Australian Government
National Health and Medical Research Council

NHMRC

NHMRC Report

Peer Review Analysis Committee

Australian Government
National Health and Medical Research Council

NHMRC

NHMRC Report of the work of the Peer Review Analysis Committee

Introduction

Background

The National Health and Medical Research Council (NHMRC) relies on the work of thousands of peer reviewers to guide the selection of grant applications for funding. Peer review is demanding: it draws both on the specific expertise of assessors to evaluate complex scientific proposals and on their judgement of applications against criteria such as significance and innovation. NHMRC therefore seeks to maximise the rigour of peer review while minimising the burden of this work on individual assessors and the health and medical research sector as a whole.

This is a difficult balance to strike and depends on the scale (number of applications) and scope (range of research disciplines and fields) of the grant scheme. NHMRC has paid particular attention to streamlining peer review of applications to its two largest schemes, Investigator Grants and Ideas Grants, which together receive more than 3,500 applications covering a wide range of fields each year and are allocated about 65% of the available budget for grants. Both schemes were introduced in 2019 for funding from 2020 as part of a substantial reform of NHMRC's grant program.[1]

Since 2020, NHMRC has progressively moved the assessment of Investigator Grant and Ideas Grant applications from traditional Grant Review Panels (where sets of applications are allocated to a panel of assessors who meet in person) to 'application-centric' peer review (where each application is reviewed independently by a unique set of best-fit assessors).[2] This change was initiated in response to feedback from many assessors that their Grant Review Panels did not have the breadth of expertise needed to assess their allocated applications and evidence from within and outside NHMRC that panel discussion had only a modest impact on final scores compared with spokespersons' pre-panel scores.[3,4] Progression to application-centric peer review was accelerated when the COVID-19 pandemic limited in-person meetings in 2020 and 2021.

With reduced use of Grant Review Panel meetings at a time of low funded rates, applicants have expressed concerns about the robustness of peer review processes, particularly the potential impact of divergent ('outlier') scores on funding outcomes. To support consideration of these issues, NHMRC formed the Peer Review Analysis Committee (PRAC) in 2020 to analyse and provide advice on the peer review processes used in the first two rounds of the Investigator Grant and Ideas Grant schemes.

---

[1] New grant program | NHMRC

[2] Application-centric peer review | NHMRC

[3] Peer review for Ideas Grants in 2021 | NHMRC

[4] Carpenter AS, Sullivan JH, Deshmukh A, Glisson SR and Gallo SA. A retrospective analysis of the effect of discussion in teleconference and face-to-face scientific peer-review panels. *BMJ Open* 2015, 5(9). http://dx.doi.org/10.1136/bmjopen-2015-009138

Scope and conclusions of the Peer Review Analysis Committee

With the support of the Office of NHMRC, PRAC analysed data from the 2019 and 2020 rounds of the Investigator Grant and Ideas Grant schemes to investigate the impact of scoring behaviour on application outcomes. Specifically, PRAC considered:

- distributions of assessor scores
- the frequency of outlier scores, i.e. those that differed markedly from other assessors' scores of the same application
- the proportion of assessors who tended to score higher (more generously) or lower (more critically) than other assessors of the same applications
- the predicted effect of systematically excluding each assessor or each assessment on the funding outcomes
- the percentage of variation in scores due to application quality (signal-to-noise ratio) versus systematic assessor effects (generosity or negativity) and random variation between assessors
- methods to adjust scores to remove systematic assessor effects, by either normalising (to the same mean and standard deviation) or rescaling (to the same mean only) the scores of all assessors reviewing a category of applications.

Based on these analyses and members' knowledge of international practices and relevant literature, PRAC drew a number of conclusions about the peer review processes used in 2019 and 2020 that have broad relevance for NHMRC's current processes for the review of Investigator Grant and Ideas Grant applications, as summarised below.

There is no gold standard for peer review processes.

While independent peer review against published criteria is the international gold standard for making grant funding decisions, there is little agreement across the research sector, nationally or internationally, on the gold standard for the design of the peer review process and how peer review recommendations should be used to make funding decisions.

Peer review is inherently variable.

Variation between assessors' scores for an application is an inherent part of the peer review process as it relies on individual judgement and is also affected by random variation. In part it reflects the challenge of distinguishing between highly competitive and complex applications.

The system is sensitive to variation in scores near the funding cut-off.

Scores are clustered and small differences in score around the funding cut-off can determine whether or not an application is funded.

Outlier scores can affect funding outcomes and are not necessarily incorrect.

Outlier scores (i.e. scores that are 2 or more points away from an application's mean or median assessor score on a 7-point scale) can affect funding outcomes. More minor variation in scores can also be influential near the funding cut-off.

Outlier scores may reflect the specific expertise or judgement of the assessor and are not necessarily incorrect. Assessor training may help to reduce variation but would not be expected to eliminate outliers. As outlier scores may reflect differences in scientific

judgement, they should not be excluded unless there is clear evidence to justify their exclusion. All assessors should be reassured of the importance of their work (i.e. every review counts).

In PRAC's analyses, most influential scores (i.e. those that changed funding outcomes if they were excluded) were not outliers.

Some assessors are more generous or more critical than others.

Some assessors tended to score higher or lower than other assessors of the same applications. However, the assessors who had the greatest influence on funding outcomes were generally not from these groups.

Some systematic assessor effects could be removed by normalising or rescaling and this would change the funding outcome for some applications. Given the inherent variation in peer review, without experimentation it is not clear that these adjustments would give a more repeatable or accurate outcome.

Changes to NHMRC's peer review processes

NHMRC has valued the careful consideration, insight and advice provided by PRAC. In response to PRAC's discussions and feedback from the sector, NHMRC has implemented several changes to its peer review processes to increase quality, transparency and accountability. Notable amongst these are: screening to identify and verify outlier scores; requiring assessors to provide comments to explain their scores; sharing of assessor comments with other assessors of the same application; better matching of applications to assessors; and strengthened support for assessors.

NHMRC will continue to draw on PRAC's work and other data, feedback from applicants and assessors, and the experience of other national and international funders in the further development of its peer review processes and training activities. Issues for ongoing consideration by NHMRC include efforts to measure and understand assessor variability, ways to improve calibration between assessors and clearer differentiation between category scores.

1. Purpose

PRAC was established on 1 October 2020 under section 39 of the *National Health and Medical Research Council Act 1992* to advise the NHMRC CEO on aspects of the peer review process. Members of the committee (Appendix A) were appointed based on the expertise required to analyse the statistical and procedural aspects of peer review and scoring of grant applications.

NHMRC has implemented significant changes to its peer review processes for two major grant schemes, Investigator Grants and Ideas Grants, in recent years (Appendix B). One of those changes has been a shift to obtaining multiple independent assessments of each grant application, together with reduced use of Grant Review Panels to discuss applications and finalise scores. This change has highlighted the importance of having robust processes for scoring and ranking applications in which the research community can have confidence. Low funded rates have also contributed to greater scrutiny of peer review.

PRAC was formed to advise the CEO on several aspects of the peer review process, including recommending possible changes, while ensuring that NHMRC's peer review principles are upheld and the objectives of the grant program are met. The focus was primarily on the Investigator Grant and Ideas Grant schemes.

PRAC's terms of reference were to advise on:

- the significance of divergence between scores of independent reviewers and the effect of divergent scores on outcomes
- appropriate methods to identify and manage divergence
- methods for determining final scores for applications within a panel and for final ranked lists from multiple panels
- mechanisms to allow peer reviewers to calibrate their scores against those of other peer reviewers of the same grants
- any other related issues as requested by the CEO.

2. Defining the problem

The challenge of recognising and rewarding the 'best' science is felt throughout the research sector, affecting decisions on grant funding, publications and appointments and promotions. For research funding agencies, the peer review processes that underpin funding decisions need to demonstrate fairness, transparency, robustness and consistency. Pressure on funding agencies to make the 'right' decisions increases as grant funded rates decrease.

Peer review by independent experts against published criteria is considered the international gold standard for allocating research grants. However, there is little agreement across the research sector on the ideal design of scientific peer review or how peer review recommendations should be used to make funding decisions. Design of peer review processes can reflect a range of factors, including the scope, scale and strategic objective of the grant scheme, assessment criteria, availability of expert reviewers, resources (funds and staff), time available to make decisions, and other local or historical conditions.

This diversity of views is reflected in the proliferation of peer review systems used by Australian and international research funding agencies, all of which are likewise contested to some degree by their own research communities. Peer review and funding decisions also have an inherent level of subjectivity and variability, as they rely on the opinions and judgements of people who have different experiences and expertise.

Against this background, PRAC was asked to advise on the peer review processes used in the first two rounds of NHMRC's Investigator Grant and Ideas Grant schemes.

3. Investigations

PRAC applied an evidence-based approach based on detailed analyses of NHMRC data and members' knowledge of published research on peer review. Each item in the terms of reference, along with anecdotal feedback and perceptions reported from the sector, was addressed by analysing data gathered over the first two rounds (2019 and 2020) of the Investigator Grant and Ideas Grant schemes (Appendix B).

The analyses considered by PRAC covered three broad themes:

1. General Observations – These analyses sought to characterise the outcomes of the two rounds in detail and provide context for further analyses.
2. Assessor effects and outlier assessments – These analyses profiled the behaviour of assessors to understand their impact, as individuals within a larger pool of assessors in each round, on the outcome of applications.
3. Methods to modify assessor scores – These analyses investigated how the outcome of applications would change when statistical methods were applied to modify assessor scores.

3.1 General observations

NHMRC presented an overview of application score distributions in the 2019 and 2020 rounds of Investigator Grants and Ideas Grants (see Section 3.1.1) to provide context for subsequent analyses. NHMRC also presented an analysis of scoring against each Investigator Grant assessment criterion to determine whether there were any differences in consistency in how the criteria were applied (see Section 3.1.2).

In 2019 and 2020, five independent assessments were sought for each Investigator Grant application (with a minimum of four achieved) and four independent assessments were sought for each Ideas Grant application (with a minimum of three achieved). Final scores were derived from the mean of all assessments for each application. Grant Review Panel meetings were generally not held (with the exception of the 2019 Ideas Grant round) (see Appendix B). Within each competition, funding recommendations were based on the ranked list of final scores with funding allocated to the highest ranked applications until the total budget for the competition was reached.

