PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

<table>
<thead>
<tr>
<th>TITLE (PROVISIONAL)</th>
<th>Using simplified peer review processes to fund research: a prospective study</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTHORS</td>
<td>Herbert, Danielle; Graves, Nicholas; Clarke, Philip; Barnett, Adrian</td>
</tr>
</tbody>
</table>

VERSION 1 - REVIEW

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Guy D. Eslick</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The University of Sydney, Australia</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>28-Apr-2015</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS    | This is an interesting and important study. Much time and effort is wasted in preparing Grant applications as part of a researchers career. Even more time and effort is wasted by national organisations reviewing such proposals. This process must be streamlined and simplified! This study goes some way to initiating this process and I commend the authors for producing a simple and viable option for Grant bodies to follow in the future. The cost savings are substantial! |
|                     | Table 2, a column showing a p-value comparing the 'Simplified' vs 'Journal' panel would be of value, if only to show that there is no difference between the groups. This method would also need to be added to the 'statistical analysis' section of the METHODS. |
|                     | I must say I like the idea of the 'Journal panel', nice and easy, but how would applications be ranked/selected using a two-person approach for a large number of applications (n~4,000)? Any suggestions? |
|                     | Some comparison with the NIH (USA) system would be nice in the Discussion as I believe they have reduced the size of the Grant proposals recently. |

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Stephen Gallo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>American Institute of Biological Sciences</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>01-May-2015</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS    | The authors have conducted a prospective study to look at levels of agreement between three different grant review processes: the standard onsite NHMRC review panel, the simplified and smaller onsite review panel process, and a journal-style review (2 independent reviewers). There are three panels in total, all reviewing the same set of applications. The two experimental panels had shorter applications, fewer reviewers (with and without discussion) |


and greatly reduced costs (in both travel time and in reviewer hours). The results are primarily the percentage levels of inter-panel agreement in funding status of proposals and the costs associated with each review setting.

Optimizing the funding processes of scientific research is crucial to the success of the scientific enterprise. While peer review is the most commonly used decision-making process to determine funding, it is poorly studied. In addition, most of the data in the literature concerning peer review is retrospective; there is a glut of prospective studies. Therefore, the authors are to be congratulated on their efforts to conduct such studies.

However, there are a few concerns that should be addressed in the manuscript.

Firstly, the funding rates for Journal style review vs the NHMRC are very different. The journal panel only found 8.3% of proposals worthy of funding while NHMRC and Simplified panels would find 20.8% and 23.6% of proposals worthy of funding, respectively. This indicates that the journal panel was much harsher in its evaluation of proposals. It is unclear if this is just this particular group of reviewers, or a function of the review setting itself. A relatively recent analysis of grant proposal scoring data from the US National Institutes of Health has suggested that score shifts as a result of panel discussion have practical importance over the funding status of 13% of applications (Martin MR et al, PLoS One. 2010; 5(11): e13526). Thus, it may be that lack of discussion may play a role in the differences seen between Journal and NHMRC panels. In any case, it certainly seems that the statement: “our results indicate that the current lengthy process could be simplified without impacting greatly on funding outcomes” is probably not completely correct for the Journal-review setting, as it seems there is a substantial impact on funding outcomes when switching to the Journal-style review panels. This issue should be discussed in the text.

Secondly, of the proposals that the NHMRC panel did decide to fund, only 40% were in agreement with the Simplified panel’s decisions and only 7% were in agreement with the Journal panel’s decisions. Yet in contrast, 64% of the proposals that the NHMRC panel did NOT decide to fund were in agreement with the Simplified panel’s decisions and 72% were in agreement with the Journal panel’s decisions. The authors state:

“The agreement about which proposals to fund was lower than the agreement about which proposals not to fund. This is partly because many more proposals were not funded than funded.”

