THE UNIVERSITY OF SYDNEY

University of Sydney feedback on the *ERA 2023 Benchmarking and Rating Scale - Consultation Paper*

22 April 2022

The University of Sydney acknowledges the thought and effort that the Australian Research Council (ARC) has invested in refreshing the Excellence in Research for Australia (ERA) exercise in recent years. Since its first trial and full implementation in 2009-10, ERA has played an important role in identifying and confirming the quality of Australian universities' research outputs by Field of Research (FoR) compared to world standards. However, ERA also requires an immense amount of ongoing effort on the part of Australian researchers, universities and the ARC, which comes at a considerable cost (both financial and opportunity) for all stakeholders. At the University of Sydney for example, participating in each ERA exercise consumes more than 40,000 hours of staff time and costs the University well in excess of $2 million in salaries alone. The full economic and opportunity cost of participation is much higher than this, however, as time spent by our researchers and staff meeting the requirements that arise from ERA participation is time that cannot be dedicated to our teaching, research, and its translation for societal benefit.

The reforms proposed in the *ERA 2023 Benchmarking and Rating Scale - Consultation Paper* are consistent with the recommendations of the ERA and EI Review and the Government's expectations that ERA will help 'drive the quest for excellence by Australia's universities' (p.3 & p.6). The proposed modifications to the rating scales aim to account for changes in the global research environment by realigning ERA to the high-quality top end, and provide greater granularity to enable differentiation between institutions and identification of areas of research strength. We understand the intent of the proposed changes, however, we do question whether the benefits of introducing them into the tight timeline to ERA 2023 outweigh the opportunity cost to Australian research.

Our responses to the consultation questions follow.

Section 2 - Options for a more granular rating scale

Q2.1 Which rating option (A or B) is preferred?

The proposed rating options share two main reforms: the merger of the two categories at the bottom of the rating scale into one band, and the introduction of the High Performance Indicator (HPI). Deleting less-used categories at the lower end of the rating scale and adding more gradation at the upper end of performance is aligned with improving performance information and encouraging excellence. We note that **Option A** was informed by the work of an eminent group of experts that comprised the ERA Benchmarking and Rating Scale (BRS) Working Group. The University of Sydney prefers the approach outlined in **Option A** for the following reasons:

- **Five rather than six rating bands** carry less risk of introducing errors through potentially unreliable over-magnification in the citation disciplines. There are concerns about whether the proposed new High Performance Indicator (HPI) will be able to reliably separate top band performance into three sub-ratings as featured in **Option B**.
- It creates **less change relative to previous ERA rounds**, and therefore makes interpretation easier than **Option B**, especially given the overlying complication of new FoR codes in ERA 2023.

We acknowledge that **Option B** provides greater granularity of performance at the top of the rating scale compared to **Option A**. However, there is considerable concern as to whether three sub-ratings in this top band are required and whether that level of granularity would be able to be robustly implemented. The sub-ratings in **Option B** would need to be very tightly defined and applied, and the results publicised (with analysis) to build trust in the reliability of outcomes using this system.

We note that the rating system labels are subject to change. The existing numerical labels worked well, and we do not have a preference for this aspect of either of the proposed options. To some, **Option A**'s labelling system seems easier to interpret as the labels themselves explain the rating relative to world standard/world leading (whereas the labels in Option B need a 'key' or set of definitions to be understood). Others consider **Option B**'s alphabetical rating system to be simpler and clearer, as well as more recognisable, relatable and likely to be comprehended by a wider audience than the world-based ratings proposed for **Option A**. In some disciplines, such as music, an abstract rating system is considered important because they are compared against institutions that are not connected with universities and their performance and practice-based research does not necessarily compare with the citation-based research of a 'music' department at an international university.

Q2.2 Are there particular features of either option that should be adopted or modified?

Option A:

- The five performance bands, which relinquish less-used categories at the lower end in favour of adding more gradation at the upper end of performance, should be **adopted**.