3.1.1 Overview of score distributions

PRAC considered distributions of application final scores and summary statistics, including the mean (with standard deviation), median and cut-off scores.

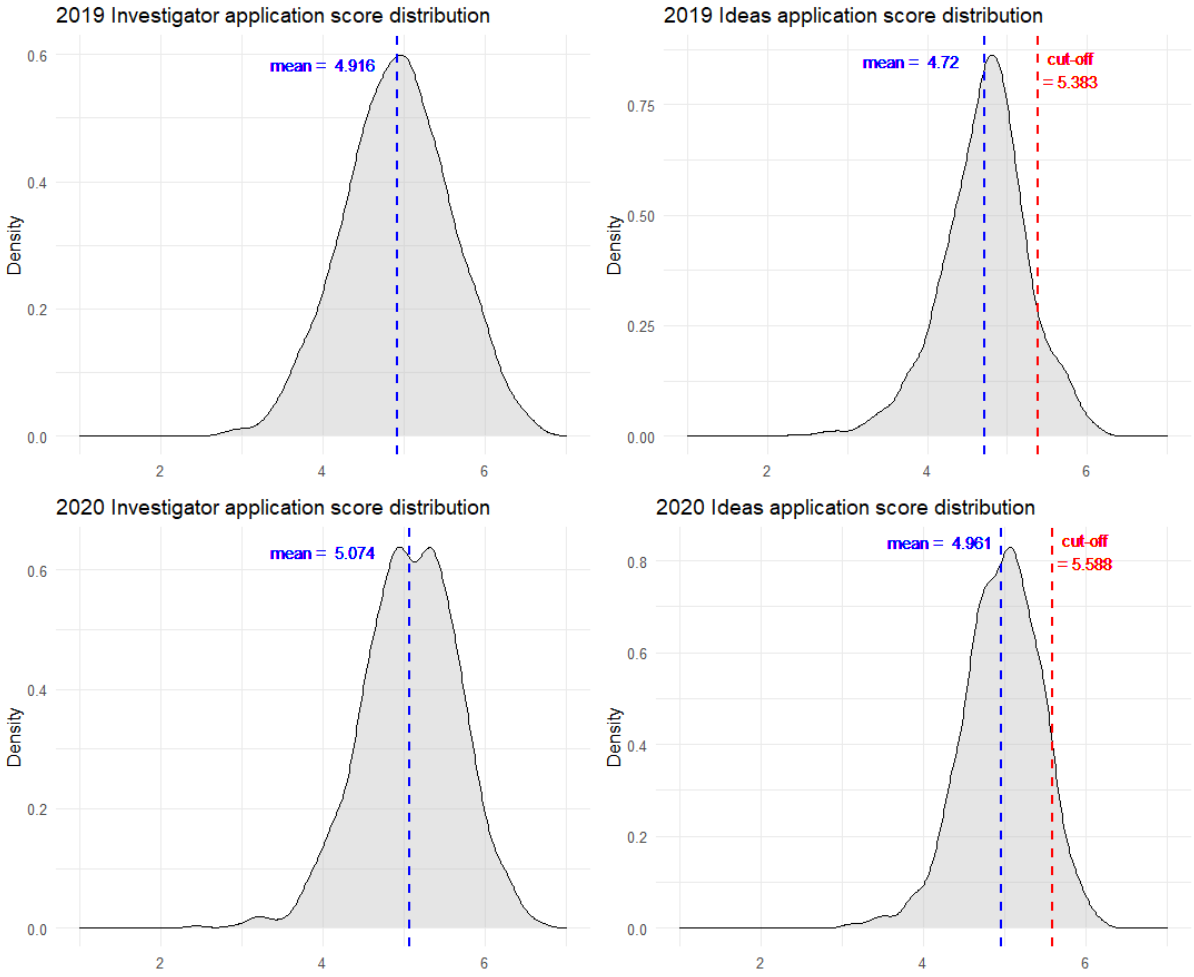
This analysis revealed that:

- The distribution of final scores in Investigator Grants followed an approximately normal distribution in 2019 and to a lesser extent in 2020.
- The distribution of final scores in Ideas Grants followed an approximately normal distribution in both 2019 and 2020.
- For both schemes in both years, the majority of scores were clustered between 4 and 6 and all funding cut-off scores were well above the mean.

- In both 2019 and 2020, the distribution and mean scores of Emerging Leadership level 2[5] (EL2) and Leadership (L) Investigator Grant applications were similar. EL1 level applications had a slightly lower mean score than EL2 or L applications.

The results are summarised below in *Figure 1*, *Table 1, Table 2* and *Figure 2*.

Figure 1: Investigator Grant and Ideas Grant final score distributions in 2019 and 2020[6]



[5] The Investigator Grant scheme comprises two categories: Emerging Leadership (EL) and Leadership (L). The EL category is restricted to researchers who are ≤10 years post-PhD or equivalent and has two levels (EL1 and EL2) which differ in salary and Research Support Package (RSP). The L category has three salary levels (L1, L2 and L3) and four tiers of RSP (LT1, LT2, LT3 and LT4). The scheme is run as three discrete competitions (EL1, EL2 and L), each with a predetermined budget.

[6] There were no application scores less than 1.

Table 1: Investigator Grant and Ideas Grant mean and median final scores and funding cut-offs in 2019 and 2020

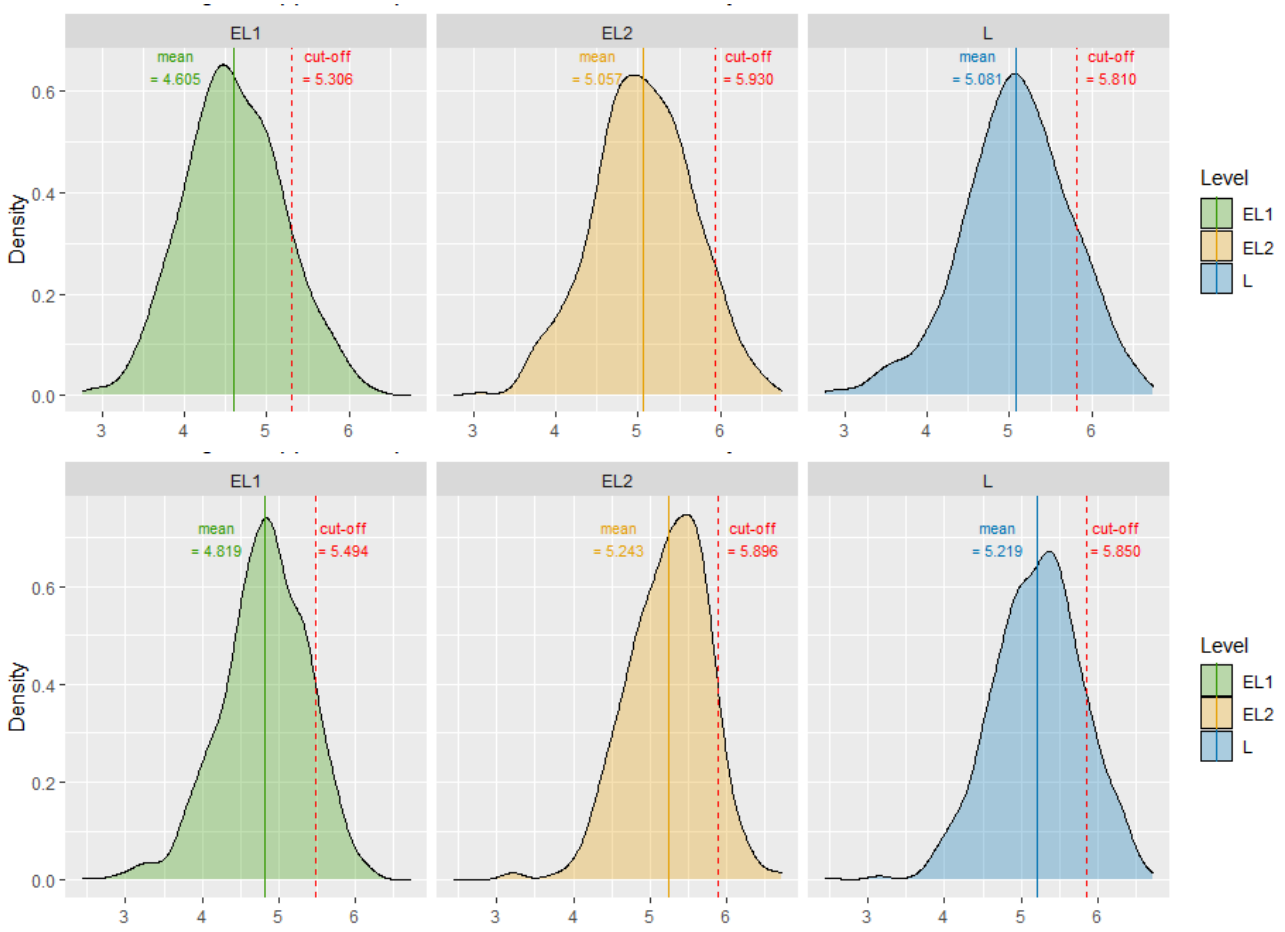
Scheme	Year	Mean score ± standard deviation	Median score	Funding cut-off
Investigator Grants	2019	4.916 ± 0.661	4.920	Refer Figure 2
	2020	5.074 ± 0.597	5.082	Refer Figure 2
Ideas Grants	2019	4.720 ± 0.545	4.754	5.383
	2020	4.961 ± 0.484	4.988	5.588

Table 2: Investigator Grant mean and median final scores and funding cut-offs by level in 2019 and 2020

Year	Level	Mean score ± standard deviation	Median score	Funding cut-off
2019	EL1	4.605 ± 0.596	4.573	5.306
	EL2	5.057 ± 0.600	5.060	5.930
	L	5.081 ± 0.659	5.090	5.810
2020	EL1	4.819 ± 0.566	4.850	5.494
	EL2	5.243 ± 0.519	5.288	5.896
	L	5.219 ± 0.584	5.244	5.850

It is important to note that the funding cut-off is not set in advance but rather it indicates the lowest score of a successful application in the round. For the purposes of this analysis, applications that were funded through the structural priority initiative or the Electromagnetic Energy (EME) program were excluded when determining the funding cut-off.