While this is surely the case, it does call into question how reviewers make decisions for high versus low quality proposals. It may be very likely that reviewers are better at separating well-constructed proposals from those that are flawed than they are at discriminating between good and great proposals. It may be that decisions based in catching logical flaws and methodological errors may have higher inter-panel reliability than those which are based predominantly in the subjective prediction of success. As the authors point out, others have observed the heterogeneity in reliability across proposal quality. For instance, in the prospective inter-panel reliability study of Fogelhom (J clin epi 65.1 (2012): 47-52.), which the authors quote in
this manuscript, it also found that reliability is “score” dependent:

“The results suggest that reliability is better at the low end of the scale, that is, the consistency in identifying good proposals is lower than for average and poor proposals. The distribution of the scores (success rate) does not affect the sensitivity and specificity. Therefore, our findings suggest that the criteria (or consensus) for poor proposals are clearer than the criteria for good proposals.”

In any case, before a strong case can be made that the experimental review mechanisms are roughly equivalent to the NHMRC with regard to funding outcomes, the authors should further justify why the bulk agreement parameter is used to assess the different review panels. Also, the difference in agreement with the list of fundable NHMRC proposals for Simplified and Journal (40% vs 7%, respectively) should be discussed. It may be that acceptable agreement thresholds for both unfundable and fundable populations may be needed.

Finally, the authors only explore the bulk level of inter-panel agreement and cost as parameters to justify the adoption of one peer review setting over another. However, depending on the stakeholder point of view, there may be other parameters that could also be viewed as important to optimize. For instance, it is unclear if review setting affects the promotion of innovation. It may be that one review setting is superior to the others in fostering the selection of highly impactful, innovative proposals, or perhaps is better at minimizing bias. Obviously the “funded” applications in the simplified and journal panels were not actually funded, so the output of these selected grants cannot be measured. Nevertheless the authors should go further to justify why only these two parameters were chosen and clarify any potential limitations this may have on their conclusions.

REVIEWER
Rüdiger Mutz
Professorship of Social Psychology and Research on Higher Education, ETH Zurich, Switzerland

REVIEW RETURNED
11-May-2015

GENERAL COMMENTS
In my view such prospective or observational studies to test the effect of different peer review procedures are very helpful and very interesting in peer review research. Therefore, the manuscript merits publication in BMJ Open, however, after a revision, which should address the following aspects:

- Conclusions: Apparently, the results are not quite as expected. The agreement between simplified and official panels did not exceed the threshold. However, some far-reaching decision were made that the NHMRC should simplify their peer review processes and shorten applications. This conclusions seems to be based more on reasons of efficiency than on the empirical results. Further, a sample size of 32 reviewers and 76 proposals might not justify such far reaching decisions. I recommend to formulate the conclusions more cautiously or modestly.

- Study design: It would be helpful for the readers, if some more details of the study design could be reported. There are overall 32 reviewers, who were assigned with equal proportions to the two
review procedures. In each procedure 72 proposals were reviewed (Basic Science, 36 Public Health). For me it is not quite clear, how the 16 reviewers were assigned to the 7-person Simplified panel? How many panels were generated (2 or 3)? Further, measurement dependencies must be assumed due to the fact that the same proposal was reviewed by different reviewers, and different reviewers review the same proposal or different reviewers were in the same panel. Measurement dependencies might have some impact on the statistical analyses. I think this circumstances should be mentioned in the revision and, possibly, excluded per assumption (e.g., “We do not consider any measurement dependency in our analyses”).

- Selection bias: In experimental designs the units are randomly assigned to the treatments, i.e., different peer review procedures and the official peer review procedure as reference group, respectively. This seems in the study at hand not to be the case. Some information about the groups is given (Table 2), which make it plausible that the two simplified peer review treatment groups are comparable at least in some covariates. There is, however, no information about the official peer review group. Therefore, I recommend to mention the problem of comparability of groups as a limitation of the study in the discussion.

- Measurement error: It is also clear that rating scales are not free from measurement error. Even if the same reviewers rate the same proposal again, one will have to take into account random fluctuations of the judgments beyond lack of inter rater agreement. Therefore, I recommend to mention the problem of measurement errors in the discussion as further limitation.

- Agreement coefficient: As agreement coefficient the percentage of agreement was used. If the reported confidence intervals were used for interpretation, the threshold of 75% would lay within the 95% confidence interval (i.e. repeated sampling with the same procedure). Therefore the result will be perhaps not as bad as it sounds, if only the absolute value is interpreted.

