NZIER
NEW ZEALAND INSTITUTE OF ECONOMIC RESEARCH

The Policy Project
Responsive today, shaping tomorrow

Policy Quality Framework

Independent Panel Review of Papers

Report to the Department of the Prime Minister and Cabinet

October 2023

Authorship

This Report was prepared by Cathy Scott from NZIER – with input from the Panel set up by DPMC to review papers – Julie Keenan, Ministry of Education; Howard Fancy, Howard Fancy and Associates; and Helen Lockyer, DPMC.

The assistance of Sarah Spring at NZIER is gratefully acknowledged.

Registered office: Level 13, Public Trust Tower, 22–28 Willeston St | PO Box 3479, Wellington 6140
Auckland office: Ground Floor, 70 Shortland St, Auckland
Tel 0800 220 090 or +64 4 472 1880 | econ@nzier.org.nz | www.nzier.org.nz

Key points

This exercise aimed to investigate whether or not the Policy Quality Framework (PQF) was being applied consistently. And if not, provide some suggestions on what improvements could be made to the guide on panels and processes for assessing policy advice papers. Any areas identified as potential problems can be discussed in more detail with agency panels. This has arisen from a concern, following the interim evaluation of the Policy Project, that the PQF may be applied differently by different agencies, and that its application might change over time.

The Department of the Prime Minister and Cabinet set up a panel to assess 52 papers from 26 agencies that had been assessed as part of the 2021/22 policy advice reviews.[1] Half of these were scored a 3, and half were scored a 4 by agencies.

Key findings from this analysis are:

- Papers were largely fit for purpose and would allow Ministers to make informed decisions. They covered a wide range of types and issues – some of which were highly technical and quite tricky.
- On average, the Panel scored the papers lower than the original score across the whole sample.
- This difference was significant in the papers originally scoring a 4 – the Panel scored them on average half a mark lower.
- However, there was a good deal of consistency in the scores of papers which originally scored a 3. 44% were scored a 3 by the Panel, and 88% were scored within half a mark of a 3. The average score from the Panel was 3.06 (compared to the average original score of 3.00).
- The Panel did not have access to any contextual information about the papers other than what was in the paper itself. This may have impacted on scores.
- The Panel found that there were a number of key elements of the PQF often missing in papers, which accounted for the lower scores the Panel gave. In particular, papers often lacked:
  - A Treaty of Waitangi analysis, a te ao Māori perspective or an assessment of the implications for Māori.
  - A consideration of implementation requirements, issues and risks.
  - Any reflection of diverse views and perspectives, the views of stakeholders, or other agencies.
  - Options assessments – when options could have been offered and assessed.
  - Clarity early in the paper about what the Minister/s needed to do, e.g. make a decision, discuss with other Ministers, or note the content. The next steps in the process for the Minister also weren't as clear as they could be at times. Papers needed to be more centred on the needs of the Minister.
- There was slightly more consistency in the scoring of standard policy papers, i.e. those seeking decisions from a Minister, compared to noting papers, operational papers or those discussing process matters.

[1] Around half of these were scored by agency panels and half by NZIER as part of the review process.

The Panel could not comment on consistency between original scores due to the sample size (two papers per agency) from each agency.

We concluded that there were some differences in sampling and how the scoring was done – particularly for higher-scoring papers.

We should note that, in the main, the feedback written up on each paper designed to go to authors and managers was reasonable, cognisant of the PQF criteria, recognised good practice, and offered practical suggestions for improvement. Good feedback is an important tool in improving the quality of policy advice.

The plan is now to investigate these issues further in discussion with agency panels. This will help to make recommendations on possible changes to the Guidance for panels and assessing papers to improve consistency. Guidance or further background on aspects of the PQF standards to be used when assessing papers (and developing them) can also improve consistency and overall quality.

Contents

Appendices

Figures

Tables

1 Introduction

The purpose of the Policy Quality Framework is twofold

Firstly, for annual performance reporting – essentially an accountability purpose.

And secondly, to measure performance in order to manage and improve the quality of policy advice – part of which is to identify strengths and innovations and make them more widespread and to identify and address areas for improvement.

This assessment is considered in light of these two factors.

This exercise was spurred by the findings of an interim evaluation of the Policy Project

The Policy Quality Framework (PQF) has been mandated for use by Government agencies in assessing and reporting on the quality of policy advice since 2020.