Figure 2：Investigator Grant final score distribution by level in 2019 (top) and 2020 (bottom)



3.1.2 Analysis of scores for individual assessment criteria

Investigator Grant applications are assessed against two major criteria: Applicant Track Record (70%) and Research Program (30%). The Track Record score is made up of 5 elements, each of which is scored separately: Publications (35%), Research Impact (comprising Research and Significance of the Research Impact, Research Program's Contribution to the Research Impact and Applicant's Contribution to the Research Program) (20%) and Leadership (15%). The Research Impact elements were included in NHMRC's track record assessment framework for the first time in the 2019 round.

PRAC asked whether the variance of scores differed between assessment criteria. The analysis below compares the variance of scores for the Publications criterion against each of the three Research Impact criteria to investigate whether the Research Impact criteria were less consistently applied than the Publications criterion.

The approach used in this analysis was:

- for each application, calculate the sample variances[7] of assessor scores for the three Research Impact criteria and the Publications criterion respectively, and
- compare[8] the variance of each of the three Research Impact criteria scores against the variance of Publications scores.

The analysis revealed that:

- For the great majority (>90%) of applications, there was no statistically significant difference in score variance for any one of the Research Impact criteria compared to the Publications criterion. However, these results should be treated with caution as the limited number of assessors per application (four or five) inhibit the statistical variance test from identifying small differences.
- Where there was a difference in the variance between the criteria, a higher proportion of applications had significantly higher variance in the Research Impact criteria scores than the Publication criterion scores.

The results are summarised below in *Table 3*.

When drawing conclusions from this analysis, it should be noted that the variance test is insensitive to a small difference between variances due to the limited number of assessors per application. The average number of assessors per application is approximately five, which means each sample variance is calculated using five assessment scores at most; this is a small sample size.

Table 3: Comparison of variance in scores for Research Impact and Publications criteria

Research Impact criteria	Comparison to Publications criterion	Number of applications	Percentage of applications
Reach and Significance of the Research Impact	No significant difference	1,683	90.7%
	Reach and Significance scores have significantly higher variance	126	6.8%
	Publications scores have significantly higher variance	46	2.5%
	Total	1,855	100.0%
Research Program's Contribution to the	No significant difference	1,693	91.3%
	Program's Contribution scores have significantly higher variance	115	6.2%
	Publications scores have significantly higher variance	47	2.5%

[7] This statistical test establishes how similarly the assessors scored each application against these criteria and thus infers the degree to which assessors applied the assessment criteria consistently.

[8] F-tests were conducted with a significance level of 5%.

Research Impact	Total	1,855	100.0%
Applicant's Contribution to the Research Program	No significant difference	1,696	91.4%
	Applicant's Contribution scores have significantly higher variance	112	6.0%
	Publications scores have significantly higher variance	47	2.5%
	Total	1,855	100.0%

3.2 Assessor effects and outlier scores

Noting that assessor scores are inherently variable because they rely on the judgements of individuals with different perspectives and expertise, PRAC discussed what level of repeatability (i.e. the percentage of applications that would have the same funding outcome when an application is reviewed independently by two panels of assessors) would be acceptable to the sector. It was noted that one study[9] showed that surveyed NHMRC applicants were willing to accept a wide range in repeatability, with the most common response being 75%.

PRAC sought to understand how the scoring behaviour of individual assessors influenced the outcome of applications. This issue was investigated using a number of approaches as outlined below.

PRAC examined the frequency of outlier assessment scores (i.e. instances where an assessor scored notably differently from their peers for the same application) which some researchers had suggested could be inappropriately affecting application outcomes (see Section 3.2.1).

PRAC also investigated the tendency of individual peer reviewers to score consistently higher or lower than the other assessors to identify possible assessor effects (i.e. assessors tending to be more generous or critical in their scoring; see Section 3.2.2).

To examine whether scoring divergence might have a strong influence on application funding outcomes, two methods were used: characterising the effects on outcomes of excluding all scores from an individual assessor and excluding any individual score (see Sections 3.2.3 and 3.2.5 respectively).

PRAC also investigated whether those assessors who exhibited assessor effects in the analysis presented in Section 3.2.2 were also those most influential on outcomes, as identified in the analysis in Section 3.2.3. This cross-sectional analysis is presented in Section 3.2.4.

[9] Herbert DL, Barnett AG, Clarke P and Graves N. On the time spent preparing grant proposals: an observational study of Australian researchers. *BMJ Open* 2013;**3:**e002800. doi: 10.1136/bmjopen-2013-002800

Finally, PRAC members used a signal-to-noise[10, 11] estimation procedure to quantify the robustness of the peer review system (see Section 3.2.6).

3.2.1 Outlier scores

Researchers had raised concerns about the possibility that outlier scores could advantage or disadvantage applications. Concerns have been centred on the risks of lower assessment scores.

Based on technical advice from PRAC members, the method for this analysis was:

- define an outlier assessor score based on how far an assessor's score is from an application's final score
- for each assessment, calculate the number of points away from application final score (i.e. the difference between assessor score and application final score in absolute value) in a selected round
- identify divergent scores whose difference from the application final score is larger than a certain 'points away' threshold.

Table 4 below shows the results for 2019 Investigator Grants under four different 'points away' thresholds. This analysis revealed that:

- Most individual assessor scores were less than one point away from the final score.
- If a threshold of "two or more points away" was used , there would be 16 (out of 3,103) assessments at the EL1 level, 6 (out of 2,325) assessments at the EL2 level and 5 (out of 3,698) assessments at the L level that would be considered divergent/outliers.

Table 4: Number and percentage of assessments by threshold (absolute difference from final score) and level in the 2019 Investigator Grant round

Investigator level	Number (percentage)				
	0–0.999 point away	1–1.499 points away	1.5–1.999 points away	2+ points away	Total
EL1	2,594 (86.1%)	330 (11.0%)	73 (2.4%)	16 (0.5%)	3,013 (100%)
EL2	2,061 (88.6%)	225 (9.7%)	33 (1.4%)	6 (0.3%)	2,325 (100%)
L	3,356 (90.8%)	291 (7.9%)	46 (1.2%)	5 (0.1%)	3,698 (100%)

[10] Marsh HW, Jayasinghe UW and Bond NW. Improving the peer-review process for grant applications. Reliability, validity, bias, and generalizability. generalizability. *Am Psychol* 2008, 63(3):160-8. doi: 10.1037/0003-066X.63.3.160.

[11] Visscher PM and Yengo L. The effect of the scale of grant scoring on ranking accuracy [version 2; peer review: 2 approved with reservations] *F1000Research* 2023, 11:1197 https://doi.org/10.12688/f1000research.125400.2

Total	8,011 (88.7%)	846 (9.4%)	152 (1.7%)	27 (0.3%)	9,036 (100%)

Possible limitations of this analysis should be taken into consideration. The definition of an outlier score in this analysis is based on points away from the mean of all assessors' scores (i.e. application final score). PRAC noted that this working definition is not ideal because the mean score is itself affected by the outlier score. It might be more natural to define, though more complex to calculate, an outlier based on a given number of points away from the mean of the remaining assessors' scores.

3.2.2 Assessor Generosity

While there might be valid reasons for assessors scoring differently, such as expertise in a niche field of research or knowledge of key issues with the proposed research, divergent scores might also be due to systematic assessor behaviour where assessors tended to score consistently differently from their peers. Such a finding might support the use of mechanisms to calibrate assessors' scoring.

PRAC discussed ways to identify divergent assessors. Members suggested looking at whether there were assessors who tended to score lower or higher than other assessors of the same applications.

NHMRC analysed the scoring behaviour of assessors and categorised them based on their tendency to score more generously or critically than their peers.

The method to conduct this analysis was:

- define a new variable called 'DeviationRatio':
  - a 'DeviationRatio' measures the extent to which an assessor's score for an application deviates from the final score of that application. It is calculated by subtracting the application final score from the score of a single assessment and then dividing it by the application final score (see Appendix C for formula).
- calculate 'DeviationRatio' for each assessment (i.e. each assessor's score deviation from the final score for each application they review).
- classify assessors into five categories by the proportion of their assessments with positive and negative values for DeviationRatio. The categories are set using a 95% threshold[12] as the limit for the most distal categories and are defined in *Table 5* below.
- visualise all DeviationRatios to categorise assessor behaviour as in *Figure 3*.

[12] The threshold is used to manage the confidence level of an assessor's generosity or negativity. A higher threshold indicates more consistent scoring behaviour.

Table 5: Classification of assessors by their proportion of Deviation Ratio

Classification	Criteria
Mostly Generous	Positive DeviationRatio for between ≥95% and 100% of assessments
Likely Generous	Positive DeviationRatio for between >60% and <95% of assessments
Balanced	Equal DeviationRatio for assessments (≤60% positive and ≤60% negative)
Likely Critical	Negative DeviationRatio for between >60% and <95% of assessments
Mostly Critical	Negative DeviationRatio for between ≥95% and 100% of assessments

Figure 3：Distribution of assessors' DeviationRatio with a threshold of 95% in the 2019 Investigator Grant round

