 82% of this elite listing (in 2018 this was even higher at 85%). Among the more than 52,000 units of assessment in the higher education system in India, many leading and aspiring institutions now participate in the NIRF exercise. The NIRF score is computed from five broad parameters, of which one is a peer review-based perception score for participating institutions. Using its teaching, learning and resources (TLR) parameter as a proxy for teaching and learning resources input, and its research and professional practices (RPC) and graduation outcomes (GO) parameters as proxies for teaching and research outputs or outcomes, we also compute a quality or excellence proxy and from this compute a second-order X-score[2,3]. The three scores, NIRF, perception and X are used in the context of construct validity (CV) to construct two-dimensional maps to determine how the top colleges are placed with respect to each other. The CV maps enable us to recognize possible biases in the peer review perception scores and also recommend that the X-score, which is based on an input–output model, may give a better representation of reality. A quantitative estimate is obtained using Peirce's measure of predictive success[4] to determine if the use of one construct measure to predict another is acceptable or not.

The NIRF has just released its 2019 ranking of higher educational institutions across India (https://www.nirfindia.org/2019/Ranking2019.html). It arrives at a single NIRF score using scores from five broad generic groups: TLR – Teaching, learning and resources; RPC – Research and professional practices; GO – Graduation outcomes; OI – Outreach and inclusivity and PR – Perception.

This five-dimensional model is developed from sub-heads with weights assigned to each broad head, and also to the sub-heads within each head. For each sub-head, a score is generated using suitably proposed metrics, and the sub-head scores are then added to obtain scores for each individual head. The overall score is computed based on the weights allotted to each head. This score can take a maximum value of 100. Thus, a multi-dimensional input and output problem is compressed into a single score and institutions, irrespective of size or resources, are finally rank-ordered based on these scores.

How valid is this score? This is a difficult question to answer, as we are dealing with the issue of validity in a complex social system. There is indeed no such thing as an independent ground truth. Perhaps, the closest we have here is the perception score, which is an observed variable and not a latent variable emerging from a mathematical model. In the NIRF operationalization, this is included as an input to get a final NIRF score, although with a very low weighting, i.e. 10%. From the CV point of view (an elaboration of which follows in the next section), it is more meaningful to use the perception score as a baseline with the NIRF score to get a better appreciation of the

*e-mail: gangan_prathap@hotmail.com

various biases, prejudices and mis-measures involved[5]. It is possible to use an argument based on separating input scores from output and outcome scores and size-independent from size-dependent scores[2,3], to obtain second-order X-score as an independent measure of performance. We now have three scores, the perception score, the NIRF score and the X-score which can be compared in a three-way CV exercise using CV maps. Note that the perception score is an observed variable and can be interpreted as a 'ground truth' (however faulty it is; this will become evident when the CV maps are drawn) against which the X-score and the NIRF score will be validated. In the language of psychometrics, the X-scores and the NIRF scores are latent variables emerging from the mathematical models. Peirce's quantitative measure of predictive success helps determine whether to see if the use of one construct measure to predict another is acceptable or not[4].

For each institution, category-wise, the NIRF makes available a set of five parameters. Thus, as an example, for Presidency College, Chennai, we have TLR = 77.49, RPC = 98.02, GO = 46.40, OI = 50.56 and perception = 56.50. For each indicator, the maximum score possible is 100. Note that from an econometric and scientometric evaluation protocol[3,6], TLR is an input score related to teaching and learning resources, while RPC and GO are output or outcome scores related to research and teaching performance. All three are used to compute the X-score. OI is only of sociological relevance and is not considered here for computing the X-score. Both OI and perception are used along with TLR, RPC and GO to compute the NIRF score.

We examine data for the top 100 institutions in the 'Colleges' category. In each case, we treat the TLR term as a single input, and RPC and GO as two-dimensional output terms. We use a totalization strategy[7], so that for each ith institution we get a normalized input term

$$I(i) = \text{TLR}(i)/\text{SUM}(\text{TLR}(i), i = 1100), \quad (1)$$

and a normalized output term

$$O(i) = (\text{RPC}(i)/(\text{SUM}(\text{RPC}(i),$$
$$i = 1100) + \text{GO}(i)/(\text{SUM}(\text{GO}(i), i = 1100))/2. \quad (2)$$

The quality term in each case is

$$q(i) = O(i)/I(i). \quad (3)$$

Note the significance of the input term appearing in the denominator. This further implies that $q = 1$ is the norm or average performance of all the top 100 institutions taken together. The second-order X-score is simply $X(i) = q(i)O(i) = q(i)^2 I(i)$. This serves as a single-valued scalar measure for the research and teaching performance of each institution and is a second-order exergy term

from the input, output and quality (excellence) indicators[8], $X = q^2 I = qO$. This exercise is done for the top 100 colleges in the NIRF 2019 ranking. In order to facilitate comparison on a 0–100 scale, the X-scores are rescaled so that the highest-scoring institution (Presidency College, Chennai) is assigned a score of 100. The highest-scoring college based on the NIRF score is Miranda House, New Delhi.

