Incubator for AI

Consult Evaluation: DWP's Pathways to Work consultation

27/08/25

Executive summary

Consult is a tool for analysing public consultations, which uses AI to generate themes for each question and map responses to those themes.

This report outlines findings from an evaluation of Consult using a consultation run by the Department for Work and Pensions on the reform of the health and disability benefits system and employment support.

Headlines:

- ➔ We conducted a robust analysis. A total of 87,738 responses to consultation questions were reviewed by 125 people.
- ➔ Consult's theme mappings were consistent with people. Its mapping was more similar to reviewers than reviewers' mappings were to each other.
- ➔ The reviewer made no changes to Consult's mapped themes for 73% of responses. When two reviewers were compared to each other, they gave identical theme mappings for 65% of responses.
- ➔ The overall performance (F1) score was 0.816 (out of 1) when comparing Consult to a single reviewer. When two reviewers were compared to each other, then the equivalent score was 0.710.
- ➔ We found no evidence of bias. Consult's theme mapping performance did not differ with self-reported disability status.

Contents

Introduction

Background

What is the Incubator for Artificial Intelligence (i.AI)?

i.AI rapidly designs, tests, and delivers AI products for government. We take pride in being a highly technical team, composed of experts with deep knowledge and experience from both the public and private sector. This allows us to tackle complex challenges and innovate continuously towards our overall mission: to harness the opportunity of AI for public good.

What is Consult?

Each year the government runs around 600 public consultations, with some receiving more than 100,000 responses. Analysing these consultations takes hundreds of thousands of hours of civil service time, or is contracted out at substantial cost. The time-intensive process also delays the findings and, ultimately, the development of policy.

Consult, an AI-powered tool developed by i.AI, aims to alleviate this burden by facilitating quicker analysis, of comparable quality, for a fraction of the cost. It does this using a multi-step process with a human-in-the-loop:

- 'Theme Generation': Consult identifies common themes in the responses to generate a list of themes for each question.
- 'Theme Sign Off': The policy team running the consultation reviews a long list of generated themes and decides which to use for the analysis, with the ability to add, edit or remove themes.
- 'Theme Mapping': Consult classifies the full set of responses using one or more of these generated themes.

Why are we running this evaluation?

Consult has the potential to save substantial time and money for government, but it is critical that this is not at the expense of high quality analysis that captures the public's views. We are therefore carrying out evaluations to ensure that Consult is suitable for wider use, and

continually improving its delivery. We have conducted several retrospective evaluations using historic consultation data. We have also conducted an evaluation on a live consultation from the Scottish Government and a live call for evidence from the Independent Water Commission.

The purpose of this specific evaluation was to gather evidence of Consult's performance using a sample of responses to help the Department for Work and Pensions (DWP) decide whether to use Consult to analyse the full set of responses when their consultation closed.

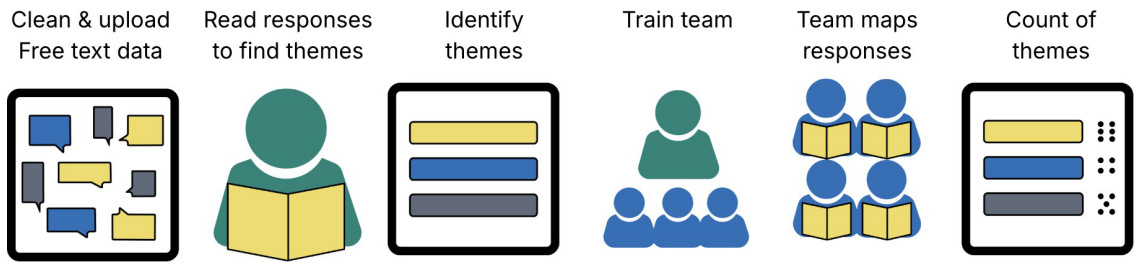
This work also builds on our previous evaluations in multiple ways:

- This is the third evaluation we have done on a live consultation or call for evidence, allowing us to better understand how Consult performs in practice.
- Unlike the first live consultation, responses were double reviewed, giving us a better idea of how two people differ when asked to review the same responses.
- This consultation addresses a new topic. To be confident in Consult's general use, we need to evaluate it across a range of different topics.
- This is the first live consultation for which we have also completed supplementary equality analysis to assess whether Consult exhibits bias.

How does Consult work?