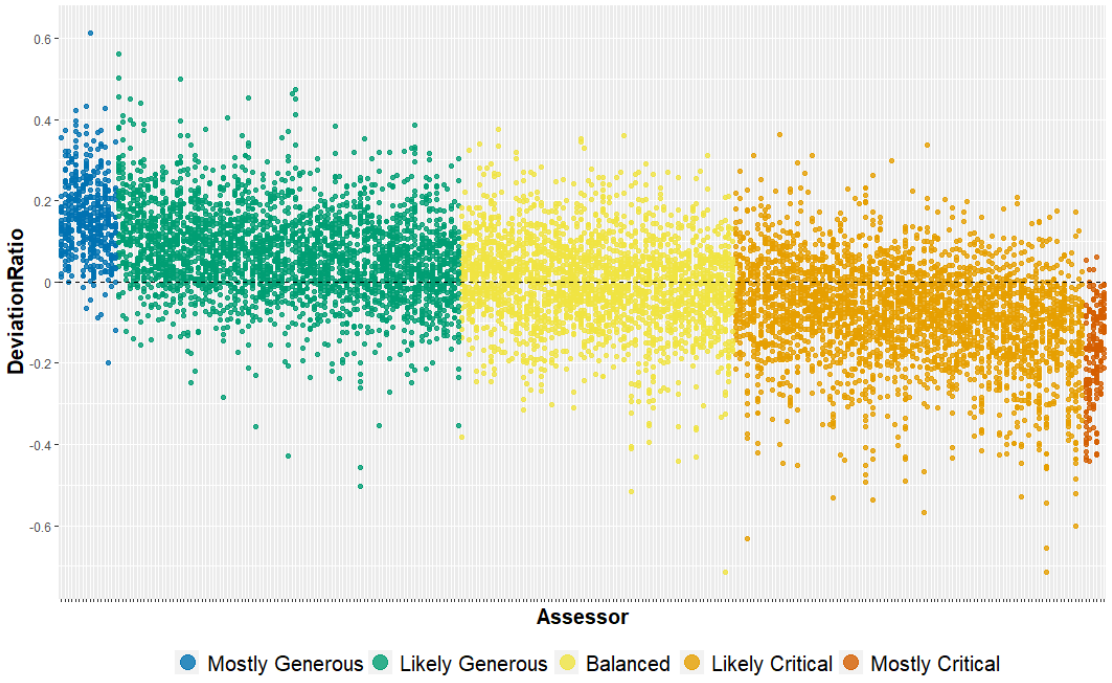


Table 6 compares results using four different thresholds (80%, 85%, 90% and 95%). It shows the number and proportion of assessors based on their scoring profile and reveals that:

- The largest populations of assessors are those who are Likely Generous and Likely Critical followed by those who are Balanced. Few assessors score generously for at least 95% of the applications they review while even fewer score critically for at least 95% of the applications they review.
- In the 95% threshold scenario, 5.5% of individuals are classified as 'Mostly Generous' and 2.1% are classified as 'Mostly Critical', which means in total 7.6% of assessors gave a score above or below the average score of their peers for each application they reviewed in more than 95% of instances in the 2019 Investigator Grant round.

Table 6: Number and percentage of assessors by scoring behaviour classified under different thresholds.

Threshold (X)	Number (percentage)				
	Positive DeviationRatio for ≥X% of assessments	Positive DeviationRatio for between >60% and <X% of assessments	Positive or Negative DeviationRatio for between 40% and 60% of assessments	Negative DeviationRatio for between >60% and <X% of assessments	Negative DeviationRatio for ≥X% of assessments
	Mostly Generous	Likely Generous	Balanced	Likely Critical	Mostly Critical
95%	16 (5.5%)	95 (32.8%)	76 (26.2%)	97 (33.4%)	6 (2.1%)
90%	24 (8.3%)	87 (30.0%)	76 (26.2%)	87 (30.0%)	16 (5.5%)
85%	35 (12.1%)	76 (26.2%)	76 (26.2%)	74 (25.5%)	29 (10.0%)
80%	50 (17.2%)	61 (21.0%)	76 (26.2%)	60 (20.7%)	43 (14.8%)

The results presented only provide observations for a single year and may not reflect long-term trends. This analysis compared the scoring behaviour of the four to five assessors for each application. Inter-panel generosity (between the groups of assessors) was not examined.

3.2.3 Leave One Assessor Out

PRAC discussed concerns raised by the sector that individual assessors may have an undue or disproportionate impact on the outcome of some applications – that is, that one assessor could cause many applications to be funded or not due to their unique scoring behaviour.

Based on technical advice from members, NHMRC analysed whether excluding the scores of any single assessor could affect the final score of any application enough to change its funding outcome.

The method to conduct this analysis was:

- systematically select each assessor and exclude all scores they provided for applications that they assessed in the 2019 Investigator Grant round
- calculate new final scores (i.e. Leave One Assessor Out scores) for the affected applications using the remaining scores from the other assessors
- rank all applications using their Leave One Assessor Out score (where calculated) or final score
- classify applications as 'funded' (in descending rank order) until the same number of applications (by level) is identified as were actually funded in the round (note that this is a simplified funding process)

- compare the hypothetical Leave One Assessor Out funding outcome to the actual outcome for those applications that were reviewed by the excluded assessor.

This analysis revealed that:

- 37.6% of assessors had no direct impact that would change potential funding outcomes. This means that excluding the scores of any one of these individuals did not change the final score for any application enough to influence whether the application was funded in this hypothetical scenario.
- 54.8% of assessors directly affected the outcome of up to 10% of the applications they assessed.
- 5.9% of assessors directly affected the outcome of 10–15% of the applications they assessed.
- 1.7% of assessors directly affected the outcome of more than 15% of the applications they assessed.

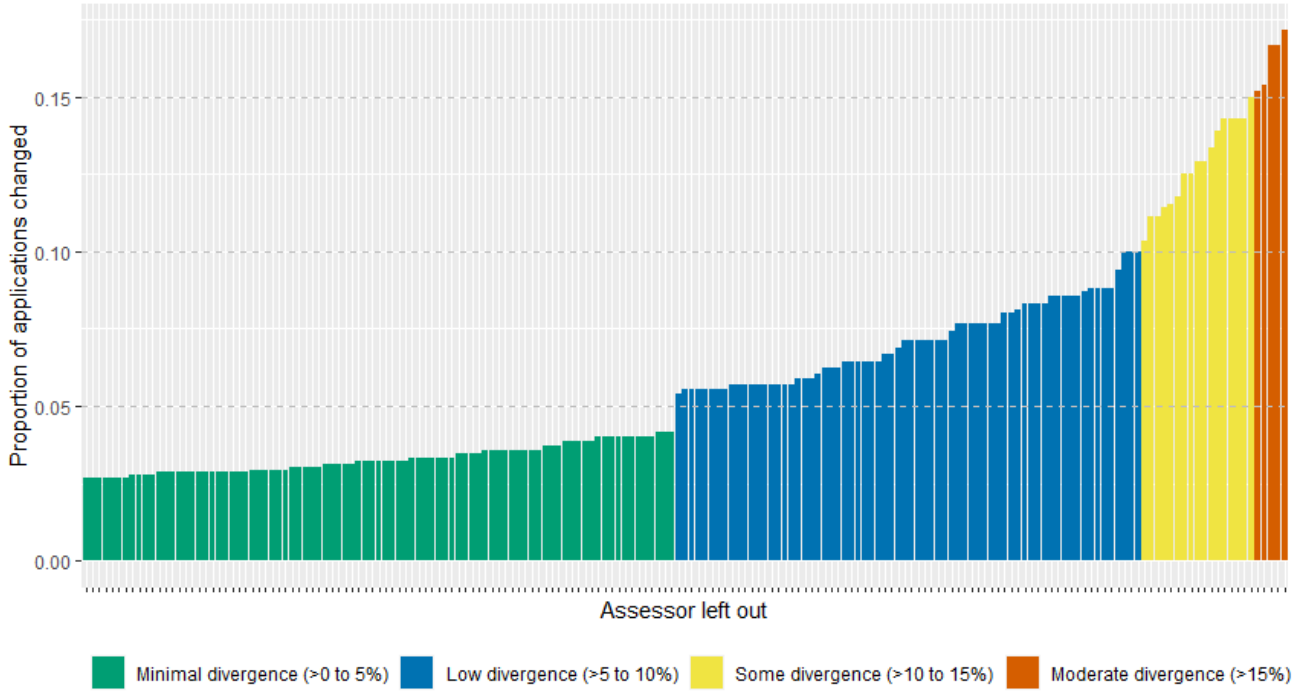
Table 7 below shows the number and percentage of assessors categorised by the percentage of applications that would have a different potential outcome if their scores were excluded.

Table 7: Number and percentage of assessors by divergent assessor category in Leave One Assessor Out analysis

Assessor category (Percentage of funding outcomes changed)	No divergence (0%)	Minimal divergence (>0–5%)	Low divergence (>5–10%)	Some divergence (>10–15%)	Moderate divergence (>15%)	Total
Number of assessors	109	88	71	17	5	290
Percentage	37.6%	30.3%	24.5%	5.9%	1.7%	100.0%

Figure 4 below shows the proportion of applications reviewed by each assessor that had a different potential funding outcome (i.e. changed from funded to unfunded or vice versa) if their scores were excluded in this analysis.

Figure 4: Proportion of applications reviewed by each assessor that have a different potential outcome when the Leave One Assessor Out method is applied.



The impact of excluding an assessor's scores depends on the proximity of an application to the funding cut-off. Applications with a final score close to the cut-off are more sensitive to changes in their scores and more likely to have a different outcome if an assessor's score is excluded. Alternatively, if an application is far above or below the funding cut-off, then an assessor could score differently to the other assessors without affecting the funding outcome.

When four or five assessors' scores are combined, each assessor contributes 25% or 20% respectively of an application's final score. This means that, if an assessor who scores more than 3 away from the mean final score of an application is excluded, the application's final score would change by a maximum of 0.75.

3.2.4 Influential Assessors

PRAC enquired about the overlap between the divergent assessors in the Assessor Generosity analysis and those who had the most impact on outcomes, as identified in the Leave One Assessor Out analysis.

NHMRC investigated the intersection of these two analyses to quantify the number of individuals who had a strong impact on outcomes.

The approach used in this analysis was:

- group assessors by their generosity classification (using 95% threshold as shown in Table 6)
- determine the divergent assessor category (as shown in Table 7) for each assessor
- tabulate the results.

This analysis revealed that:

- Most assessors (251/290 in 2019 and 279/321 in 2020) affected the outcome of no more than 10% of the applications that they assessed and did not show any obvious patterns in their scoring behaviour (i.e. they were classified as Likely Generous, Balanced or Likely Critical) (green cells in Table 8 below).
- One assessor in 2019 changed the outcome of more than 15% of the applications they reviewed and was classified as Mostly Critical (orange cell).