A brief note about CV will be meaningful here. It can be defined as the extent to which an operationalization measures the concept it is supposed to measure[9,10]. The NIRF scores are intended to measure and comparatively rank the performance of higher educational institutions through a complex operationalization. For each institution, the NIRF collects and collates a large number of objective metrics from both the input (e.g. TLR data) and output sides (e.g. RPC and GO data), and also subjective metrics (e.g. the perception score). The NIRF exercise then reduces these multi-faceted input and output terms into a single number called the NIRF score. When multi-dimensional scores are thus compacted, there is always the danger of a single story[2].

In this study, instead of a single score, we have three measures of constructs to evaluate research and teaching excellence in colleges. One is based on peer review (perception) and is largely subjective in nature, while the others (NIRF and X) are more objective in nature and based on publication and citation data (RPC) and GO. We need to ascertain if any two similar constructs correspond with each other. This can be done by a process called convergent validity[5,11]. One construct is used as the baseline against which the other is compared. This is necessary because in the socio-cultural universe, unlike the physical universe, we do not have a ground truth. Bornmann et al.[5] used peer assessment as the baseline, but here we shall proceed with a three-way comparison in each category: NIRF score–perception, NIRF score–X-score and X-score–perception. For each of the three leading states, we shall draw CV maps in addition to using correlation coefficients. Departure from a $y = x$ line is seen as the absence of evidence of convergent validity. The CV maps enable us to recognize possible biases in the peer-review perception scores and also recommend that the X-score, which is based on an input–output model, may give a better representation of reality.

For the top 100 colleges in the NIRF 2019 ranking, perception scores range all the way from 0 (Nirmalagiri College, Kannur, Kerala) to 100 (Loyola College, Chennai). The NIRF scores are telescoped into a narrow band and vary from 49.03 (Ramanujan College, Delhi) to 73.72 (Miranda House, Delhi). The X-scores range from 3.23 (Government Arts College, Thiruvananthapuram, Kerala) to 100.00 (Presidency College, Chennai). Note that each of these three scores can be interpreted as a 'ground truth' from the CV point of view[5,11]. While the perception and NIRF scores range freely from 0 to 100,

Table 1. Pearson's correlation for the three construct validity (CV) measures of colleges from the three leading states in the NIRF top 100

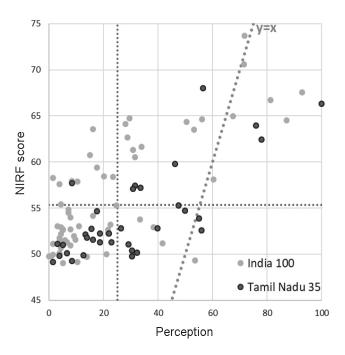| Pearson's correlation		Perception	NIRF score	X-score
Tamil Nadu 35	Perception	1.00	0.77	0.52
	NIRF score	0.77	1.00	0.66
	X-score	0.52	0.66	1.00
Delhi 29	Perception	1.00	0.85	0.33
	NIRF score	0.85	1.00	0.65
	X-score	0.33	0.65	1.00
Kerala 18	Perception	1.00	0.41	0.00
	NIRF score	0.41	1.00	0.28
	X-score	0.00	0.28	1.00

Figure 1. Construct validity map between National Institutional Ranking Framework (NIRF) scores and perception scores for the top 100 colleges in India and 35 from Tamil Nadu.

this is not seen for the NIRF scores which ranges over a narrow band, as reported earlier[2]. The three scores can be 'validated' against each other, leading to a three-way representation. One protocol for validation is to compare correlation coefficients[5,11]. Table 1 shows Pearson's correlations for the three CV measures of colleges from the three leading states in the NIRF top 100. Agreements vary with correlation coefficients ranging from 0.85 (Delhi colleges between the NIRF scores and perception scores) to 0 (Kerala colleges between the X-scores and perception scores), showing biases in the perception judgements and the possible weaknesses in the NIRF score computation. The perception bias positively favours Delhi colleges and is heavily biased against Kerala colleges.