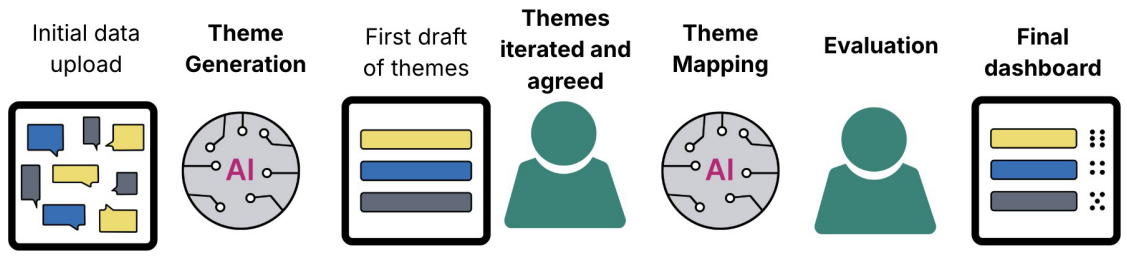
How teams currently analyse consultation responses

Conventional analysis of public consultations is a multi-stage process. Although the approach varies by consultation (for example, based on their size) it generally involves an initial review of a subset of responses to identify common themes. Once these are identified, a team is trained on the theme framework (to ensure they have a common understanding of each theme) before reading through all of the responses and mapping each response to the list of possible themes. At the end, a summary shows how frequently each theme appeared.



How Consult analyses responses with a human in the loop

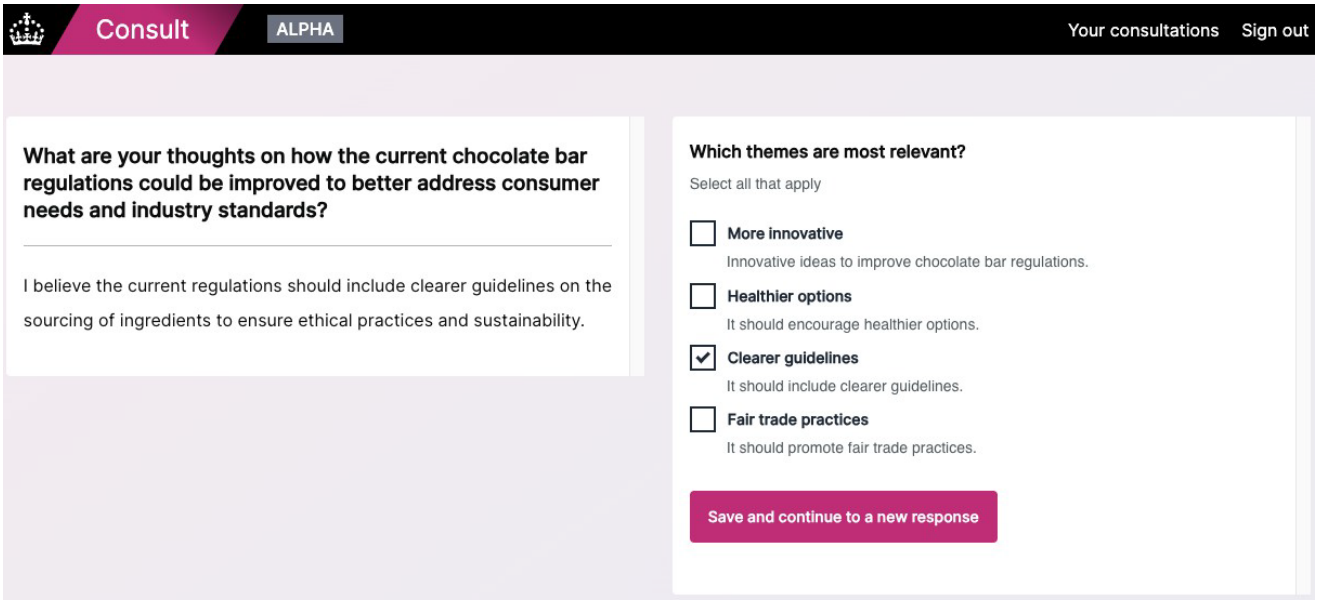
Consult broadly comprises two steps: theme generation and theme mapping. First, it uses a topic modelling approach to identify common themes for each question (using all the available responses, rather than a sample). It then maps each response to the available themes using Large Language Models (LLMs). The mappings are summarised in a dashboard allowing policymakers to see how often each theme appeared, read the original response content, and synthesise their findings.



This two stage process allows us to include a 'human in the loop'[1] at each stage. In the first stage (theme generation), expert reviewers working on the policy area can review and approve the themes before they are used in the mapping process.

The second stage (theme mapping) was the focus of this evaluation. Reviewers verified (and, if necessary, amended) the theme mappings that Consult provided for each response. Using the Consult application, reviewers were presented with an individual free text response to an individual question and the themes that Consult had mapped to it. If they deemed it necessary, reviewers made changes to the themes which Consult had mapped to the response using the tickboxes. Once they were content with the set of themes mapped to the response, they saved the mapping and continued to the next response. You can see a hypothetical example below.