Table 8: Intersection of assessor generosity and divergence (Leave One Assessor Out analysis) in the 2019 Investigator Grant round

	Mostly Generous	Likely Generous	Balanced	Likely Critical	Mostly Critical	Total
No divergence (0%)	2	45	35	27	0	109
Minimal divergence (>0–5%)	4	27	23	33	3	90
Low divergence (>5–10%)	8	19	16	26	0	69
Some divergence (>10–15%)	2	4	2	7	2	17
Moderate divergence (>15%)	0	0	0	4	1	5
Total	**16**	**95**	**76**	**97**	**6**	**290**

Table 9: Intersection of assessor generosity and divergence (Leave One Assessor Out analysis) in the 2020 Investigator Grant round

	Mostly Generous	Likely Generous	Balanced	Likely Critical	Mostly Critical	Total
No divergence (0%)	4	43	33	30	1	111
Minimal divergence (>0–5%)	1	35	34	34	1	105
Low divergence (>5–10%)	3	24	18	28	1	74
Some divergence (>10–15%)	1	7	4	12	3	27
Moderate divergence (>15%)	0	1	0	3	0	4
Total	**9**	**110**	**89**	**107**	**6**	**321**

As this analysis is based on the results of analyses in Sections 3.2.2 and 3.2.3, the same considerations apply.

3.2.5 Drop One Assessment Out

PRAC discussed concerns raised by the sector that an application's outcome can be overly influenced by an individual assessment – that is, among all assessors reviewing the same application, one assessor could cause the application to be funded or not funded.

Based on technical advice from members, NHMRC analysed whether application funding outcomes change if one assessment is excluded from the final score calculations.

This Drop One Assessment Out analysis differs from the Leave One Assessor Out analysis in Section 3.2.3. The current analysis examines the impact of unique assessments (i.e. one assessor's review of one application) on application final scores whereas the previous Leave One Assessor Out analysis examined the impact of individual assessors across the scheme (i.e. one assessor's review of multiple applications).

The method to conduct this analysis was:

- systematically select each application and exclude one of its assessments at a time in a selected round
- every time an assessment is removed, calculate a new final score (i.e. Drop One Assessment Out score) for that application by averaging the remaining scores from the other assessors

- classify applications as 'funded' if the new final score is higher than the actual funding cut-off for that round or 'unfunded' if the new final score is lower than the actual funding cut-off for that round
- compare the hypothetical Drop One Assessment Out funding outcome to the actual outcome for the application.

This analysis revealed that:

- 87.8% of 2019 Investigator Grant *applications* would have had the same outcome and 12.2% a different outcome if any one of their assessments was excluded.85.9% of 2020 Investigator Grant *applications* would have had the same outcome and 14.1% a different outcome if any one of their assessments was excluded.
- Approximately 4.0% of *assessments* in either grant round changed funding outcomes if excluded.

Table 10 below shows the number and percentage of applications that cross the funding cut-off (i.e. change from funded to unfunded or vice versa) when any one assessment is dropped out in the 2019 and 2020 Investigator Grant rounds.

Table 11 shows the number and percentage of assessments that are responsible for the change in application outcomes in the 2019 and 2020 Investigator Grant rounds.

Table 10: Number and percentage of applications moving across the funding cut-off when any one assessment is dropped out

Year	Level	Moving above the line[13]	Moving below the line	Moving above or below the line
2019	EL1	50/616 (8.1%)	37/616 (6.0%)	87/616 (14.1%)
	EL2	34/475 (7.2%)	16/475 (3.4%)	50/475 (10.5%)
	L	49/764 (6.4%)	41/764 (5.4%)	90/764 (11.8%)
	Total	133/1,855 (7.2%)	94/1,855 (5.1%)	227/1,855 (12.2%)
2020	EL1	63/669 (9.4%)	35/669 (5.2%)	98/669 (14.6%)
	EL2	40/391 (10.2%)	15/391 (3.8%)	55/391 (14.1%)

[13] Applications that changed outcomes to 'funded' are classified as 'Moving above the line' and those that changed to 'unfunded' are classified as 'Moving below the line'. The sum of these two groups is classified as 'Moving above or below the line'.

	L	58/718 (8.1%)	39/718 (5.4%)	97/718 (13.5%)
	Total	161/1,778 (9.1%)	89/1,778 (5.0%)	250/1,778 (14.1%)

Table 11: Number and percentage of assessments that, if excluded, cause applications to move across the funding cut-off

Year	Level	Moving above the line	Moving below the line	Moving above or below the line
2019	EL1	75/3,013 (2.5%)	63/3,013 (2.1%)	138/3,013 (4.6%)
	EL2	50/2,325 (2.2%)	27/2,325 (1.2%)	77/2,325 (3.3%)
	L	76/3,698 (2.1%)	74/3,698 (2.0%)	150/3,698 (4.1%)
	Total	201/9,036 (2.2%)	164/9,036 (1.8%)	365/9,036 (4.0%)
2020	EL1	86/3,268 (2.6%)	57/3,268 (1.7%)	143/3,268 (4.4%)
	EL2	51/1,892 (2.7%)	34/1,892 (1.8%)	85/1,892 (4.5%)
	L	76/3,520 (2.2%)	68/3,520 (1.9%)	144/3,520 (4.1%)
	Total	213/8,680 (2.5%)	159/8,680 (1.8%)	372/8,680 (4.3%)

The impact of removing an assessment depends on the proximity of an application to the funding cut-off. Applications with a final score close to the funding cut-off are more sensitive to changes in their scores and more likely to have a different outcome if an assessment is excluded. Alternatively, if an application is far above or below the funding cut-off, then an assessor could score differently to the other assessors without changing the outcome.

This analysis investigated changes to application scores with reference to the funding cut-off to infer whether applications might have a different outcome if an assessment is ignored. However, in practice, changing the outcome of any application will also affect the outcome of other applications due to the ceiling on the scheme budget – that is, the cost of funding one application must be counteracted by not funding another application.

3.2.6 Estimation of the Signal-to-Noise Ratio

Based on technical advice from PRAC members, ONHMRC estimated the degree to which assessors provide true scores (i.e. scoring based on the actual quality of applications) compared to assessment error (i.e. any systematic assessor behaviour and/or random noise that results in scores differing from the unobserved true quality of applications).

The methods provided by members estimated the variance in assessor scores due to two factors:

- application quality (i.e. true differences in the merits of applications)
- assessor behaviour (i.e. tendencies of assessors to score higher or lower than their peers based on their interpretation of the assessment criteria and category descriptors) and 'noise' (i.e. unexplained random error).

This involved fitting random/mixed effect models (see Appendix C for full method) based on the level of Investigator Grant applications each assessor reviewed to quantify how accurately assessors scored applications.

These models were used to calculate two key metrics:

- The percentage of variation in scores due to application quality (the signal-to-noise ratio). The signal-to-noise ratio ranges from 0% to 100%. The larger the ratio, the less variation is due to possible assessor effects or random noise, thus the more accurately assessors score applications. This metric is called single-rater reliability in some studies.[14]
- The percentage of variation in scores due to systematic assessor effects. This reflects the extent to which assessor generosity or negativity affects the scores they provide.

In an ideal (yet unachievable) system, the proportion of variation in scores due to application quality would be 100% and there would be 0% variation due to assessor behaviour or noise.

This analysis revealed that:

- L assessors had larger signal-to-noise ratios than EL assessors in both the 2019 and the 2020 Investigator Grant rounds (see Table 12 below). This implied that L assessors tended to score more accurately to the true quality of the applications they reviewed.
- The percentages of variation due to systematic assessor effects were higher for EL assessors than for L assessors in both rounds.

The signal-to-noise ratio can also be used to calculate reliability with a defined number of assessors, using the Spearman-Brown equation.[15] This would show that with four or five assessors the reliability is quite high.

---

[14] Marsh HW, Jayasinghe UW and Bond NW. Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *Am Psych* 2008, 63(3):160-168. doi: 10.1037/0003-066X.63.3.160

[15] Visscher PM and Yengo L. The effect of the scale of grant scoring on ranking accuracy [version 2; peer review: 2 approved with reservations] *F1000Research* 2023, 11:1197 https://doi.org/10.12688/f1000research.125400.2

Table 12: Key metrics in the 2019 and 2020 Investigator Grant rounds by level

Year	Investigator level	Average number of assessors per application	Percentage of variation due to application quality	Percentage of variation due to systematic assessor effects	Percentage of variation due to noise
2019	EL	4.89	31.6%	26.1%	42.4%
	L	4.84	42.2%	22.7%	35.1%
2020	EL	4.87	34.2%	19.7%	46.2%
	L	4.90	39.2%	17.0%	43.8%

This analysis assumed that assessor effects and the quality of applications are independent of each other. However, in reality this might not always be the case (e.g. an assessor might tend to score more critically overall but then score very generously for one or more specific applications). This analysis does not investigate whether there is an interaction effect of these variables.

This analysis does not attempt to identify individual assessor effects (such as whether an assessor scores more generously or not) or which applications had outlier scores.

Although a higher signal-to-noise ratio means that the final score is closer to the true quality of an application, interpretation of signal-to-noise ratio values is subjective as there is no fixed threshold to determine whether the values are problematic or not.

3.3 Methods to modify assessor scores

The previous section showed that there are possible assessor effects in the data analysed here. In the following section, several methods to modify assessor scores were considered to investigate their effects on funding outcomes.

3.3.1 Normalising assessor scores

PRAC members discussed concerns that systematic assessor behaviour (i.e. some individuals consistently scoring more generously or critically than their peers) could affect which applications are funded.

PRAC members considered possible methods to normalise assessor scores so that all assessors in each round (and level in the case of Investigator Grants) had the same mean score and standard deviation. In principle, this would ensure that assessors have a similar distribution of scores.

The method used for this analysis was:

- classify assessors into three groups based on which Investigator level application they reviewed (EL1, EL2 and L)
- calculate the mean score and standard deviation for each assessor within each level

- calculate the combined mean score and standard deviation across all assessors within each level
- calculate normalised assessor scores (see Appendix C for formula) so that all assessors have the same mean score and standard deviation within each level; these match the values calculated in the step above for the combined mean score and standard deviation for each level
- derive each application's normalised score by averaging the normalised assessor scores for the four or five assessors who reviewed each application
- classify applications as 'funded' (in rank order) until the same number of applications (by level) is identified as were actually funded in the round (note that this is a simplified funding process)
- compare the hypothetical funding outcome to the actual outcome for each application.

The analysis revealed that:

- Normalising assessor scores in the 2019 Investigator Grant round would not change the outcome of 95.4% of applications; 4.6% (86) of applications would have a different outcome.
- Normalising in the 2020 Investigator Grant round would not change the outcome of 93.8% of applications; 6.2% (110) of applications would have a different outcome.
- The number of applications that would change their funding outcomes in both years is not substantial given that differences in outcomes are strongly influenced by the clustering of many application scores around the funding cut off.