An interim evaluation of the Policy Project was undertaken by Allen + Clarke (published September 2021).[2] This evaluation highlighted concerns about the confidence that agencies have in the policy quality review results, particularly the consistency of scoring policy papers. Recommendation 7 of the interim evaluation recommends that the Policy Project: *'investigate the annual quality assessment, so that agencies understand how these assessments operate, have confidence in the results, and use it to drive performance improvement.'*

A three-phase approach was established – this exercise was focused on the part of the first phase of that approach

Considering the feedback of the Policy Capability Leads' Group and discussions with the Deputy Chief Executive (Policy) DPMC and NZIER, the Policy Project proposed a three-phase strategy for addressing concerns about consistency and confidence in agency policy scoring. This is set out in Appendix A.

The outputs expected from this part of the work are to:

- Confirm whether or not there is an issue with the consistency of scoring across agencies (and perhaps across time, types of paper, etc.)
- Identify any changes required to the approach and Guidance for panels[3] needed to help address these issues.

This will also feed into discussions with Panels and the proposed webinar for Panels.

[2] www.dpmc.govt.nz/publications/overview-of-interim-evaluation-of-policy-project

[3] www.dpmc.govt.nz/publications/using-policy-quality-framework-assess-papers

2 Methodology

2.1 Paper sampling

Each agency was asked to select two papers from their 2021/22 review sample – one randomly selected from the papers that scored a 3 and one randomly selected from the papers that scored a 4.[4, 5] This ensured that we were comparing like with like; the size of the sample was manageable, and the papers could easily be drawn from agencies' samples.

The Panel received 52[6] papers from 26 of the 28 agencies with policy appropriations by the time the review commenced.

2.2 Panel process

The Panel then reviewed these blind (i.e. without knowing the score) using the Policy Quality Framework.

The Panel was made up of people with considerable experience in applying the Policy Quality Framework. It included:

- Cathy Scott, NZIER (chair) – involved in the reviews undertaken by NZIER, as well as membership on a number of other agency review panels.
- Helen Lockyer, Policy Project, DPMC – involved in developing the PQF, and in a number of panels in DPMC and other agencies.
- Julie Keenan, Ministry of Education, Chair of the Ministry's Policy Quality Panel
- Howard Fancy, Howard Fancy and Associates, Chair of the Ministry for the Environment's Policy Quality Panel, and involved in a range of other agency panels.

The original scores and the Panel review were then compared – alongside the comments made by both the Panel and the agency/NZIER reviews.

A spreadsheet has been provided to DPMC to accompany this report. It looks at the following:

- raw scores from the Panel
- differences in scores between the original score and the Panel score
- differences in scores for papers originally scoring a 3, and those originally scoring a 4 and Panel scores
- differences in scores by paper type
- differences in scores by the relative level of complexities in papers.

It was not possible to compare and contrast scoring approaches across different agencies with only a sample of two papers per agency – any conclusions would not be statistically significant.

This report presents summary data. It also includes an analysis of the factors behind differences between scoring – based on what the Panel saw in the sample.

[4] Not all agencies did this. Some agencies didn't have papers that fitted in these categories (particularly 4s – so provided an alternative paper).

[5] One agency used a different scoring regime – papers had "intermediate" marks – rather than sticking with full and half marks, e.g. 3.6 and 3.8. In these cases, scores were rounded to the nearest 3 or 4 to categorise them for this analysis.

[6] Around half were reviewed by agency panels, and the other half reviewed by NZIER.

3 Results

3.1 Overall, the Panel scored papers lower than the original score

The average original score across the sample was 3.51, and for the Panel, 3.14. The range of the Panel scores was between 2 and 4.[7]

Table 1: Panel scores

On average, the Panel scored papers lower than the original score.