Option B (modify):

- For the proposed AAA, AA and A ratings, the meanings of the terms 'very small number', 'clearly above', 'comparable to other high performing institutions' respectively are unclear. If the intention is for ratings of A, AA or AAA to indicate that a university's performance in the assessed discipline is comparable with the performance of research institutions of all types in the top 10% globally, the definition for each sub-rating could be made clearer by introducing specific numerical performance thresholds, for instance:

  Performance assessed as comparable:
  AAA – top 2% of research institutions worldwide
  AA – top 3-5% of research institutions worldwide
  A – top 6-10% of research institutions worldwide.

  Alternatively, if **Option B** is to be adopted, the proposed three top-end ratings should be collapsed into two as follows:

  Performance assessed as comparable:
  AAA – top 5% of research institutions worldwide
  AA – top 6-10% of research institutions worldwide.

- For proposed **Rating B**, the definition 'universities that are below the higher performers but above the standard of other universities worldwide' is vague, especially as this rating appears likely to include a relatively large range of research institutions (~30%) given that a HPI rating covers the top 10%, while **Rating C** represents 'around the average standard of other universities worldwide' and **Rating D** includes those institutions with performance assessed as 'below the average standard of universities worldwide'. We also query why the term 'universities worldwide' is introduced for **Ratings B, C and D**, when the term 'institutions' is used for the HPI ratings, and the Consultation Paper notes that ERA assesses Australian universities' performance against all research institutions worldwide, not just other universities (p.9).

- For proposed **Rating C**, the term 'around the average standard of universities worldwide' is vague – at least for a lay reader/user of ERA outcomes. Should it be replaced with something like: 'at or within 5% of the average standard of research institutions worldwide'? In relation to 'world average', it is important to note that the Consultation Paper states: "'World average' is potentially easier to understand, and in citation disciplines can be unambiguously defined by an RCI of 1.0 and RCI class and centile distributions matching the world average (explained further at Appendix C)."

Q2.3 How will the change in ratings shift university research efforts?

As a basic design principle, the purpose of ERA is to serve as one means by which the Australian Government periodically measures - for the benefit of interested Australians and other stakeholders - the

quality of research undertaken within Australian universities. ERA should not be an end in itself, and should not be used to try to drive individual universities to pursue particular types of research or research strategies.

To have the trust, support and understanding of the Australian university and wider community, ERA must also be as simple, understandable, transparent, objective, evidence-based and cost-effective as possible. ERA assessments should encourage universities to 'shift' their research efforts, in a responsible way so as to maximise the excellence of their research efforts. It would be extremely unfortunate if institutions were to lack confidence in the ERA system and ignore its intent and outcomes and therefore fail to make changes to address weaknesses. ERA must be rigorously defensible and to enable this, the rating system must be transparent as well as clearly and tightly defined.

It is hard to predict how universities' research efforts may change as a direct result of changing the ERA ratings. However, from our experience with ERA to date, we anticipate that universities will continue to consume considerable time and resources - which would be better invested supporting research itself - adjusting their policies, systems, processes and approaches to compiling ERA submissions to maximise their prospects of performing well in ERA 2023 and each future iteration of the exercise. This considerable effort could be mitigated by providing the community with all data and decisions taken by the ARC to establish benchmarks, rather than institutions compiling these data themselves. This would also provide all universities with the same access this information rather than just those with capacity to resource this activity.

The change in ratings may encourage more focus on very high-end publications in internal performance standards, however it is unlikely to motivate a large shift of effort. This is because the large majority of university researchers already aim to reach a world-leading standard. It is important to note that a significant number of major paradigm shifts and breakthroughs are achieved through a mix of effort and collaboration *across*, rather than within, institutions, and such cross-institution collaborations are not really supported by ERA. It is also worth noting that many important research breakthroughs arrive by chance as well as effort.