- Minor p. 9 (3rd section): The term “esoteric” in “esoteric criteria-based scoring system” is somewhat pejorative. Perhaps a less pejorative term can be found.

p. 6 and p. 8: The two titles “Simplified panels” on page 6 and “Simplified peer review” on page 9 is somewhat irritating, especially that the text following the title on page 9 is about the official panel and text before “Journal panels” is about the “simplified peer review”. The revision should better structure the methodological part (superordinate title and other titles).

Table 5: It should be mentioned in the tables that mean (or total sum) values per proposal were reported. Additionally, it would be helpful to shortly mention in the text, how the information for calculation of the peer review costs of 1109 Dollar was extracted from the table (e.g., =412+160+537).
The subject matter of this study is of great practical importance, particularly in an age in which many funding agencies are struggling to maintain reasonable success rates in the face of escalating costs and stagnant budgets. Any study that can contribute to a simplification and streamlining of the application process for agencies and researchers is both needed and helpful.

In the Methods section, there is a description of the simplified and journal panels, including Table 2 that summarizes participant characteristics. However, the information on how panels categorized grants is not given until after the description of both panels. It might be helpful to explain the scoring system used with the description of the review process.

There is no section in the Methods section that describes the official (traditional) review process/panel. For example, in the abstract it states that the official process uses panels of 12 people; however, there is no information on the size of the official panels in the Methods section.

Table 2 describing the characteristics of the simplified and journal panels should really be moved into the results section, as it reflects the results of the recruitment process. Would it be possible to add another column giving this information for the official review panels? Even if the three panels are too small to compare statistically, it would be interesting to see whether the convenience samples created for the simplified and journal panels reflect the characteristics of the official panel.

As it currently reads, there is no explanation of how the scores of the official panels were utilized to correspond with the three-category system used by the simplified and journal panels (definitely fund, possibly fund, or definitely do not fund). So in Table 3, is it safe to say that "Funded by NHMRC" means actually received funding, rather than viewed by the panel as fundable? There is a critical difference between the two. Unless NHMRC has a bottomless bucket, not all fundable projects (i.e., proposals with a score indicating at least good science) actually receive funds. In most funding agencies, proposals are ranked by score and then the number actually funded depends upon how far the money stretches. This can vary according to the proposed budgets, the approved budgets, the number of applications, and the size of the funding envelope for that year (as well, for some agencies, commitments from previous years).

As a result of these issues, comparing “funded” to “fundable” as in Table 3 is difficult to interpret. In fact, if you calculate a Cohen's kappa, between the simplified panel and the official process for both fields agreement is only slight (Cohen's k=.198). Whereas raw agreement suggests agreement is better for the basic science than the public health panels (69% vs. 75%), Cohen's kappa suggests agreement is slightly better for the basic science application (k=.2205, a fair effect) than public health proposals (k=.182, a slight effect).

For the comparison of the journal panel to the official panel, Cohen's kappas are negative as the observed concordance is less than expected by mean chance. It is possible, for example, that the poor results shown in Table 3 reflect the fact that some of the projects the simplified or journal panels believed were fundable fell below the NHMRC funding cut-off.

The comparison between the simplified and journal panels (as shown in Table 4) is probably more helpful as the outcome
measures were consistent (fundable vs. not fundable). However, the Cohen’s kappa for the basic science (Cohen’s k= .217), public health (k=.300) and combined (k=.256) areas are only fair.

In both Tables 3 and 4, using the percent raw agreement also gives a somewhat misleading picture of how well the various systems compare. If you look at the actual numbers, there is good agreement on proposals that should not be funded but poor agreement on those that should. This does not bode well for identifying which proposals should be – or will be – funded.

In summary, some revision is needed before this study is ready to publish. The authors need to 1) add a description of the official process, 2) reconsider or justify how to compare experimental panel to official process results, and 3) use more appropriate statistics than simply raw agreement. The information on cost of the various review processes is very interesting and informative and should be of interest to a large number of funding agencies.