To understand this visually, in Figures 1–3, we see the three possibilities for the 35 colleges from Tamil Nadu. Instead of relying only on correlation coefficients, we can get good visual appreciation by marking on each map the $y = x$ line (represented by green dotted lines) that represents a perfect agreement between the two scores. The red dotted lines represent the mean scores in each case and help divide the map into four quadrants. The top right-hand quadrant is the true positive and the bottom left-hand quadrant is the true negative, where the x-axis is taken as the predictor for the observed y-axis. The top left-hand quadrant and the bottom right-hand quadrant are those associated with the type-I and type-II errors (i.e. false negative and false positive cases respectively) in a convergent validity exercise, if this is indeed needed[5,11]. This exercise was repeated for the 29 colleges from Delhi and the 18 colleges from Kerala (map not shown here). The nine maps show that the NIRF scores are telescoped into the 45–75 range, while the perception scores and X-scores span the whole range from 0 to 100. In Figure 1, we can take either the perception score or the NIRF score as the ground truth, and validate one against the other. We find that the perception scores considerably underestimate performance at the lower end of the spectrum where most of the institutions are placed. Particularly, colleges from Kerala are the most affected by perception bias. Conversely, one can also argue that the NIRF scores considerably overestimate performance at this end of the spectrum. A similar picture emerges from Figure 3; the NIRF scores favour the low-performing institutions by assigning higher scores than computed through the X-scores model. Many institutions from Delhi are prominently outliers in a type-I or type-II sort of way – the perception score in an egregiously false positive way favours them compared to an X-score evaluation and are unduly favoured by the NIRF scores.

A quantitative estimate can be obtained using Peirce's measure of predictive success to determine if the use of one construct measure (say perception) to predict the event (say X-score) is acceptable or not[4]. This measure is independent of Pearson's correlation[12] and gives a more

quantitative refinement of the type-I and type-II classification of Smith[13].

In each case, we use the mean of the predictor scores (x) and event scores (y) for all 100 colleges as the thresholds (x_t, y_t respectively) for marking the four quadrants. This is done a priori, and is nothing more than a promise[14]. For each institution, the event is said to have taken place if $y \geq y_t$. Peirce's measure allows us to assess the
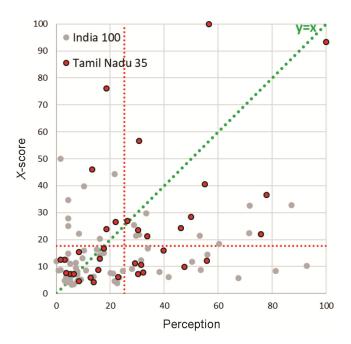

Figure 2. Construct validity map between X-scores and perception scores for the top 100 colleges in India and 35 from Tamil Nadu.
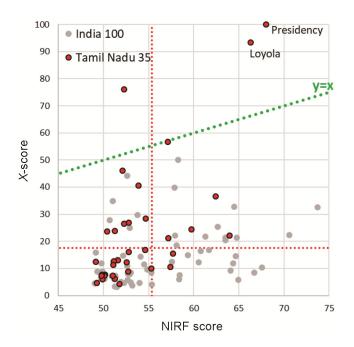

Figure 3. Convergent validity map between X-scores and NIRF scores for the top 100 colleges in India and 35 from Tamil Nadu.

success of our decision rule ($x \geq x_t$) in predicting the event ($y \geq y_t$). Figure 4 shows the four quadrants of the predictor–event space following Peirce[4]. The TT quadrant signifies all cases where the predictor has promised a top college (T for true), and the event shows that this has been realized (T for true). Similarly, FF quadrant collects all cases where the predictor has rejected the college (F for false), and the event shows that this has been correctly predicted (F for false, i.e. non-event). The FT quadrant is therefore the one that represents type-I error – incorrectly predicted events where we have rejected a case that should have been accepted[13]. The TF quadrant represents all type-II errors – incorrectly predicted non-events where we have accepted cases which should have been rejected. The Peirce's measure of 'the science of the method' is given by the simple formula

$$i = TT/(TT + FT) - TF/(TF + FF). \qquad (4)$$

We can show using simple calculations that i will range from 1 (the decision rule is 100% successful and there are no type-I and type-II errors), through 0 (TT/TF = FT/FF) to –1 (all are type-I or type-II errors).

Table 2 summarizes Peirce's measure of predictive success for comparison of the three CV measures of colleges from the three leading states in the NIRF top 100. We see from the last row that the perception bias for 18 colleges from Kerala takes zero value, indicating that peer review evaluation is unable to anticipate the NIRF score or the X-score. From the last column we see that the NIRF score is a better predictor of the X-score for Delhi, showing that the NIRF scores have been biased towards Delhi colleges and giving them undue prestige. This was also seen when we compared correlation coefficients (Table 1).

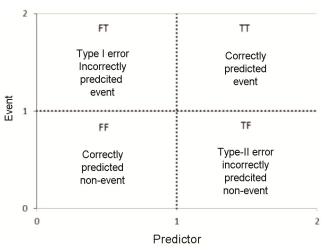The biases are also clearly noticeable visually if departure from the $y = x$ lines or occupation of the false

Figure 4. Four quadrants of the predictor–event space following Peirce[4].