Fig 1: Example of the interface used by reviewers to assess and amend themes.



The theme mapping evaluation was conducted in two phases using 125 reviewers in total.

- In the first 'double review' phase, two reviewers independently judged the mapping that Consult had provided for a sample of 1000 responses for each of the 17 questions in the consultation.
- In the second 'single review' phase, one reviewer assessed the mapping that Consult had provided for up to 5000 responses for each question. If there were not yet 5000

[1] 'Human in the loop' is when human review and decision-making is incorporated into an AI process. This ensures that the outputs are regularly checked, and aims to improve the quality of the process.

responses for a given question, then the full set of available responses were used (details in the Annex).

Between the two phases, experts working on the policy area provided further feedback on the themes for each question.

All of our code is open source

We have two open source repositories:

- ThemeFinder is a python package containing the AI capabilities (theme generation and mapping). We separated these into their own package to make it easier to test, develop and share with others.
- Consult is our application, containing an interface for both evaluation and the final dashboard. We plan to continue developing and improving this into a self-serve tool.

Research questions

This evaluation was focussed on answering four questions about Consult's efficacy:

- **How similarly do two people map themes to responses?**
- **How similarly does Consult map themes to the reviewers?**
- **To what extent do differences between Consult and the reviewers impact the headline findings of the consultation?**
- **Does Consult's performance differ by disability status?**

Theme mapping evaluation

Methodology

Once the assigned themes had been reviewed, we had a dataset that included (for each response) the themes that were assigned by Consult and the themes assigned by the reviewer(s).

Table 1: Illustrative data snippet (theme names replaced with letters for simplicity).

Respondent	Question #	Consult themes	Reviewer 1	Reviewer 2
1	1	A, B	A, B, C	A, B
2	1	B, D	B, D	B

Research questions

Our main analysis aimed to answer the following questions:

- **How aligned were the Consult-mapped themes with the reviewer-mapped themes?**
- **To what extent do differences between Consult and the reviewer themes impact the implied headline findings?**

For every response, we also gave the reviewers the opportunity to choose the following two options (which were not available to Consult):

- 'None of the above': The response discusses a topic not covered by the listed themes.
- 'No reason given': The response does not provide a substantive answer to the question.

As well as providing us with useful data for our equality analysis (details provided in a subsequent section), this also helped us get an indication of whether the theme framework (which had been signed off after expert feedback) was missing themes which covered a large proportion of responses.

Finally, we used data from Consult to answer the question:

- **How long did it take reviewers to review Consult's theme mappings?**

For our analysis, we primarily investigated two comparisons.

- Using the data from the 'double review' phase, we compared Reviewer 1 to Reviewer 2 to estimate a 'baseline' level of person-to-person variation on this inherently subjective task.
- Using the data from the 'single review' phase, we compared Consult to a reviewer to assess how aligned the Consult-mapped themes and reviewer-mapped themes were.

Measuring alignment

To assess the alignment between sets of mapped themes, we calculated an F1 score – a common measure of a model's performance. Broadly, it considers how often the model correctly identifies a theme (a true positive), versus how often it identifies a theme that is not correct (a false positive), or misses a theme that should be there (a false negative). When reporting Consult's performance, we take the reviewer-identified themes to be the 'correct' answer, and compute the F1 score against this standard[2]. We used a multi-label F1 score that can credit partial matches (where one theme is correct but another is not), and penalises both false positives and false negatives.

We calculate the F1 score at a response level, and then average across all the responses. For example, row 1 below has an F1 score of 0.8 and row 2 has a score of 1 (perfect alignment), so the average score is 0.9. By calculating the score at a response level and averaging them, longer responses with more themes do not disproportionately drive the score.

Table 2: Example F1 score calculation for a single pairwise comparison.

Respondent	Question #	Consult themes	Reviewer Group 1	F1
1	1	A, B	A, B, C	0.8
2	1	B, D	B, D	1
3	1	-	-	-
			Average	0.9

For a fuller explanation of the F1 score and how it is calculated, see the Annex.

[2] In reality, there is nothing saying the reviewer label is 'correct'. Theme mapping is often subjective, and we use the reviewer-to-reviewer comparison from the double review evaluation phase as our baseline to decide whether Consult's performance is acceptable.