**VERSION 1 – AUTHOR RESPONSE**

**Reviewer Name Guy D. Eslick**
**Institution and Country** The University of Sydney, Australia

This is an interesting and important study. Much time and effort is wasted in preparing Grant applications as part of a researchers career. Even more time and effort is wasted by national organisations reviewing such proposals. This process must be streamlined and simplified! This study goes some way to initiating this process and I commend the authors for producing a simple and viable option for Grant bodies to follow in the future. The cost savings are substantial!

Table 2, a column showing a p-value comparing the 'Simplified' vs 'Journal' panel would be of value, if only to show that there is no difference between the groups. This method would also need to be added to the 'statistical analysis' section of the METHODS.

RESPONSE: With 16 members per group we only have a 90% power to detect a difference between groups of 47%. It is therefore likely that all the p-values would be over 0.05, but that could be misinterpreted by some readers as the groups being the same. We prefer to keep them as simply descriptive statistics.

I must say I like the idea of the 'Journal panel', nice and easy, but how would applications be ranked/selected using a two-person approach for a large number of applications (n~4,000)? Any suggestions?

RESPONSE: It is highly likely that with just two reviewers per proposal and many applications there would be tied scores. We note that these would not occur across 4000 applications as applications would still be stratified by field. In this situation we could use either a third journal review or lottery to split these ties. We have added this to the discussion.

Some comparison with the NIH (USA) system would be nice in the Discussion as I believe they have reduced the size of the Grant proposals recently.

RESPONSE: We are not aware of any major recent changes to the NIH system in relation to making applications simpler; there was some reform of the peer review process but that does not save time for applicants. In 2012 the Canadian Institutes of Health Research made major changes to their system to make it less time-consuming for researchers. However, we are not aware of any research or other publication that reviews the impact of these changes.

**Reviewer Name Stephen Gallo**
**Institution and Country** American Institute of Biological Sciences

The authors have conducted a prospective study to look at levels of agreement between three different grant review processes: the standard onsite NHMRC review panel, the simplified and smaller onsite review panel process, and a journal-style review (2 independent reviewers). There are three panels in total, all reviewing the same set of applications. The two experimental panels had shorter
applications, fewer reviewers (with and without discussion) and greatly reduced costs (in both travel
time and in reviewer hours). The results are primarily the percentage levels of inter-panel agreement
in funding status of proposals and the costs associated with each review setting.

Optimizing the funding processes of scientific research is crucial to the success of the scientific
enterprise. While peer review is the most commonly used decision-making process to determine
funding, it is poorly studied. In addition, most of the data in the literature concerning peer review is
retrospective; there is a glut of prospective studies. Therefore, the authors are to be congratulated on
their efforts to conduct such studies.

However, there are a few concerns that should be addressed in the manuscript.

Firstly, the funding rates for Journal style review vs the NHMRC are very different. The journal panel
only found 8.3% of proposals worthy of funding while NHMRC and Simplified panels would find 20.8%
and 23.6% of proposals worthy of funding, respectively. This indicates that the journal panel was
much harsher in its evaluation of proposals. It is unclear if this is just this particular group of
reviewers, or a function of the review setting itself. A relatively recent analysis of grant proposal
scoring data from the US National Institutes of Health has suggested that score shifts as a result of
panel discussion have practical importance over the funding status of 13% of applications (Martin MR
et al, PLoS One. 2010; 5(11): e13526). Thus, it may be that lack of discussion may play a role in the
differences seen between Journal and NHMRC panels. In any case, it certainly seems that the
statement:

“our results indicate that the current lengthy process could be simplified without impacting greatly on
funding outcomes”

is probably not completely correct for the Journal-review setting, as it seems there is a substantial
impact on funding outcomes when switching to the Journal-style review panels. This issue should be
discussed in the text.

RESPONSE: We have qualified the relevant “Strengths and limitations” bullet point. We have also
discussed the lower acceptance rate for the journal panel and included the useful reference.