Finally, it is important to note that most universities are not organised and structured cleanly in alignment with the named Fields of Research (FoR) that are used for the statistical purposes of classifying research outputs and related activities, and according to which ERA submissions are made and ratings determined. Rather, researchers based in diverse faculties, schools, research centres or affiliated research institutes - and who may have limited or no contact with each other - produce research outputs that are captured in the same FoR within a university's ERA submission. The capacity of large research institutions to significantly influence the focus, direction and the quality of research outputs by FoR for ERA purposes is limited by the reality that their constituent parts - employed and affiliated researchers - operate with large degrees of autonomy and in accordance with the principles of academic freedom. Strategic institution-level efforts to improve research quality can be pursued at the faculty, school, research centre or group level, but the results feed indirectly into ERA performance and can take considerable time to deliver changes in performance as measured by ERA.

Q2.4 To what extent would the proposed options be more challenging for universities than the existing ERA rating scale?

The existing ERA ratings are broad, but meaningful. The new ratings will challenge universities to identify world-leading and breakthrough research at the upper end of the scale. This may be more transparently achieved for citation than for peer review FoR.

Changing the system now will require a lot of re-education in a short timeframe. It is also difficult to undo decisions already made, and strategies already adopted to meet the existing rating system.

Q2.5 What changes, if any, are required to the characteristics that accompany each rating level?

As outlined in our responses to previous questions, clarity and transparency in the definitions and application of the rating scale are paramount to ensuring a robust and trusted system. For example, it will be challenging to make a real distinction between the 'high performer' and 'world benchmarks' as what distinguishes the two is currently not sufficiently clear. Lack of clarity in these and other definitions will make these measures hard for assessors to apply and outcomes hard for the research and wider community to trust.

It is currently unclear how a Unit of Evaluation (UoE) that substantively publishes in national venues can be treated as 'world leading'. A feature of the examples of 'world leading' (MIT, ETH Zurich) is that these groups produce outputs published in venues of international reach. This aspect of the characteristics - national venues amounting to world-leading status - is unclear and contradictory. In some disciplines, for example in the social sciences, what research will become paradigm-shifting is rarely understood at the time it is published, and the 6-year ERA reference period often acts against the identification of world-leading research. Research published in top venues from world leading institutions may never generate further research, attract citations or create impact in translation and application. At the same time, every discipline has examples of research, later identified as paradigm-shifting, that was rejected by the leading venues or lay dormant for many years. This category may have the unintended consequence of discouraging truly innovative research by favouring well-recognised methods and publication venues that peer-review panels can justify as world leading.

Q2.6 Would it be feasible for expert reviewers to draw meaningful distinctions between each rating points using the characteristics provided?

As stated in responses to previous questions, we believe this will be challenging unless the clarity of the wording is significantly improved so that distinctions between rating points are able to be clearly identified.

We also note that no institution or research group produces *only* world-leading research. One challenge will be to identify how much world-leading research a UoE needs to produce to be rated as 'world leading'.

Q2.7 What kind of additional training or guidance may be required in ERA 2023 to support the revised rating scale?

Clearer definitions of the categories and how to distinguish between them will be essential. In relation to the citation disciplines, further guidance in dynamic RCI usage and calculations is required; and in peer review disciplines, further guidance is required on how to select the 30% sample set. Time pressures will, however, make training and guidance hard to deliver effectively at this stage.

This will be especially challenging for the peer review codes given their subjective nature. For peer review (and especially creative) disciplines, a full understanding of what institutions are being measured against is needed. The guidelines provide examples that are specifically tailored towards empirical, citation-based research (eg MIT), however, they do not provide clarity on how creative or humanities disciplines, which may tend towards more parochial concerns in their research focus, are measured against one another. What is, for example, the "best of the high performing institutions" for music practice and research? Does The Juilliard School (NYC) or the Royal Academy of Music (London) compare with the University of Sydney Conservatorium of Music's research into the music of Australia's First Nations people or of the Asia-Pacific?

To ensure objectivity and better alignment in deliberations of the expert panels, the ARC should build into the evaluation process a robust discussion amongst members of those institutions that sit around the benchmark thresholds across each rating point and HPI range. This will ensure a shared understanding of what it takes to exceed a threshold across various institutional profiles for particular disciplines.