Table 13: Number and percentage of applications moving across the funding cut-off after normalising application scores in the 2019 and 2020 Investigator Grant rounds

Year	Level	Moving above the line	Moving below the line	Moving above or below the line
2019	EL1	15/616 (2.4%)	15/616 (2.4%)	30/616 (4.9%)
	EL2	9/475 (1.9%)	9/475 (1.9%)	18/475 (3.8%)
	L	19/764 (2.5%)	19/764 (2.5%)	38/764 (5.0%)
	Total	43/1,855 (2.3%)	43/1,855 (2.3%)	86/1,855 (4.6%)
2020	EL1	24/669 (3.6%)	24/669 (3.6%)	48/669 (7.2%)
	EL2	13/391 (3.3%)	13/391 (3.3%)	26/391 (6.6%)

	L	18/718 (2.5%)	18/718 (2.5%)	36/718 (5%)
	Total	55/1,778 (3.1%)	55/1,778 (3.1%)	110/1,778 (6.2%)

By ensuring that all assessors have the same mean and standard deviation across the applications they review, normalisation may adjust for some aspects of assessor behaviour. It assumes, however, that assessors review the same distribution in application quality. This assumption might not be appropriate because assessors were not randomly assigned to applications but were assigned based on their suitability and absence of conflicts of interest.

Currently there is no way of evaluating whether changes in application outcomes due to normalising assessor scores are appropriate or not.

This analysis investigates the impact of normalising scores with reference to a fixed number of applications being funded. However, funding outcomes are more complex and changing the outcome of any application will affect the outcome of other applications due to the ceiling on the scheme budget – that is, the cost of funding one application must be counteracted by not funding another application.

3.3.2 Rescaling assessor scores

PRAC members considered another method to rescale assessor scores in which all assessors have the same mean score across their assessments in each year, but each assessor has their own standard deviation. This method adjusts for assessors who are mostly generous or mostly critical on average but does not standardise the spread of assessors' scores.

This analysis is an extension of the normalisation analysis in Section 3.3.1.

The method used for this analysis was:

- classify assessors into three groups based on which Investigator level application they reviewed (EL1, EL2 and L)
- calculate the mean score for each assessor within each level
- calculate the combined mean score across all assessors within each level
- calculate the rescaled assessor scores (see Appendix C for formula) so that all assessors have the same mean score within each level; these match the values calculated in the step above for the combined mean score for each level
- derive each application's rescaled score by averaging the corresponding assessor scores for the four or five assessors who reviewed the application
- classify applications as 'funded' (in descending rank order) until the same number of applications (by level) is identified as were actually funded in the round (note that this is a simplified funding process)
- compare the hypothetical funding outcomes to the actual outcome for each application.

This analysis revealed that:

- Rescaling assessor scores in the 2019 Investigator Grant round would not change the outcome of 94.0% of applications; 6.0% (110) of applications would have a different outcome.
- Rescaling in the 2020 Investigator Grant round would not change the outcome of 93.2% of application; 6.8% (120) of applications would have a different outcome.
- By comparison, 4.6% (86) and 6.2% (110)[16] of applications in the 2019 and 2020 Investigator Grant rounds would have a different outcome if the previous normalisation method was used in which assessors shared the same mean and standard deviation across their assessments.

Table 14: Number and percentage of applications moving across the funding cut-off after rescaling application scores to the same mean in the 2019 and 2020 Investigator Grant rounds

Year	Level	Moving above the line	Moving below the line	Moving above or below the line
2019	EL1	17/616 (2.8%)	17/616 (2.8%)	34/616 (5.6%)
	EL2	10/475 (2.1%)	10/475 (2.1%)	20/475 (4.2%)
	L	28/764 (3.7%)	28/764 (3.7%)	56/764 (7.4%)
	Total	55/1,855 (3.0%)	55/1,855 (3.0%)	110/1,855 (6.0%)
2020	EL1	23/669 (3.4%)	23/669 (3.4%)	46/669 (6.8%)
	EL2	14/391 (3.6%)	14/391 (3.6%)	28/391 (7.2%)
	L	23/718 (3.2%)	23/718 (3.2%)	46/718 (6.4%)
	Total	60/1,778 (3.4%)	60/1,778 (3.4%)	120/1,778 (6.8%)

Unlike normalisation, rescaling only relies on the assumption that assessors review applications of the same quality on average. However, this assumption might still not be appropriate because assessors were not randomly assigned to applications.

[16] See Table 13 in Section 3.3.1 for more details on the normalisation results.

This analysis investigates the impact of rescaling scores in reference to a fixed number of applications being funded. However, funding outcomes are more complex and changing the outcome of any application will affect the outcome of other applications due to the ceiling on the scheme budget – that is, the cost of funding one application must be counteracted by not funding another application.

4. Considerations

The following material was provided by NHMRC to inform PRAC's discussions.

4.1 International examples

4.1.1 How do we compare?

An environmental scan was performed to determine how international research funding agencies use peer reviewer assessments to reach funding decisions, in order to inform discussion of policies for normalisation or calibration of peer reviewer scores in NHMRC grant schemes. This scan focused on major project funding schemes (similar to NHMRC Ideas and Investigator Grants) run by international funding agencies but findings are broadly relevant to other schemes of the respective funding agency.

Summary of findings

A minority of funding agencies normalise or calibrate scores using several methods. Normalisation may be used to adjust for differences in scoring between peer review committees (e.g. Health Research Council, New Zealand); to adjust for 'score creep' over time for a peer review committee (e.g. National Institutes of Health, USA); or to distribute applications across all scoring levels (e.g. Australian Research Council). However, most agencies use methods similar to NHMRC, where applications are ranked based on a mean or median of raw peer review scores and this ranking is used to decide funding outcomes.

Normalisation of scoring

- The **Health Research Council** (HRC; New Zealand) applies statistical normalisation to peer reviewer scores following meetings of the peer review committees at both the expression of interest and full application stages. This is done to minimise the effect of scoring variation between committees, if two or more committees are appointed to assess applications within a single Research Investment Stream of a funding scheme. Scores are normalised using the mean and standard deviation of scores across the individual committee, corrected for the mean and standard deviation of the larger distribution of scores across all committees.
    - Scores are used to produce a ranked list of applications, identifying those that are of insufficient quality to be funded.
    - Final scores and application rankings from committees within each Research Investment Stream are collated and presented to the Grant Approval Committee for funding decisions.
    - Normalisation of scores is not used in HRC schemes using lottery systems, such as Explorer Grants.

- The **Australian Research Council** (ARC) normalises peer review scores from General Assessors. Normalisation redistributes peer reviewers' raw scores across the entire spectrum of scores, so that their distribution matches the percentage of applications that should fall into each scoring band, as identified in the scoring matrix for each ARC funding scheme. This process is intended to improve discrimination between scores of applications in order to determine their rank and is not specifically intended to correct for scoring differences between peer reviewers or peer review committees.
    - Normalisation of general assessor scores occurs before peer review committee meetings. Overall application scores are calculated as the median of the average of General Assessor ratings and the average of Detailed Assessor ratings. Overall application scores are then used to rank applications and these rankings inform funding decisions.
    - Scores for smaller grant opportunities are not normalised.

Committee percentile rankings

- The **National Institutes of Health** (NIH; USA), for some schemes, assign applications a rank percentile based on their overall impact score. A percentile is the approximate percentage of applications that received better impact scores than that particular application from the peer review committee during the past year. This ranks applications relative to other applications reviewed by the peer review committee over its last three meetings in order to counter 'score creep' (where committees tend to give higher scores over time) and makes it easier to discriminate between exceptional applications. This is not intended to correct for differences in methods of scoring between peer reviewers or committees.
    - Peer reviewers assign applications a score for each scheme criterion and an overall impact score. No formula is used to derive the overall impact score from the individual criterion scores and reviewers are instructed to weight the criteria as they see fit.
    - Reviewers are encouraged to use the entire scoring range in order to distinguish accurately between applications.

No scoring adjustment

- **Canadian Institutes of Health Research** (CIHR) peer review committees first determine a consensus score for each application (either through agreement by the assigned peer reviewers or as a mean of individual reviewer scores). All committee members then vote ± 0.5 of the consensus score to reach the final score. Applications are ranked based on their final score and this ranking is considered by the Scientific Council to recommend funding decisions.
  The now discontinued Foundation Grant scheme (which provided large grants to elite researchers) previously asked reviewers to produce individual rankings to prevent clustering and help distinguish between exceptional applications.
- The **Medical Research Council** (MRC; United Kingdom) calculates a median score of all individual peer reviewer scores, rounded to the nearest whole number. This median score is used to rank all proposals under consideration for funding decisions. (Note: If NHMRC adopted this approach it would likely result in a large number of tied applications.)

4.2 Changes to the assessment process

PRAC asked whether aspects of the assessment process could be changed to address some of the issues raised. PRAC's questions and the information provided by NHMRC to help inform PRAC's discussions are provided below.

4.2.1 Could more assessors be recruited per application?

PRAC asked whether more assessors could be recruited to review each application to reduce the impact of individual assessors on application outcomes.

Information provided by NHMRC:

The potential benefit of increasing peer reviewer numbers needs to be balanced against peer review burden and the sustainability of higher peer review demands on the research sector.

NHMRC recognises that participating in peer review, particularly for the large schemes such as Investigator Grants or Ideas Grants, is a significant undertaking in time, effort and opportunity cost. NHMRC also recognises that changes made to peer review processes to improve rigour are often a trade-off between their anticipated value and peer review burden. It seeks to find an appropriate balance.

For example, in the 2021 round of Ideas Grants, the number of assessors assigned to each application was increased from four to five. While the intention of this change was to improve confidence in the scoring outcome and decrease the impact of outlier scores, this change alone required 136 more assessors than in 2020, despite almost 400 fewer applications being submitted to the 2021 round. The additional assessment performed for each application was estimated to add a combined 329 days of time away from research and other activities to undertake peer review.[17]

4.2.2 Should outlier assessments be excluded?

PRAC asked whether outlier assessments should be excluded to eliminate their impact on application outcomes.