Secondly, of the proposals that the NHMRC panel did decide to fund, only 40% were in agreement
with the Simplified panel’s decisions and only 7% were in agreement with the Journal panel’s
decisions. Yet in contrast, 64% of the proposals that the NHMRC panel did NOT decide to fund were
in agreement with the Simplified panel’s decisions and 72% were in agreement with the Journal
panel’s decisions. The authors state:

“The agreement about which proposals to fund was lower than the agreement about which proposals
not to fund. This is partly because many more proposals were not funded than funded.”

While this is surely the case, it does call into question how reviewers make decisions for high versus
low quality proposals. It may be very likely that reviewers are better at separating well-constructed
proposals from those that are flawed than they are at discriminating between good and great
proposals. It may be that decisions based in catching logical flaws and methodological errors may
have higher inter-panel reliability than those which are based predominantly in the subjective
prediction of success. As the authors point out, others have observed the heterogeneity in reliability
across proposal quality. For instance, in the prospective inter-panel reliability study of Fogelhom (J
clin epi 65.1 (2012): 47-52.), which the authors quote in this manuscript, it also found that reliability is
“score” dependent:

“The results suggest that reliability is better at the low end of the scale, that is, the consistency in
identifying good proposals is lower than for average and poor proposals. The distribution of the scores
(success rate) does not affect the sensitivity and specificity. Therefore, our findings suggest that the
criteria (or consensus) for poor proposals are clearer than the criteria for good proposals.”

RESPONSE: We agree with this interpretation. It may well be easier to consistently find flaws than
discriminate between degrees of excellence. We have now included a discussion of this issue.

We conducted interviews with all our panel members after the peer review process and some of them
spoke about this issue. We hope to soon publish the results of a qualitative analysis of these
interviews.

Given the choice we would rather have a system that reliably rejected the poor grants and struggled
with good grants as any “mistakes” amongst the good grants have less consequence.
In any case, before a strong case can be made that the experimental review mechanisms are roughly
equivalent to the NHMRC with regard to funding outcomes, the authors should further justify why the
bulk agreement parameter is used to assess the different review panels. Also, the difference in
agreement with the list of fundable NHMRC proposals for Simplified and Journal (40% vs 7%,
respectively) should be discussed. It may be that acceptable agreement thresholds for both
unfundable and fundable populations may be needed.
RESPONSE: We use overall agreement as a simple statistic that is meaningful and is likely to be
acted on by the key decision makers. The cross-tabulations are also an important result as they show
every possibility and readers can use the tables to draw their own summaries (as three reviewers
have done).
Finally, the authors only explore the bulk level of inter-panel agreement and cost as parameters to
justify the adoption of one peer review setting over another. However, depending on the stakeholder
point of view, there may be other parameters that could also be viewed as important to optimize. For
instance, it is unclear if review setting affects the promotion of innovation. It may be that one review
setting is superior to the others in fostering the selection of highly impactful, innovative proposals, or
perhaps is better at minimizing bias. Obviously the “funded” applications in the simplified and journal
panels were not actually funded, so the output of these selected grants cannot be measured.
Nevertheless the authors should go further to justify why only these two parameters were chosen and
clarify any potential limitations this may have on their conclusions.
RESPONSE: This is a good point, and there may be systematic differences between the approaches
rather than just noise. A funding agency would want to look at these issues carefully before making
any change to their peer review process. We might expect that the journal panel that did not use track
record was more likely to fund ideas from less experienced researchers, and there is evidence from
some fields that younger researchers have more innovative ideas. This issue is beyond the scope of
this study, but we could imagine a prospective study where applications were scored using two
schemes, randomly funded according to one scheme or the other, and then compared at a later date
in terms of their scientific output.
Another potential motivation for changing to a simpler system is that any money saved (from needing
less staff to run the system and less travel and accommodation costs) could be added to the funding
pool. This gives the simpler system a buffer, so if it does make slightly worse decisions these may be
more than compensated for by the larger number of applications that can be funded.
We have added a caveat about funding agencies generating their own data to the last line of the
paper.
Reviewer Name Rüdiger Mutz
Institution and Country Professorship of Social Psychology and Research on Higher Education, ETH
Zurich, Switzerland
In my view such prospective or observational studies to test the effect of different peer review
procedures are very helpful and very interesting in peer review research. Therefore, the manuscript
merits publication in BMJ Open, however, after a revision, which should address the following
aspects:
- Conclusions: Apparently, the results are not quite as expected. The agreement between simplified
and official panels did not exceed the threshold. However, some far-reaching decision were made that
the NHMRC should simplify their peer review processes and shorten applications. This conclusions
seems to be based more on reasons of efficiency than on the empirical results. Further, a sample size
of 32 reviewers and 76 proposals might not justify such far reaching decisions. I recommend to
formulate the conclusions more cautiously or modestly.
RESPONSE: We agree that the recommendations were based more on efficiency than strong
evidence of no change. We have toned down our language, including the last sentence of the abstract
and the “Strengths and limitations” bullet point.
- Study design: It would be helpful for the readers, if some more details of the study design could be
reported. There are overall 32 reviewers, who were assigned with equal proportions to the two review
procedures. In each procedure 72 proposals were reviewed (Basic Science, 36 Public Health). For me it is not quite clear, how the 16 reviewers were assigned to the 7-person Simplified panel? How many panels were generated (2 or 3)? Further, measurement dependencies must be assumed due to the fact that the same proposal was reviewed by different reviewers, and different reviewers review the same proposal or different reviewers were in the same panel. Measurement dependencies might have some impact on the statistical analyses. I think this circumstances should be mentioned in the revision and, possibly, excluded per assumption (e.g., “We do not consider any measurement dependency in our analyses”).