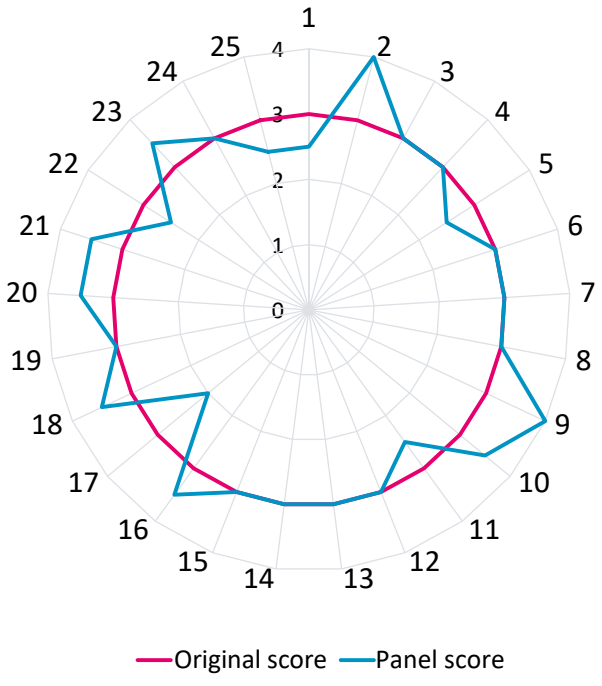
Papers	Number	Average Panel score
Original score 3	25	3.06
Original score 4	27	3.22
Whole sample	**52**	**3.14**

Source: NZIER

3.1.1 There was considerably greater consistency in papers scoring a 3

For papers originally scored a 3, there was some variation in scores between those and the Panel scores. But the average score was largely consistent. This is illustrated below.

Figure 1: Papers which originally scored a 3 compared to Panel scores



Source: NZIER

44% of the papers that were originally scored a 3, the Panel also scored a 3.

88% of the papers that originally scored a 3, the Panel scored within half a mark of a 3 (i.e. 2.5, 3 or 3.5).

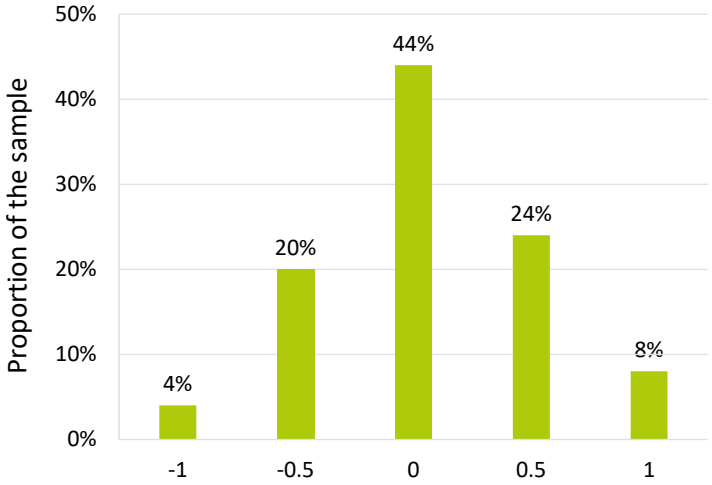
[7] The trends were broadly similar for the subset of papers reviewed by NZIER. However, there was even greater consistency with papers scoring a 3, and slightly improved consistency with papers scoring a 4, when compared to papers scored by agencies (and the Panel).

For 32% of the papers, the Panel scored higher than the original score.

Overall, the Panel scored these papers relatively consistently with agencies.

Figure 2: Difference in scores between the original and the Panel

Most papers scored within half a mark of the original score.

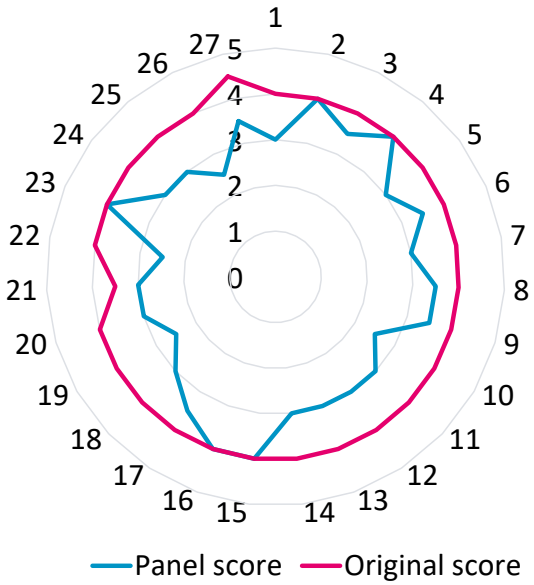


Source: NZIER

3.1.2 There was more variation in papers which originally scored a 4

The variation in scoring between the Panel and the original score was greater in papers that originally scored a 4.[8] The Panel scored these papers lower than the original score – on average, just over half a mark lower. The Panel's average score for these papers was 3.22. The original score was 4.00.