Information provided by NHMRC:

NHMRC's view is that every assessment should count for two reasons.

First, peer review draws significantly on the health and medical research sector and Australia has a small sector compared to many other countries. NHMRC relies on researchers (particularly those who receive NHMRC funding) and the Administering Institutions that employ them to make time available for peer review. The expertise and experience that researchers across all career stages and research areas bring to peer review are invaluable and underpin the selection of the best applications for funding. Researchers sacrifice time away from their own research and personal lives to undertake

[17] One additional assessor for 2635 applications x average 3 hours per review: 2635 x 3 hours = 7,905 hours = 988 workdays (at 8 hours per day)

peer review. NHMRC wants to ensure that the effort each assessor dedicates to this task is valued and appropriately used to inform the outcome of an application.

Second, as outlined in Section 3.2.2, there are many possible reasons for outlier scores. For example, an assessor may have specific expertise that leads them to recognise strengths or weaknesses in the project design that might not be obvious to assessors with different expertise. Outlier scores are not necessarily incorrect.

PRAC noted that it is not currently possible to determine programmatically whether divergent scores are genuine (i.e. an assessor had a valid reason for scoring differently), malicious (i.e. intentional bias or gamesmanship), careless (i.e. an assessor did not read the application properly) or mistakes (i.e. an assessor misinterpreted the seven point scale or inadvertently entered the wrong value in Sapphire). A process to identify whether scores are valid would allow a decision to be made to retain valid outlier scores and to exclude incorrect or inappropriate scores. In the absence of such a mechanism, PRAC agreed that 'every review counts' given the time, dedication and expertise required of peer reviewers to provide them, and that it would not be appropriate to discount or discard outlier assessments at this time.

4.2.3 Could borderline applications (or those whose outcomes would be changed by exclusion of an assessment) be reviewed by additional assessors or a panel?

PRAC asked whether a further round of assessment could be implemented for applications close to the funding cut-off or those that have received outlier scores, to improve confidence in the final score and/or to resolve any differences of opinion among the first-round assessors. For example, this could be done by forming a Grant Review Panel to review the first round of reviews, or by undertaking a second stage of review by different assessors, with a focus on applications around the funding cut-off.

PRAC noted that more assessors (i.e. a second peer review stage undertaken by a new set of assessors) would not necessarily be better as the quality of assessment depends on appropriate expertise. A second round of review would also introduce another layer of chance and could be considered unfair if it was not applied to all applications. PRAC also noted that the funding cut-off is only known once the peer review process is complete – particularly for those schemes where individual grant budgets vary (e.g. Ideas Grants) so that the number of grants that can be awarded is not known in advance.

Information provided by NHMRC:

NHMRC outlined the impact that an additional round of assessment would have on assessor recruitment and timelines.

To outline the current process for NHMRC's two largest schemes (Investigator and Ideas Grants), assessor recruitment commences several months ahead of the rounds opening for applications because of the number of assessors required to review up to 2,000 Investigator Grant applications and up to 3,000 Ideas Grant applications. Initial recruitment is based on modelling of application numbers and the spread of expertise required in previous rounds. Following submission of minimum data, additional assessors are usually recruited to ensure adequate coverage of the number and spread of expected applications. Once submissions close, assessors are asked to declare their suitability and conflicts of interest (CoI) against a subset of applications before being

assigned the group of applications they will independently review. The matching of applications to assessors aims to make best use of each assessor's expertise while also balancing workloads among assessors.

Assigning additional assessors or holding an assessment panel meeting (even if only for applications around the funding cut-off) would require significant additional assessor recruitment beyond that already required for large schemes. For example, Ideas Grants drew on more than 700 assessors in the 2021 round with four assigned per application. In 2022 with the increase to five assessors per application, more than 800 were required, despite a significant drop in the number of submitted applications.

Significant additional time would also be required. The second round of recruitment would need to occur after the conclusion of the initial assessment, when applications that fall around the funding cut-off can be identified. Time would also be required for suitability and CoI declarations before the second stage of peer review could commence. An additional assessment period would rely on high sector commitment to deliver within further compressed timeframes.

In 2021, more than 34% of assessors invited to review Investigator Grant applications and about 28% of assessors invited to review Ideas Grant applications were not used because they declined, did not respond to ONHMRC's invitation, withdrew after acceptance (often during the assessment period) or declared low suitability (and thus could not be matched with a minimum number of applications for assessment). Assessors who withdrew had a wide range of reasons for doing so, compounded in the 2021 rounds by COVID-19-related issues (such as extended lockdowns in several jurisdictions). The impact of so many assessors withdrawing (particularly if it was late in the assessment phase of the Ideas Grant round) was significant for final assessment numbers, as replacement assessors needed to be found for approximately 200 Ideas Grants applications. This placed additional burden on other assessors who were asked to pick up additional applications late in the process. Due to conflicts of interest, low suitability and withdrawals at a very late stage, replacement assessors could not be found for all applications where one or more of the assigned assessors withdrew.

These challenges need to be taken into account when considering introduction of additional peer review processes given the already high peer review demand on the sector and the impact on peer reviewer time, availability and commitment.

5. Conclusions

PRAC was formed to advise NHMRC on the use of assessor scores to award competitive grants, particularly in the two largest schemes, Investigator Grants and Ideas Grants, where grants are now awarded based on the scores of independent assessors without panel meetings.

Over 14 months from October 2020, PRAC reviewed a range of analyses of Investigator Grant and Ideas Grant scores in the 2019 and 2020 rounds to investigate the impact of scoring behaviour on application outcomes. Members explored the effects of several approaches to the statistical management of scores, including the exclusion of 'outlier' scores and mechanisms to recalibrate scores to reduce variance.

5.1 Specific conclusions

PRAC drew a number of specific conclusions from the analyses while noting the limitations and caveats outlined throughout Section 3 above.

Score distributions and outlier scores

- Final scores of Investigator Grant and Ideas Grant applications were approximately normally distributed and clustered around the mean. As small differences in final score around the funding cut-off can determine whether or not an application is funded, the system is sensitive to variation between individual assessor scores near the cut-off.
- In the case examined (2019 Investigator Grants), 0.3% of assessor scores were outliers (i.e. they differed markedly from other assessors' scores of the same application) when defined using NHMRC's traditional threshold of 2 or more points away from an application's mean assessor score.
- Some assessors tend to score higher or lower than other assessors of the same applications ('Generosity analysis'). In the case examined (2019 Investigator Grants), 5.5% of assessors were defined as 'mostly generous' and 2.1% as 'mostly critical' based on positive or negative deviation of at least 95% of their scores from the mean assessor score for the same applications.

The impact of outlier scores on application outcomes

- Systematic exclusion of each *assessor* in the 2019 Investigator Grant round ('Leave One Assessor Out' analysis) showed that about 62% of assessors affected the outcome for some of the applications they assessed; 7.6% of assessors affected the outcome for more than 10% of the applications they assessed.
- Comparison of the Leave One Assessor Out and Generosity analyses for the 2019 and 2020 Investigator Grant rounds showed that the most influential assessors (i.e. those who affected the outcome of more than 10% of the applications they assessed) were generally not from the 'mostly generous' or the 'mostly critical' group.
- Systematic exclusion of each *assessment* in the 2019 and 2020 Investigator Grant rounds ("Drop One Assessment Out') showed that about 12% and 14% of applications respectively would have had a different outcome if any of their assessments was excluded. About 4% of assessments changed the application outcome if excluded.
- Therefore there was no change in funding outcome for over 85% of applications when individual assessments were left out.
- As expected, in both the Leave One Assessor Out and Drop One Assessment Out analyses, applications whose score is close to the funding cut-off are more sensitive to changes to their outcome when an assessment is excluded.
- Modelling to estimate the percentage of variation in scores due to application quality (signal-to-noise ratio) versus systematic assessor effects (generosity or negativity) and random variation between assessors (e.g. due to differences in

interpretation of category descriptors) suggested that random noise was a major contributor to score variation, especially for Emerging Leadership applications, while systematic assessor effects contributed the least. The signal-to-noise ratio was higher for Leadership than for Emerging Leadership applications.

- Even under an ideal scoring system, where there are no systematic assessor effects and the only random element is unexplained random noise, we would expect to see variation between assessors of significant magnitude to alter grant outcomes if an assessor or assessment is excluded.
- Outlier scores should not be excluded in the absence of a way to determine whether they were an appropriate reflection of differences in scientific judgement, unless there was clear evidence to justify their exclusion (such as assessor error).

Adjustment of scores to remove systematic assessor effects

- Recognising that systematic assessor effects (generosity or negativity) contribute to some of the variation in scores for a given application, two methods were tested to adjust scores to remove these effects.
  - Normalising assessor scores within each category of Investigator Grant applications (EL1, EL2 and Leadership) so that all assessors have the same mean and standard deviation as the actual combined mean and standard deviation for the category would change the outcome for some applications (3.8 to 5.0% in 2019 and 5.0 to 7.2% in 2020).
  - Rescaling assessor scores within each category so that all assessors have the same mean as the actual combined mean for the category but retain their own standard deviation would also change the outcomes for some applications (4.2 to 7.4% in 2019 and 6.4 to 7.2% in 2020).
  - The value of modifying scores to adjust for assessor differences was not clear. Normalisation or rescaling would change outcomes and should only be considered following experimentation, further analysis and discussion with the sector.
- There is noise in the system which cannot be reduced to zero. Some noise will be due to the challenge of distinguishing between highly competitive and complex applications. Methods to rescore those applications would not necessarily give a more repeatable or more accurate outcome.
- Recruitment of additional assessors or discussion for applications close to the funding cut-off to reduce variation in scores would introduce inequities and significantly increase the time taken to complete the funding round.

5.2 General observations

PRAC made the following general observations.

- While independent peer review against published criteria is the international gold standard for making grant funding decisions, there is little agreement across the research sector, nationally or internationally, on the gold standard or best practice for the design of the peer review process and how peer review recommendations should be used to make funding decisions.