RESPONSE: The two panels were assembled based on expertise. We agree that there will be correlations within scores from the same reviewer and between reviewers on the same panel. The NHMRC system ignores these dependencies and aggregates scores using a simple mean, and we therefore did the same. We have added a more detailed description of the NHMRC system.

- Selection bias: In experimental designs the units are randomly assigned to the treatments, i.e., different peer review procedures and the official peer review procedure as reference group, respectively. This seems in the study at hand not to be the case. Some information about the groups is given (Table 2), which make it plausible that the two simplified peer review treatment groups are comparable at least in some covariates. There is, however, no information about the official peer review group. Therefore, I recommend to mention the problem of comparability of groups as a limitation of the study in the discussion.

RESPONSE: We have used publicly available data to compare the characteristics of our panels to the official NHMRC panel members (Table 2) and discuss the differences.

- Measurement error: It is also clear that rating scales are not free from measurement error. Even if the same reviewers rate the same proposal again, one will have to take into account random fluctuations of the judgments beyond lack of inter rater agreement. Therefore, I recommend to mention the problem of measurement errors in the discussion as further limitation.

RESPONSE: We agree that measurement error is an issue, and this is contained within the overall difference between the panels. We have now explicitly mentioned measurement error in the discussion.

- Agreement coefficient: As agreement coefficient the percentage of agreement was used. If the reported confidence intervals were used for interpretation, the threshold of 75% would lay within the 95% confidence interval (i.e. repeated sampling with the same procedure). Therefore the result will be perhaps not as bad as it sounds, if only the absolute value is interpreted.

RESPONSE: Yes, the range of agreement does comfortably contain the 75% threshold. However, given the recommendation to use more cautious language we prefer not to make any claims about this.

- Minor

p. 9 (3rd section): The term “esoteric” in “esoteric criteria-based scoring system” is somewhat pejorative. Perhaps a less pejorative term can be found.

RESPONSE: Agreed and changed to “more complex”.

p. 6 and p. 8: The two titles “Simplified panels” on page 6 and “Simplified peer review” on page 9 is somewhat irritating, especially that the text following the title on page 9 is about the official panel and text before “Journal panels” is about the “simplified peer review”. The revision should better structure the methodological part (superordinate title and other titles).

RESPONSE: We have changed the “simplified peer review” subheading to “simplified scoring” and have added a separate “Official panels” sub-section.

Table 5: It should be mentioned in the tables that mean (or total sum) values per proposal were reported. Additionally, it would be helpful to shortly mention in the text, how the information for calculation of the peer review costs of 1109 Dollar was extracted from the table (e.g.. =412+160+537).