Figure 3: Papers originally scoring a 4 compared to Panel scores



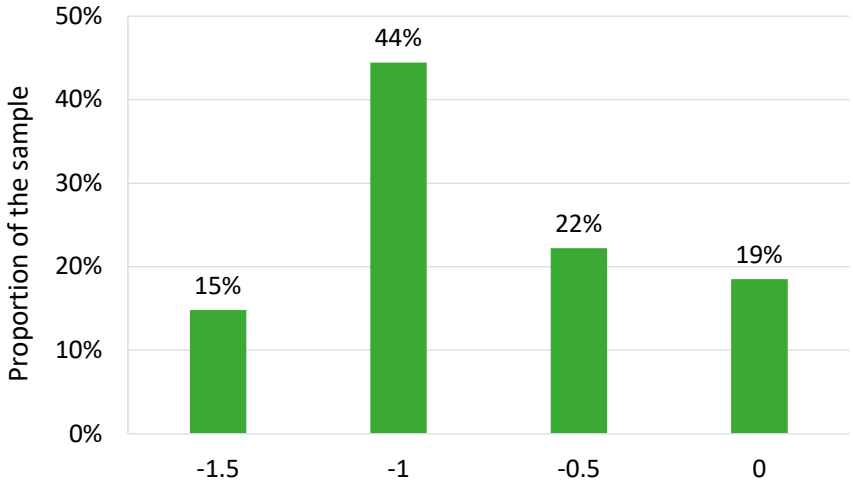
Source: NZIER

None scored more than a 4, only 19% scored a 4, but 41% were within half a mark of a 4 (i.e. scored a 4 or a 3.5).

[8] As noted previously there were some minor variations in original scores – papers counted in this category due to rounding included one papers scoring a 3.5, one a 3.6, one a 3.8 and one a 4.5.

Figure 4: Difference between the original score and the Panel score

Scores were generally lower and had greater variation.



Source: NZIER

3.1.3 Standard policy papers scored slightly more consistently

There was considerable diversity in the types of papers in the Panel's sample – as might be expected given the Guidelines.[9]

Half of the papers reviewed were policy briefings which sought decisions from Ministers. 65% scored within half a mark of the original scores (whether or not they were originally 3s or 4s). This is consistent with the sample as a whole.

A quarter of the papers were Cabinet papers (including Cabinet LEG papers) with or without cover notes. These typically scored lower than the original score – on average by half a mark. However, the overall proportion of papers within half a mark was similar to the sample overall.

The number of other types of papers was too small to assess individually. But, fewer of these scored within half a mark of the original score, i.e. there was a greater spread of scores.

In section 3.2.4 of this report, we discuss the factors that seem to lie behind the differences in scoring consistency for different types of papers.

Table 2: Score differences by paper types

Standard policy papers had a more consistent score.

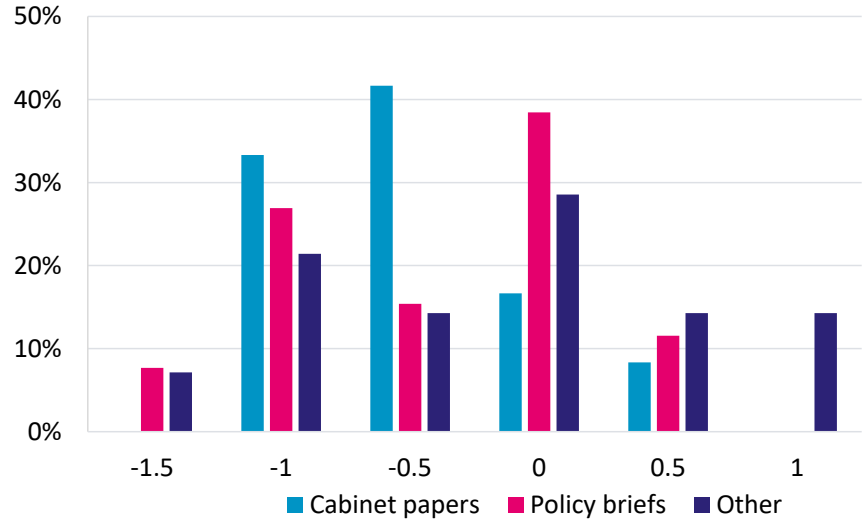
Paper type	Number	Average Panel score	Average original score	Within half a mark
Policy briefs	26	3.19	3.60	65%
Cabinet papers	12	3.00	3.48	67%
Other types[10]	14	3.18	3.36	57%
Whole sample	**52**	**3.14**	**3.51**	**64%**

Source: NZIER

9 Page 6 of www.dpmc.govt.nz/sites/default/files/2021-08/policy-quality-framework-guide.pdf.

10 Includes aides-mémoire, operational papers, event briefings, meeting briefs, update reports and covering notes over substantive documents for decision.