Table 2. Peirce's measure of predictive success for comparison of the three CV measures of colleges from the three leading states in the NIRF top 100

Peirce's measure	NIRF – perception	X – perception	X – NIRF
Tamil Nadu 35	0.50	0.38	0.37
Delhi 29	0.65	0.36	0.52
Kerala 18	0.00	0.00	0.20

positive and false negative quadrants is taken into account. In all three states, and overall for the whole country, the perception score considerably underestimates most institutions and the NIRF score generally overestimates them.

We examined the 2019 scores from NIRF for the top 100 colleges in India from the CV point of view. The NIRF exercise provides a final score (the NIRF score) from five broad parameters for participating institutions. One parameter is an observed variable and is a peer-review-based perception score. Using the TLR parameter as a proxy for teaching and learning resources input, and the RPC and GO parameters as proxies for teaching and research outputs or outcomes, we independently computed a second-order X-score. The NIRF scores and the X-scores are latent variables that emerge from mathematical models. The three scores are compared in the context of CV, and weaknesses and biases can be recognized while validating such multi-dimensional evaluation exercises. Peirce's quantitative measure of predictive success is used to determine if one construct measure can predict the other.

One interesting but unsurprising takeaway is that the perception scores favour colleges based in Delhi and that this bias is carried over to the NIRF scores as well. Again, not unsurprisingly, the perception scores underestimate the performance of most of the institutions which are at the lower end of the spectrum. The bias is particularly felt by the colleges from Kerala.

Precisely the reverse picture emerges from the NIRF scores. Now, the performance of institutions at the lower end of the spectrum is shored up considerably – these scores are noticeably higher than those computed from the X-scores model. The X-score which is based on an input–output model, where the input term, TLR appears in the denominator, may give a better representation of reality than the NIRF model, where TLR is added to the output parameters to arrive at a final score.

1. All India survey on higher education (2016–17); http://aishe.nic.in/aishe/viewDocument.action?documentId=239 (accessed on 12 April 2019).
2. Prathap, G., Danger of a single score: NIRF rankings of colleges. *Curr. Sci.*, 2017, **113**(4), 550–553.
3. Savithri, S. and Prathap, G., Indian and Chinese higher education institutions compared using an end-to-end evaluation. *Curr. Sci.*, 2015, **108**(10), 1922–1926.
4. Peirce, C. S., The numerical measure of the success of predictions. *Science*, 1884, **4**(93), 453–454.
5. Bornmann, L., Tekles, A. and Leydesdorff, L., How well does I3 perform for impact measurement compared to other bibliometric indicators? The convergent validity of several (field-normalized) indicators. *Scientometrics*, 2019; https://doi.org/10.1007/s11192-019-03071-6.
6. Prathap, G., Making scientometric and econometric sense out of NIRF 2017 data. *Curr. Sci.*, 2017, **113**(7), 1420–1423.
7. Prathap, G., Totalized input–output assessment of research productivity of nations using multi-dimensional input and output. *Scientometrics*, 2018, **115**(1), 577–583.
8. Prathap, G., The Energy–Exergy–Entropy (or EEE) sequences in bibliometric assessment. *Scientometrics*, 2011, **87**(3), 515–524.
9. Cronbach, L. J. and Meehl, P. E., Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**(4), 281–302; doi:10.1037/h0040957.
10. Cook, T. D. and Campbell, D. T., *Quasi-Experimentation: Design & Analysis Issues in Field Settings*, Houghton Mifflin, Boston, USA, 1979.
11. Bornmann, L. and Daniel, H. D., Convergent validation of peer review decisions using the h index: extent of and reasons for type I and type II errors. *J. Informetr.*, 2007, **1**, 204–213.
12. Seglen, P. O., Why the impact factor of journals should not be used for evaluating research. *Br. Med. J.*, 1997, **314**, 498–502.
13. Smith, S. D., Is an article in a top journal a top article? *Financ. Manage.*, 2004, 133–149.
14. Prathap, G., Mini, S. and Nishy, P., Does high impact factor successfully predict future citations? An analysis using Peirce's measure. *Scientometrics*, 2016, **108**(3), 1043–1047.

Exploring the modulatory effect of albumin on calcium phosphate crystallization

Priyadarshini[1,*], Devesh Raizada[1], Pragya Kumar[1], Tanya Singh[1], Trisha Pruthi[1], Abhishek Negi[1], Lokesh Nigam[2] and Naidu Subbarao[2]

[1]Department of Biotechnology, Jaypee Institute of Information Technology, Noida 201 307, India
[2]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

Urolithiasis, also known as renal lithiasis, is a commonly occurring disease in humans. Calculi are formed in the urinary tract, including the tubular region of the kidney. Calcium phosphate crystals, especially hydroxyapatite, are responsible for nidus formation in the more commonly occurring calcium oxalate kidney stones. In the present study, effect of different concentrations of human serum albumin on

*For correspondence. (e-mail: priyadarshini@jiit.ac.in)