RESPONSE: We have added a total column to Table 5.

Reviewer Name Corinne Hodgson
Institution and Country Cancer Care Ontario, Canada
The subject matter of this study is of great practical importance, particularly in an age in which many
funding agencies are struggling to maintain reasonable success rates in the face of escalating costs and stagnant budgets. Any study that can contribute to a simplification and streamlining of the application process for agencies and researchers is both needed and helpful.

In the Methods section, there is a description of the simplified and journal panels, including Table 2 that summarizes participant characteristics. However, the information on how panels categorized grants is not given until after the description of both panels. It might be helpful to explain the scoring system used with the description of the review process.

RESPONSE: There is a lot of information to convey concerning the structure of the panels and the decisions they make. We prefer to have a separate section for scoring as this is a separate issue from the panel’s make-up.

There is no section in the Methods section that describes the official (traditional) review process/panel. For example, in the abstract it states that the official process uses panels of 12 people: however, there is no information on the size of the official panels in the Methods section.

RESPONSE: We have added a separate “Official panels” subsection and have added the size of the official panels.

Table 2 describing the characteristics of the simplified and journal panels should really be moved into the results section, as it reflects the results of the recruitment process. Would it be possible to add another column giving this information for the official review panels? Even if the three panels are too small to compare statistically, it would be interesting to see whether the convenience samples created for the simplified and journal panels reflect the characteristics of the official panel.

RESPONSE: As suggested we have added a column describing the official panels based on what information was publicly available and have moved the table to the start of the Results section. We agree the two panels are too small to compare statistically and prefer to keep the comparisons as descriptive. Our panels were similar in terms of academic level, but had more women and more people from Go8 universities compared with the official NHMRC process. We are not concerned about these differences.

As it currently reads, there is no explanation of how the scores of the official panels were utilized to correspond with the three-category system used by the simplified and journal panels (definitely fund, possibly fund, or definitely do not fund). So in Table 3, is it safe to say that “Funded by NHMRC” means actually received funding, rather than viewed by the panel as fundable? There is a critical difference between the two. Unless NHMRC has a bottomless bucket, not all fundable projects (i.e., proposals with a score indicating at least good science) actually receive funds. In most funding agencies, proposals are ranked by score and then the number actually funded depends upon how far the money stretches. This can vary according to the proposed budgets, the approved budgets, the number of applications, and the size of the funding envelope for that year (as well, for some agencies, commitments from previous years).

RESPONSE: Yes, the NHMRC results were based on actual funding. The ranking process described by the reviewer is how the NHMRC works. We have added a new subsection to the Methods giving more detail on the official process.

As a result of these issues, comparing “funded” to “fundable” as in Table 3 is difficult to interpret. In fact, if you calculate a Cohen’s kappa, between the simplified panel and the official process for both fields agreement is only slight (Cohen’s k=.198). Whereas raw agreement suggests agreement is better for the basic science than the public health panels (69% vs. 75%), Cohen’s kappa suggests agreement is slightly better for the basic science application (k=.2205, a fair effect) than public health proposals (k=.182, a slight effect).

For the comparison of the journal panel to the official panel, Cohen’s kappas are negative as the observed concordance is less than expected by mean chance. It is possible, for example, that the poor results shown in Table 3 reflect the fact that some of the projects the simplified or journal panels believed were fundable fell below the MHHRC funding cut-off.

The comparison between the simplified and journal panels (as shown in Table 4) is probably more helpful as the outcome measures were consistent (fundable vs. not fundable). However, the Cohen’s kappa for the basic science (Cohen’s k=.217), public health (k=.300) and combined (k=.256) areas
are only fair.

**RESPONSE:** Cohen’s kappa has a serious flaw as it corrects for chance agreement higher than 50%, which is impossible, this means such kappa statistics are rarely above 0.8. Better chance-adjusted agreement statistics are available, such as that proposed by Gwet (Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 2008. 61(1): 29-48).

In both Tables 3 and 4, using the percent raw agreement also gives a somewhat misleading picture of how well the various systems compare. If you look at the actual numbers, there is good agreement on proposals that should not be funded but poor agreement on those that should. This does not bode well for identifying which proposals should be – or will be – funded.