Figure 5: Difference in scores between the original and the Panel by paper types



Source: NZIER

3.1.4 The complexity of papers makes some difference in scoring

There was a mix of papers ranging from very simple to complex policy pieces.[11]

For the most complex papers – the average Panel score was nearly half a mark lower than the original score. But the spread of score differences was narrower.

For papers of medium complexity (which was most of the sample,) there was a similar difference in average scores, but there was a wider spread in the differences between scores.

And for the simpler papers, there was greater consistency in scoring.

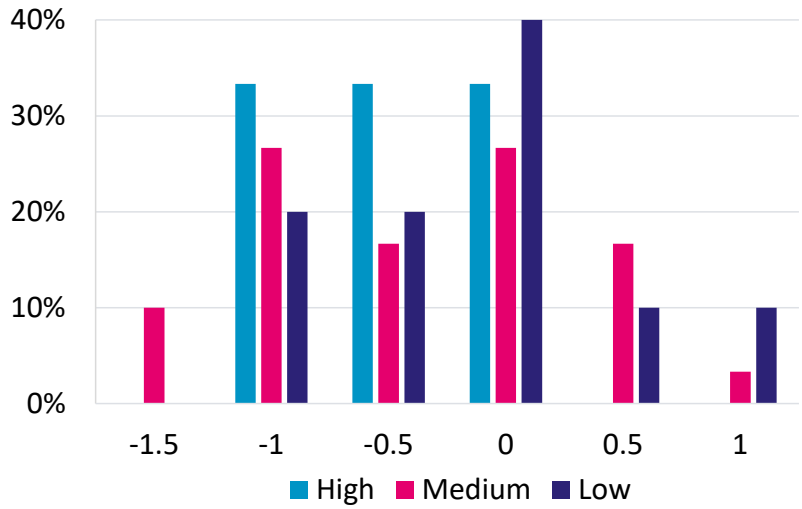
Table 3: Score differences by the complexity of the topic

Paper type	Number	Average Panel score	Average original score	Within half a mark
High	12	3.22	3.62	66%
Medium	30	3.18	3.24	60%
Low	10	3.05	3.13	70%
Whole sample	**52**	**3.14**	**3.51**	**64%**

Source: NZIER

[11] As part of the analysis the papers were categorised into different levels of complexity. This was only an approximate classification. Low = papers on simple operational, process or policy issues. These generally focused on a single issue. High complexity papers were multi-faceted policy issues involving detailed analysis, different perspectives, and a wide range of implications. Medium complexity papers fell between these two extremes.

Figure 6: Difference in scores between the original and the Panel by paper complexity



Source: NZIER

3.2 Factors behind the differences in scores

From what the Panel could see from the papers, the scores, and the feedback on the papers, there were several factors behind the differences in scores.

3.2.1 Is there a reluctance to give less than a 3?

From our experience, we know it can be hard to give a score less than a 3, i.e. that a paper didn't meet the standards, as feedback needs to be carefully managed. It may also impact an agency's reputation and credibility. We wondered about the extent to which this bias has played out.

In addition, despite the Guidance,[12] there may be more willingness to consider contextual factors, e.g. time and resource constraints, or constraints put on the analysis, if they are known.

Literature on scoring systems indicates that there is often a centrality preference. Also, a leniency bias might be in play, i.e. a tendency to rate papers as "met the standards" even when there are a number of shortcomings.

Ten (or 19%) papers the Panel reviewed scored less than a 3.

However, we note that we wouldn't expect ratings on anything like a normal curve – as many processes are in place to ensure policy papers are fit for purpose or better.

This is something that can be discussed with agency Panels and NZIER to get more information.

Section 3.3 of the report outlines the major areas for improvement the Panel saw across the sample of papers. These contributed to low scores.

3.2.2 There may be a different interpretation of what is a 4 or higher

Scoring can be interpreted using a deficit model or a strengths model.

There are different ways the definitions can be interpreted.

- A **deficit model,** i.e. identifying aspects of a paper which don't meet the standards and marking it down accordingly.
- A **strengths model** looks for elements of best practice and rates papers showing a number of these elements much more highly (despite elements which could be improved).