However, for responses with no identified themes (row 3 above) it is not possible to define an F1 score, so these responses are dropped from the calculation.[3]

Theme ranking

Whilst the number of responses mapped to each theme is important, the qualitative interpretation of the consultation is driven more by their relative frequency (e.g. whether one theme is more prevalent than another) than the absolute numbers. We used the respective theme frequency rankings as a proxy for the expected qualitative interpretation implied by Consult's theme mappings and the reviewers' theme mappings. To understand the extent to which the changes made by the reviewers affected these rankings we calculated the difference in rank for each theme. We used these differences to calculate both the 'average distance moved' and the 'maximum distance moved'. The lower these distances are, the more confident we can be that the same qualitative insights would be drawn from the consultation as a whole using either set of mappings. For this analysis, we excluded infrequent themes (those with fewer than 5% of the responses for that question mapped to them) to avoid instances where small changes in frequency lead to misleadingly large changes in the rank. This matches the intuition that low frequency themes would not impact the overall interpretation of the consultation as much as high frequency themes.

Time taken

As the theme mapping evaluation was conducted using the Consult application, we were able to store the timestamp when each response was first seen by a reviewer. We used this to calculate the duration between consecutively reviewed responses to get a proxy for the time taken to check the theme mapping provided by Consult.

[3] We could choose a metric that assigns these rows a value of 1, given the Consult and Reviewer themes are the same. An F1 score does not consider 'true negatives' like these. While there is an argument that Consult 'correctly' didn't assign any themes, we have deliberately excluded these rows to avoid biasing our performance estimates upwards.

Findings

Consult had good overall performance, and produced mappings that were more aligned with a single reviewer than two reviewers' mappings were to each other.

During the single review phase of the evaluation, the themes mapped by Consult were unchanged by the reviewer for 73% of responses. This is higher than the consistency observed during the double review phase where the two reviewers gave identical theme mappings to each other 65% of the time.

Summarising performance across all the responses checked in the single review phase, the mean F1 score for Consult was 0.816. Again, this is higher than the equivalent score when comparing two reviewers, which was 0.710.

There is no set definition of a 'good' F1 score: some sources suggest 0.8-0.9 is good,[4] others over 0.7,[5] and some suggest even lower scores may be considered good in specific contexts.[6] However, the fact that Consult achieves a higher score than the person-to-person baseline suggests that, at minimum, Consult's mapping performance is as good as a human process.

Table 3: Summary of theme mapping performance.

Evaluation Phase	# of responses reviewed	Comparison	Identical	Mean F1 Score	95% CI F1 Score
Double Review	17,000	Reviewer 1 vs. Reviewer 2	65%	0.710	[0.703, 0.717]
Single Review	70,738	Consult vs. Reviewer	73%	0.816	[0.813, 0.818]

Mean F1 scores by question, and with bootstrapped sampling distributions, are in the Annex.

Differences between Consult and the reviewers had minimal effects on the overall theme rankings.

From a policy perspective, identifying the most common themes in a consultation is usually more important than knowing the exact number of times each theme appeared. We would be

[4] https://encord.com/blog/f1-score-in-machine-learning/
[5] https://spotintelligence.com/2023/05/08/f1-score/#What_is_a_good_F1_score
[6] https://julius.ai/glossary/f1-score

more concerned, for example, if differences in theme mapping led to the top theme shifting to third place, than if the most common theme was consistent but was counted 280 times rather than 300 times. Comparing the theme rankings resulting from Consult's mapping and those resulting from the reviewers' mapping (during the single review phase), we find some differences but they are generally small with themes moving, on average, half of a rank.

Table 4: Movement of theme rankings by question.

Question number	Number of themes above 5% threshold	Average distance moved by each theme	Maximum distance moved by a theme	Number of themes moving more than 2 ranking places
1	6	0.3	1	0
2	10	0.4	1	0
3	8	0.0	0	0
4	10	1.0	2	0
5	7	0.9	3	2
6	11	0.7	3	2
7	10	0.4	1	0
8	12	0.8	2	0
9	8	0.5	2	0
10	7	0.6	2	0
11	6	0.0	0	0
12	7	0.0	0	0
13	11	0.6	2	0
14	8	0.5	2	0
15	5	0.0	0	0
16	7	0.0	0	0
17	7	0.9	2	0
All	**140**	**0.5**	**3**	**4**

Reviewers used 'None of the above' for 4.5% of responses.

When reviewers select 'None of the above' they are conveying that a response contained a substantive answer to the question that belonged to a theme not included in the listed themes. The fact that 'None of the above' was mapped to fewer than 5% of responses in the single review phase suggests that the agreed list of themes was sufficiently exhaustive. This is a proxy evaluation of the theme generation and sign-off processes which reassures us that no commonly occurring themes were overlooked.