In the case of peer review, we would expect there to be a formalised study into bias across peer review rating outcomes that can be released to the wider academic community (after) the ERA cycle, so that the approach and decision points can be evaluated and refined. This will support transparency of process, recognising the subjective nature of peer review assessments, and demonstrate to the community a process of continuous improvement.

Section 3 – How can the citation metrics support the options for a revised rating scale?

Q3.1 How appropriate is the HPI as a method of supporting the rating scale options?

The HPI will favour larger institutions with a diverse range of high-performing research. It is broadly appropriate as a citation-based metric.

Q3.2 How appropriate are the dynamic RCI classes as a method of discipline-specific benchmarking?

The dynamic RCI classes are considered appropriate as a method of discipline-specific benchmarking.

Q3.3 How would the proposed citation methodologies impact research planning?

The proposed citation methodologies may drive researchers towards 'hot' topics.

As relates to ERA planning, they could potentially increase the already very significant workload of academic experts involved in ERA preparations, which is again more time that cannot be dedicated to research, research translation and teaching, and conceivably lead to increased tension in terms of strategy.

Q3.4 Do the new citation metrics support the drive for increased performance (especially in already high-performing disciplines)?

Broadly, yes. However, if the metrics have the effect of driving researcher efforts towards 'hot' topics, it will increase performance when that is defined by citations, because there is simply more activity in the area; but it will not necessarily increase real-world impact.

Q3.5 Is any additional criteria or information required in the citation disciplines to support the ratings at the highest end of each rating option?

No.

Q3.6 Please provide any additional comments on the proposed citation methodology.

The citation methodology needs to be more transparent and objective: This can be achieved by providing definitions regarding the following citation profiles used in the FoR ranking process:

- Worldwide average of the Dynamic RCI
- HPI average of the Dynamic RCI
- Distribution of articles against worldwide Dynamic RCI classes
- Distribution of articles against world centiles threshold.

Using formulas will add transparency to the citation methodology and eliminate the subjective decision making in regard to the FoR rankings.

Section 4 – How can the peer review indicator support the options for a revised rating scale?

Q4.1 To what extent are the proposed changes to peer review guidance likely to result in reports that are useful, informative and relevant for assessment panels? Please comment on any improvements that could be made, particularly with reference to disciplinary inclusivity and relevance to the options for the revised rating scale.

The proposed changes seem relevant, however, they do not provide a great deal of detail. Improvements should be considered in relation to:

- Discipline context:
  This section is likely to be helpful to assessment panels, although what the question about 'world-leading' discipline characteristics is intending to capture is unclear. Two (2) to three (3) exemplars of discipline based best practice should be identified, with explanatory notes as to why they are considered world-leading.
- Research practice:
  The last question about 'overall approach or range of approaches' is so general as to be unanswerable and should be deleted.

- Contribution:
  All the questions except the last are likely to offer useful information to assessment panels. The last question on 'overall contribution' is unhelpful and does not add to the earlier questions. It appears to be asking for an impression.

We note this is likely to be challenging for smaller multi-method disciplines, especially because assessors are likely to have many conflicts (and therefore the pool of assessors will be small). It may be necessary to look to international assessors to ensure independence and objectivity of the assessments. Other national research quality assessment frameworks make use of international assessors with success, such as Hong Kong's Research Assessment Exercise, which used 70% of non-local scholars for their 2020 assessment.

Q4.2 How feasible would it be for peer reviewers to address the proposed peer review guidance? Please comment on any improvements that could be made, particularly with reference to clarity and workload for reviewers.

It will be very difficult for peer reviewers to accurately support the revised rating scale without very detailed guidance. The questions assume that there is more consistency in research practices at 4-digit FoR code level than there actually is. In many 4-digit codes, though not all, the outputs institutions contribute for ERA encompass a broad range of qualitative and quantitative methods. This makes simple statements about consistency with the standard disciplinary practices almost impossible. Unless the review is brought down to the 6-digit FoR level, the questions about field consistency are rather meaningless. Reference could be made in the guidelines to the problem of discipline variation across institutions. At least one example should be provided of how a reviewer is to conceive of a creative faculty in a 'world' university given the variation internationally.