[12] Page 9 of the PQF Guidance for Panels.

As worded, the rating scale currently includes elements of each. This might explain some of the higher scores given to papers, which still had some areas for improvement – but also showed some areas of best practice.

This is something that could be tested further with agency Panels.

3.2.3 It's not straightforward scoring papers which have done well in one area and not in another

We saw a number of papers which had done well in one area of the PQF and poorly in another. At its most stark, there were cases where the analysis was comprehensive, but the paper wasn't well written and contained a range of typos and other simple errors, so it was hard to understand the analysis and its conclusions. We took the approach of balancing out scores – unless it was so poor in one part that it let down the paper as a whole.

Perhaps a little more guidance is needed on how this issue is best managed. It can also be tested further with agency panels to see whether approaches differ.

3.2.4 There seem to be differences in the types of paper that agencies include in their samples

There was a wide range of types of paper included in our sample. This is in keeping with the Guidelines.[13] However, we know that there are differences in approaches taken by agencies. Some agency samples focus primarily on traditional policy papers, e.g. ministerial briefings requiring policy decisions. This accounted for half of this sample. The PQF framework is specifically geared to this sort of paper.

Assessing different sorts of papers is more complex

Assessing some of the other types of papers requires a consideration of which elements of the PQF aren't relevant. This may impact scoring – it has in this sample. In general terms, there seem to be two possible approaches. Firstly, excluding elements from the assessment means that papers score on a narrower range of criteria, potentially boosting the overall score. On the other hand, the inability to demonstrate good practice over a wider range of elements may mean that scores remain moderate. Our sample had a wider range of score differences with these types of papers. This could be discussed with agencies to understand better the approaches used, and further guidance could be provided.

Papers were included that weren't strictly policy papers

There were a small number of papers in this sample, which were more of the nature of operational/procedural papers. They don't fit the criteria. Some further clarification in the Guidelines might be needed.

How to deal with Cabinet papers

Cabinet papers consistently scored half a mark lower by the Panel than were scored by agencies.

There has been some debate about including Cabinet papers in the sample. Cabinet papers must be written, so the Minister is satisfied with their content, but the issues, analysis and recommendations are clear to other Ministers with less detailed knowledge of the matter.

They are core policy products (and a major means of providing advice and the means of making decisions) and a central focus of many agencies' work. On the other hand, they are the Minister's papers, and in some cases, the Minister(s) may considerably influence their content. This may include adding material to cover specific political issues, narrowing down the range of options

[13] Page 6 of www.dpmc.govt.nz/sites/default/files/2021-08/policy-quality-framework-guide.pdf

presented, or reducing the discussion on risks. This is not always the case, though (of course, some of these matters are covered in regulatory impact assessments where provided).

Cabinet papers, along with Regulatory Impact Assessments, are one of the policy products that typically get external feedback as part of the preparation process – at times from other agencies, but also from PAG [14] Advisors or Officials' Committees.

There is also the special case of Cabinet Legislation Committee papers. These are very formulaic and often do not provide policy advice, merely a summary of the decisions already taken – alongside assurance that they have been reflected in drafting. But they are essential to effectively implementing policy and another core policy product.

Another factor to consider is whether the covering notes (often including policy, procedural and tactical advice) should be assessed together with the Cabinet papers, done as two separate (but related) papers, or just the covering note assessed. We are aware that agencies follow different practices. We noted that the nature of covering notes varies considerably – from those providing substantial advice to those which are little more than procedural.

3.3 There were elements of the framework which often weren't done well and therefore impacted scores

While the papers were generally well-written and dealt with a wide range of complex issues, the Panel identified several framework elements that weren't done particularly well in the sample as a whole.

Agency panels may need more guidance on these matters to apply the standards effectively.

Particular elements were:

3.3.1 Context: Understanding context may make a difference

A frequently occurring shortcoming in the papers the Panel assessed was the lack of context. Many scored relatively poorly on this aspect of the PQF. It was unclear how papers fitted in the wider strategic context and how they built on previously made or foreshadowed future decisions.

The Panel did not have any additional contextual information.

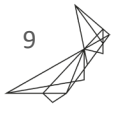
We expect that agency panels may understand more about the policy context of papers they review – even to the extent of having someone there to explain the nuances of the issues and policy processes. Agency panels may also have been provided with notes on context. NZIER received contextual information, but not in all cases.