**RESPONSE:** We agree and have included some discussion of this issue. See also our response on this issue to reviewer Stephen Gallo.

In summary, some revision is needed before this study is ready to publish. The authors need to 1) add a description of the official process, 2) reconsider or justify how to compare experimental panel to official process results, and 3) use more appropriate statistics than simply raw agreement. The information on cost of the various review processes is very interesting and informative and should be of interest to a large number of funding agencies.

**RESPONSE:** We believe that the cross-tabulations combined with raw agreement are the most appropriate statistics. The cross-tabulations tables allow people to see the complete data and calculate their own summaries (as three reviewers did). Chance is a recognised part of the funding process and we do not want to adjust for it as it will never be removed in practice. We prefer to show the raw agreement and let researchers and funding agencies decide whether this figure is acceptable.

---

**VERSION 2 – REVIEW**

| REVIEWER          | Guy D. Eslick  
The University of Sydney, Australia. |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>REVIEW RETURNED</td>
<td>02-Jun-2015</td>
</tr>
</tbody>
</table>

**GENERAL COMMENTS**

All changes made, good manuscript.

| REVIEWER          | Stephen A Gallo  
American Institute of Biological Sciences, USA |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>REVIEW RETURNED</td>
<td>29-May-2015</td>
</tr>
</tbody>
</table>

**GENERAL COMMENTS**

The authors have addressed this reviewer’s concerns.

| REVIEWER          | Corinne Hodgson  
Cancer Care Ontario  
Canada |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>REVIEW RETURNED</td>
<td>08-Jun-2015</td>
</tr>
</tbody>
</table>

**GENERAL COMMENTS**

The authors have been conscientious in addressing the concerns of the reviews except the changes appear to be more superficial than substantive. They have not reconsidered their fundamental thesis or addressed underlying problems in their study design/analysis. The issue of whether you can compare theoretically “fundable” proposals to those actually “funded” is not adequately addressed. This is an important, as few funding agencies have the resources to fund all fundable applications. This may explain the rather poor agreement between the traditional system and the simplified and journal panels.
The authors stress that raw agreement was close to 75% but almost all of the agreement occurs in the identification of non-fundable/non-funded projects. Agreement on what constitutes good research was actually very poor: 8.3% for the simplified panel and 1.9% for the journal panel. Even when comparing similar outcomes ("fundable") agreement between the two experimental methods on what was fundable was low: 5.6%. In contrast, if you look at Hodgson's study, total raw agreement on "fundable" projects (as based on a score that would make a proposal fundable if adequate funds were available) was 72.6% but it was made up of 42.3% agreement on those not fundable and 30.2% on those fundable. One reviewer requested the use of a Cohen's kappa and the argument was made that it was inappropriate as agreement above .80 is rare. Yes, A Cohen's kappa of .80 or over is considered "almost perfect" agreement and thus is relatively rare but this does not mean the test is useless. Cut-offs exist for good, fair and poor agreement. One would think that a new system should have at least good agreement with the "gold standard" if it is to be seriously taken as a potential alternative.

There are both pros and cons to publishing this study. The pro is that it would put this information in the public domain and possibly act as a stimulus for discussion among researchers and funding agencies/administrators. This could be highly beneficial. The con is that results could be taken out of context and used as "proof" that a streamlined process is feasible and reliable. That would be a shame and could lead to precipitous changes in an unwieldy but time-tested grant review process.
Using simplified peer review processes to fund research: a prospective study

Danielle L Herbert, Nicholas Graves, Philip Clarke and Adrian G Barnett

BMJ Open 2015 5:
doi: 10.1136/bmjopen-2015-008380

Updated information and services can be found at:
http://bmjopen.bmj.com/content/5/7/e008380

These include:

References

This article cites 15 articles, 5 of which you can access for free at:
http://bmjopen.bmj.com/content/5/7/e008380#BIBL

Open Access

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections

Articles on similar topics can be found in the following collections
Evidence based practice (363)
Medical publishing and peer review (40)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/