The questions themselves are detailed and represent a large workload. It would be helpful to prioritise questions for peer reviewers and give guidance about the amount of detail needed. The process should provide peer reviewers with training to minimise bias and include (as per comments in response to Q2.7) a more formal study to give recommendations on the ongoing management of this issue. The guidance material should be amended in accordance with the recommendations from such studies.

Some consideration should be given to the recruitment of a large enough pool of peer reviewers to ensure that sufficient discipline specific expertise is available, and that the quality of the reviews is adequate to support the rating of the units of evaluation.

Q4.3 How appropriate is the proposed guidance for Indigenous studies? Please comment on any improvements that could be made.

We note the guidelines and High Performance benchmark make no reference to Indigenous studies and we suggest the ARC updates the benchmarks/guidance to include reference to the importance and significance of First Nations research and knowledge, in consultation with its ERA EI Indigenous Studies Leadership Group. Due to the recent inclusion of the Indigenous studies codes, examples of world-leading research and world-leading research groups, with explanatory notes, would greatly assist with the evaluation of these new UoEs, as well as provide guidance to universities on selecting the 30% sample in these codes. Exemplars of best practice with explanatory notes would also be useful particularly to guide the selections in UoEs that have traditionally been evaluated according to citation metrics, such as the Aboriginal and Torres Strait Islander health and wellbeing code.

Given that it is a nascent FoR, this will be challenging to assess and may not yield insightful output for a few cycles.

Q4.4 How would the proposed changes to peer review guidance impact universities and/or researchers?

While the guidance is very general in nature, it may assist universities in selecting their 30% sample for submission. Additional information about disciplinary context for the reviewers and assessment panel may also help researchers feel more confident about ERA assessments.

THE UNIVERSITY OF SYDNEY

Q4.5 Is any additional criteria or information required in the peer review disciplines to support the ratings at the highest end of each rating option?

More thought could also be given to rating and distinguishing works at the highest end to identify world-leading work by world-leading research groups. Consideration should be given to the indicators that identify world-leading research. These could be formulated around relevant evidence of peer recognition and research excellence as relevant within the specific discipline. Indicators could include impact and reach of outputs and outlets where appropriate.

Examples of world-leading research and world-leading research groups, with explanatory notes, would help with this new task. Exemplars of best practice with explanatory notes may be required within FoR codes.

Q4.6 Are there any other changes to peer review that the ARC should consider?

The ARC should verify the consistency of ratings by randomly assigning a sample of submissions to several panels and comparing the ratings. There are several structured approaches for doing this and doing so in more formal manner will assist perceptions of the process's transparency and accountability. It will also demonstrate a commitment to improving the models and approach to peer review assessment, especially if made available to the community.

There is concern that **Option B** as currently structured will lead to false precision in which reviewers are asked to break apart even more fine grade distinctions in the top performance band and will rely on weak indicators to force apart assessments. This would lead to minor differences having an exaggerated effect at the higher end.

Any additional comments?

The University of Sydney understands that the Government is keen to continue ERA assessments and appreciates the serious consideration and expertise the ARC has applied to developing the two major reforms proposed in the Consultation Paper and to rethinking options for the rating scale to refresh and attempt to glean more value from the ERA exercise.

We note that the Consultation Paper does not engage with threshold questions relating to whether the benefits of ERA are worth the effort, expense and opportunity cost to Australian research that participation in the exercise requires. This is understandable given statements in the Acting Minister's December 2021 *Letter of Expectations* in relation to fast-tracking the ARC's response to the ERA and EI Review, and the timeline for ERA 2023 to which the Government and the ARC have committed. However, these are important questions for Australia to consider given the investment of resources ERA requires. As Universities Australia has recommended, there would be value in the Government giving the ARC adequate time to take proper stock of the program so as to ensure that it is fit-for-purpose and remains a cost-effective exercise.