This under-development of the context elements of the PQF could contribute to lower scores awarded by the Panel.

However, it's still important to properly tackle this part of the standards. Writing for the audience is key – Ministers may not remember all the background to an issue if it only comes across their desk infrequently; if papers are to be referred to colleagues, more context is needed; similarly, for papers going to Cabinet. Writing for the audience was a key focus during this review.

And perhaps more importantly, it needs to be clear, right up front, why the Minister is getting a paper and what they need to do with it. Considerable care needs to be taken in drafting the purpose statement – this sets the tone for the paper as a whole. In a handful of papers, it wasn't clear why the Minister was getting a particular piece of advice.

[14] Policy Advisory Group, DPMC.



The Guidelines note that assessing a stream of papers on a single or related topic can be useful. This helps focus on the fit between papers and the techniques that can be used to provide a smooth transition. It may be worth encouraging agencies to do this more often.

3.3.2 Analysis: systematic options analysis

This is a core element of the PQF. But we saw very few examples where there was a systematic assessment of a range of options against a clear set of criteria.

We know it is common practice that they are missing in Cabinet papers as some Ministers only want the paper to focus on the preferred way forward. But this doesn't explain why they are lacking in other papers.

3.3.3 Analysis: Advice: Treaty of Waitangi analysis/ te ao Māori/implications for Māori

Only a handful of papers made a decent effort at doing a Treaty analysis, looking at a te ao Māori perspective and/or setting out the implications for Māori. This was missing or underdeveloped in many papers that should have included this analysis.

This was a major shortfall and, therefore, impacted on scoring.

The Guidance on Te Tiriti o Waitangi analysis [15] has been around since 2019. But only occasionally applied in this set of papers.

In addition, looking at population impacts, including for Māori, has been a requirement for some time. We know access to data and evidence can be limited. But more is available than we saw used in this sample of papers. This is better planned early when developing the piece of advice – not left until the last minute. Agencies may have to invest in background work, data collections and capability to ensure it can be pulled through and included in their policy advice.

We understand some work is underway to provide more guidance on looking at issues from a te ao Māori perspective and incorporating this in policy advice, and it is intended to make this widely available.

3.3.4 Analysis: Advice: diverse views and perspectives not well reflected

The requirement to consider different perspectives is contained in the Analysis section of the framework. Related but different, the need to reflect diverse perspectives is a requirement in the Advice section of the framework.

With the odd exception, we saw little evidence of this in the papers that we reviewed. This impacted scores and a failure to do this can mean that key judgements, trade-offs and implementation risks may be hidden.

Even different agency views often weren't outlined. There was a list of agencies (and sometimes stakeholders) consulted – but often no indication of their views on the proposals.

3.3.5 Advice: reflecting the needs of the audience

Papers didn't seem to take enough care in delivering to the audience's needs. While we appreciate some of them were written under time duress, improvements could be made. In particular:

- They were overly long, making them hard work for busy Ministers.
- More effort was needed to tighten the Executive Summary/Key points sections (in particular, repeating elements of the background rather than focusing on the essentials of the analysis

[15] www.dpmc.govt.nz/publications/co-19-5-te-tiriti-o-waitangi-treaty-waitangi-guidance

and recommendations was a trap we saw many papers falling into); doing this well can 'save' a paper.

- Not enough distinction was made in writing for the Minister responsible (who may well have some detailed subject matter knowledge) and writing for a group of Ministers who have varied levels of understanding of the issues, the background or the technical aspects of the issue.
- There was limited use of other means of communication – diagrams, charts, tables, A3s, slide packs etc.

3.3.6 Action: Implementation matters

The analysis of implementation issues was not well developed. While there was often an indication of what further reporting to Ministers was planned. The basics of implementation weren't discussed – including:

- What an agency needed to do to deliver, e.g. staff training, IT changes, procurement, communications.
- Financial implications – for agencies – or for stakeholders.
- The implications for other parties involved in delivering changes to services, e.g. capacity and capability.
- Communications plans for the public stakeholders, users those who needed to comply with changes.
- Possible implementation risks and how they can be managed.

We note that this section of the PQF is focused on making sure those responsible know what they need to do and by when – and on monitoring and evaluation. So, this might be a gap in the PQF itself.