- Variation between assessors' scores for an application is an inherent part of the peer review process as it relies on individual judgement. There is evidence that the research sector accepts a degree of variation between assessors' scores of the same application.
- Outlier scores are not necessarily incorrect or due to systematic assessor effects but may reflect the specific expertise and judgement of the assessor. Assessor training may help to reduce variation but would not be expected to eliminate outliers.
- Every review counts. All assessors should be reassured of the importance of their work.
- PRAC did not recommend any major changes to NHMRC's current peer review processes. Members identified strategies (e.g. rescaling, normalisation) to be considered by NHMRC as part of ongoing improvements of the peer review process. PRAC recommended that any future changes should be based on evidence and, where possible, on experiments to test their effectiveness.
- PRAC also recommended increased transparency about the reasons for NHMRC's processes and analyses such as those presented in this report would increase the sector's understanding of NHMRC peer review.

5.3 Impact of the work of PRAC

NHMRC has valued the careful consideration, insight and advice provided by PRAC. While PRAC did not recommend any major changes, Members provided a range of perspectives on factors affecting NHMRC peer review, while noting the need to balance rigour with timeliness and minimising burden on the sector. In response, NHMRC has implemented several changes to its peer review processes to increase quality, transparency and accountability, as outlined in Appendix D. NHMRC will continue to draw on PRAC's work and other data, feedback from applicants and assessors, and the experience of other national and international funders in the further development of its peer review processes and training activities.

Appendix A – PRAC Membership

Name	Gender	State	Role	Institution
Professor Caroline Homer AO	F	NSW	Chair	University of Technology Sydney
Professor Emily Banks AM	F	ACT	Member	The Australian National University
Professor Adrian Barnett	M	QLD	Member	Queensland University of Technology
Professor Tony Blakely	M	VIC	Member	The University of Melbourne
Professor Tanya Chikritzhs	F	WA	Member	Curtin University
Professor Philip Clarke	M	VIC	Member	University of Oxford The University of Melbourne
Professor Peter Visscher	M	QLD	Member	The University of Queensland
Professor Tania Winzenberg	F	TAS	Member	University of Tasmania

Under the *Public Governance, Performance and Accountability Act 2013* (PGPA Act), s16A, s16B and s29, PRAC members disclosed their interests to the whole Committee and Chair. Members were provided with a list of disclosed interests before each meeting and were asked to consider if any interest could affect their capacity to bring an independent mind to bear on the matters being considered by the Committee.

Members considered a register of disclosed interests and agreed that no management strategies were required for disclosed interests.

Appendix B – Peer review processes and data selection

The following peer review processes were used in the Investigator and Ideas Grant schemes in 2019 and 2020.[18]

Investigator Grants

2019

- Potential assessors provided declarations of conflicts of interest and suitability to review groups of applications.
- Based on their declarations, each assessor was allocated 24–39 applications.
- Each application was assessed and scored independently by up to 5 assessors.
- Once they had seen the aggregated scores from all assessors for their assigned applications, assessors could nominate up to two for 'discussion by exception' (DBE) in a videoconference; this step led to rescoring of 73 applications.
- Mean DBE scores replaced the original assessor scores to produce three ranked lists of 'final scores' (for Emerging Leadership 1, Emerging Leadership 2 and Leadership applications) from which funding recommendations were developed.

2020

- Potential assessors provided declarations of conflicts of interest and suitability to review groups of applications.
- Based on their declarations, each assessor was allocated 14–33 applications.
- Each application was assessed and scored independently by up to 5 assessors.
- Mean assessor scores ('final scores') were used to produce three ranked lists (for Emerging Leadership 1, Emerging Leadership 2 and Leadership applications) from which funding recommendations were developed.

Ideas Grants

2019

- Potential assessors provided declarations of conflicts of interest and suitability to review groups of 150–200 applications.
- Based on these declarations, 39 discipline-based Grant Review Panels (GRPs) of about 15 members were formed.
- Each GRP was allocated 75–100 applications.
- Each application was assessed and scored independently by 4 GRP members.

[18] More detail on the peer review processes outlined here is available in the CEO Communique (February 2021).

- The means of these assessor scores were used to produce a single ranked list of applications.
- The top third of applications were considered by their respective GRP at a face-to-face meeting with all members scoring all applications before the panel.
- Mean GRP scores replaced the original assessor scores to produce a single ranked list of 'final scores' from which funding recommendations were developed.

2020

- Potential assessors provided declarations of conflicts of interest and suitability to review groups of 120–270 applications.
- Each assessor was allocated 25–30 applications.
- Each application was assessed and scored independently by up to 4 assessors.
- Mean assessor scores ('final scores') were used to produce a single ranked list of applications from which funding recommendations were developed.

The following data were used for the analyses considered by PRAC. All applicant and assessor data were de-identified and data were aggregated where necessary to ensure that no personal information could be deduced.

Section	Data selection
3.1.1 Overview of score distributions	Final scores for the 1,854[19] and 1,778[20] applications from the 2019 and 2020 Investigator Grant rounds Final scores for the 2,649[21] and 2,888 applications from the 2019 and 2020 Ideas Grant rounds
3.1.2 Analysis of scores for individual assessment criteria 3.2.1 Outlier scores 3.2.2 Assessor Generosity 3.2.3 Leave One Assessor Out 3.2.4 Influential Assessors	Final scores for the 1,855[22] applications in the 2019 Investigator Grant round
3.2.5 Drop One Assessment Out 3.2.6 Benchmarking and Signal-to-Noise Ratio 3.3.1 Normalising assessor scores 3.3.2 Rescaling	Final scores for the 1,855 and 1,778 applications in the 2019 and 2020 Investigator Grant rounds

[19] Three withdrawn applications were excluded from this analysis. One application that was ruled ineligible after peer review was included because the assessors' scores were considered valid data points.

[20] Two withdrawn applications were excluded.

[21] Only applications seeking NHMRC funding were examined for both 2019 and 2020 Ideas Grant rounds. Withdrawn applications were excluded.

[22] Two withdrawn applications were excluded. One withdrawn application and one ineligible application were included because these applications were removed after peer review and had criterion scores that could still be used for analysis.

Appendix C – Detailed methods and formulae

3.2.2 Assessor Generosity

DeviationRatio is defined as:

- For an application i and assessor j, if assessor j reviewed application i, the assessment score given by assessor j for application i is denoted as S_j^i.
- Let S_{final}^i denote the final score of application i.
- Hence, DeviationRatio of assessor j for application i is calculated as:

$$\text{DeviationRatio}_j^i = \frac{S_j^i - S_{final}^i}{S_{final}^i}$$

3.2.6 Estimation of Signal-to-Noise Ratio

Details of two-way random effect/mixed effect model:

This involved fitting a two-way random effect model for the L applications with assessors and applications being the random effects while assessors' scores by application were the response. The two EL levels were used as fixed effects[23] for EL applications in a mixed effect model.

The two-way random effect model for the L assessors can be expressed as: $Y_{ij} = Q_i + A_j + E_{ij}$,

The mixed effect model for EL assessors can be expressed as: $Y_{ij} = \beta I_{level} + Q_i + A_j + E_{ij}$

where

Y_{ij} = aggregate (weighted) score across scoring criteria for assessor j to application i (final score from one assessor for one application)

Q_i = quality of application

A_j = systematic assessor effect (such as always scoring higher or lower than other assessors)

E_{ij} = noise (such as measurement error and random fluctuations)

I_{level}= 0 for EL1 and 1 for EL2 (EL levels as fixed effect)

It is assumed that Q_i , A_j and E_{ij} are random continuous variables that are normally distributed.

[23] EL1 and EL2 applications were reviewed by the same group of assessors and each assessor would typically review an average of 15 applications from either EL level. To ensure there were sufficient assessments from each assessor, EL1 and EL2 application score data were combined and EL level (categorical variable) was added as a fixed effect in the model.

3.3.1 Normalising assessor scores

Details of calculating normalised assessor scores:

To calculate a normalised assessor score (X'_{ij}) where i represents the i_{th} application and j represents the j_{th} assessor who reviewed an application:

- Calculate the mean score ($\overline{X_j}$) and standard deviation (SD_j) for each assessor within the EL1, EL2 and L levels.
- Calculate the combined mean score ($\overline{X}_{Level}$) and standard deviation (SD_{Level}) across all assessors within each level.
- Calculate normalised assessor scores using the formula:

$$X'_{ij} = \overline{X}_{Level} + \frac{X_{ij} - \overline{X_j}}{SD_j} * SD_{Level}$$

3.3.2 Rescaling assessor scores

Details of calculating rescaled assessor scores:

To calculate a rescaled assessor score (X'_{ij}) where i represents the i_{th} application and j represents the j_{th} assessor who reviewed an application:

- Calculate the mean score ($\overline{X_j}$) for each assessor within the EL1, EL2 and L levels.
- Calculate the combined mean score ($\overline{X}_{Level}$) across all assessors within each level.
- Calculate rescaled assessor scores using the formula:

$$X'_{ij} = \overline{X}_{Level} + X_{ij} - \overline{X_j}$$

Appendix D – Changes to NHMRC peer review processes

Since the 2021 round, application-centric peer review has been used for both the Investigator Grant and the Ideas Grant scheme, in which each application is reviewed independently by the most suitable reviewers for that application. Five assessors are sought for each application. This change has improved the matching of applications to assessor expertise.

In addition, some changes to NHMRC peer review processes have been made in response to PRAC's analyses and advice, as well as feedback from applicants and assessors, as outlined below.

These changes have been introduced to improve the quality of peer review, to address concerns about the impact of outlier scores on funding outcomes, to increase accountability for assessor scores and to provide more feedback from assessors to applicants.

Listed below is the scheme, year and process:

Investigator Grants

2021

- Appointment of Peer Review Mentors and development of Peer Reviewer Mentor video
- Requirement for assessors to provide written comments (with strengthened guidance in 2022)
- Screening to identify outlier scores and verify their accuracy (against comments and with assessors where necessary)

2022

- Peer review briefing webinar

2023

- Sharing of assessor comments with other assessors of the same application (following successful 2022 pilot in Ideas Grants)

Ideas Grants

2021

- Appointment of Peer Review Mentors and development of Peer Reviewer Mentor video
- Increase in target number of assessors for each application from 4 to 5 (see Section 4.2.1)
- Requirement for assessors to provide written comments (with strengthened guidance in 2022)
- Screening to identify outlier scores and verify their accuracy (against comments and with assessors where necessary)

2022

- Peer review briefing webinar
- Sharing of assessor comments with other assessors of the same application

2023

- Further improvements in matching of applications to assessors