3.4 Robust feedback on papers helps to improve quality

The Panel was impressed by the feedback provided when assessing individual papers. The best followed the key elements of the PQF, highlighted areas of good practice (which could be repeated and used more widely and provided practical suggestions on how papers could be improved. This would be invaluable for authors, peer reviewers and managers, and if done well and given focus, it is a key tool in improving overall quality. At the other end of the spectrum, some agencies only provided a very short commentary. This was less helpful.

This section of the Guidelines is relatively short and could be developed further.[16]

[16] Page 11 of www.dpmc.govt.nz/sites/default/files/2021-08/policy-quality-framework-guide.pdf.

4 Areas for follow-up

4.1 Next steps

The next stage in the process is to meet with agency Panels.

This will help us assess the significance and relative importance of the issues mentioned above.

Based on this, we will develop options to address these matters, including changes to the Guidance for Panels. These will be tested with agencies and the Tier 2 Policy Leaders Network.

They will then be incorporated into the planned webinars with Panels.

The table below identifies the issues that need to be explored further.

Table 4: Next steps

Issue	Next steps				
	Clarify with agency panels	Develop options to address	Possible changes to the Guidance	Coverage in webinar	Other
Reluctance to score less than a 3?	✓	✓	✓	✓	
Differing interpretation of what a 4 or above means?	✓	✓	✓	✓	Changes to scoring definitions?
Different approaches to the sampling of policy papers assessed	✓	✓	✓	✓	
How to assess different types of paper/papers of different complexity	✓	✓	✓		
Areas where further guidance would be helpful, including those identified in this report: • Dealing with context • Treaty/te ao Māori/Māori implications • Options analysis • Presenting diverse views and perspectives • Implementation issues	✓	✓			Possible changes to the PQF, e.g. clarification of implementation requirements, additional policy advisory tools/training, more cross-panel appointments, developing a pool of expert panellists.

Source: NZIER

4.2 Further follow-up

It would be worth undertaking a similar exercise in future (e.g. in two years) to determine whether the changes put in place because of this exercise lead to greater consistency in scoring.

Consistency is important for a couple of reasons:

- Assessing the relative performance of different agencies.
- Assessing consistency over time so that improvements (or otherwise) can be tracked.

The latter may be more related to a panel's application of the tool and the capabilities of those doing the assessments. However, interagency consistency is less important for the overall goal of policy quality improvement.

How much weighting to put on the two purposes of the PQF (reporting/accountability and quality improvement) will help determine the relative investment in these sorts of activities to improve consistency in the medium/long term.

The question as to whether the development and implementation of the PQF have improved policy quality overall remains a big one – and the most important question. Looking at consistency is only one element of this question.

Appendix A – Phases of the project

A.1 Phase 1: Short term strategy (October 2022 to June 2023)

Commission an experienced policy specialist who could act as a roving ambassador for internal agency policy quality review panels and provide advice on good practice and any matters of concern.

Convene a policy paper sample review panel to review and re-score a sample of papers from the 2020/21 year – in order to determine the nature and scale of the problem with the consistency of how agencies score policy papers.

Prepare additional guidance for panels to assist their understanding of best practice and what good policy advice looks like.

Hold a webinar for panel members with the Policy Project, policy specialist, and the policy paper sample review panel to answer questions about the revised Guidance, outline tips and best practice.

A.2 Phase 2: Medium-term (July 2023 to June 2024)

A permanent independent moderation panel will be established if the review in Phase 1 above confirms a significant problem with consistency in agency quality of policy advice scoring. It will undertake an annual moderation exercise of agencies' policy scores in 2022/23 and 2023/24, then every two years.

Ongoing assistance to internal policy quality review panels from an experienced policy specialist.

Further Policy Project-hosted webinars for agency policy quality review panels.

A.3 Phase 3: Long-term (July 2024 onwards)

Moderation reviews every two years from 2023/24 (depending on the outcome of Phase 1).

Ongoing assistance to internal policy quality review panels from an experienced policy specialist.

Ongoing occasional Policy Project hosted webinars for agency quality review panel members.

Appendix B – Additional statistical information

Table 5: Panel scores

On average, the Panel scored papers lower than the original score.

Papers	Number	Average Panel score	Median score	Minimum score	Maximum score	Standard deviation
Originally scored 3	25	3.06	3	2	4	0.485
Originally scored 4	27	3.22	3	2.5	4	0.487
Whole sample	**52**	**3.14**	**3**	**2**	**4**	**0.488**

Source: NZIER