Reviewers were more likely to add themes than remove them.

Inspecting the specific changes made by the reviewers during the single review phase of the evaluation, almost three times as many theme mappings were added (19,047) as were removed (6,498). This pattern was consistent across the different questions. This suggests that the reviewers erred more on the side of including themes than Consult does.

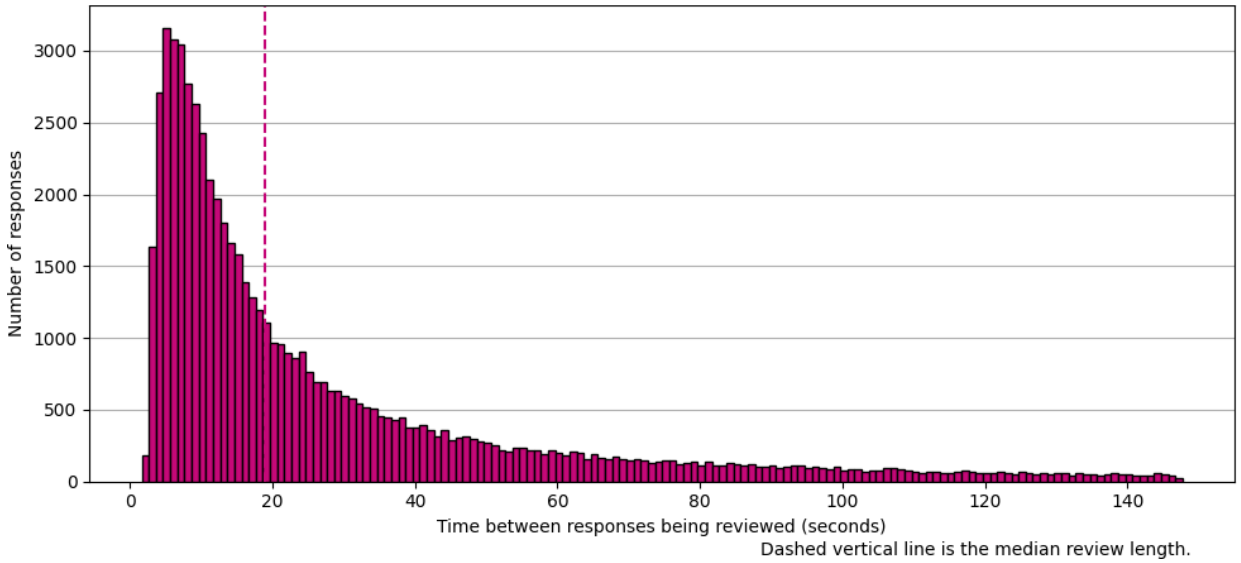
Reviewing themes generally took little time, but there was some variation to account for response complexity.

The median time taken to review a response was 19 seconds, with 78% of responses reviewed in less than a minute. The time taken to review responses was related to the length of the response and the number of changes reviewers made to the theme mapping. This suggests that, while the review process was quick, reviewers did take the time needed to assess longer and more complex answers.

Table 5: Length of response and average theme changes by time taken to review.

	Time (seconds)						
	0-5	5-10	10-20	20-30	30-40	40-60	60+
Avg length (words)	14	21	33	44	52	60	70
Avg changes (themes)	0.40	0.39	0.53	0.64	0.69	0.72	0.72

Fig 2: Histogram of time taken to review responses (below 90th percentile).



Equality analysis

Methodology

After they had provided their answers to the 17 questions comprising the consultation, respondents were also asked to provide a limited amount of demographic information. One question posed in this section was, "Do you consider yourself to have a health condition or disability?" to which respondents could answer "Yes", "No", or "Prefer not to say". For all the responses where we had demographic information, we used it to perform supplementary analysis investigating whether Consult exhibited any bias with respect to this self-reported disability status.

Research questions

The first two research questions our equality analysis aimed to answer straightforwardly relate to the theme mapping step in the Consult process:

- **Does the rate at which reviewers make changes to Consult's mappings differ by disability status?**
- **Does Consult's performance (i.e. mean F1 score) differ by disability status?**

Additionally, we aimed to answer a research question which constitutes a proxy evaluation of the theme generation step in the Consult process:

- **Does the rate at which reviewers assign 'None of the above' differ by disability status?**

We know that theme generation must necessarily be a condensation exercise to make the analysis of the responses intelligible. However, we would be concerned if the views of respondents with particular demographic characteristics were less likely to be included in the final list of themes used for the mapping step of the Consult process. We can use the data provided by the reviewers to assess the extent to which this is occurring. This is because by selecting 'None of the above' reviewers are conveying that a response contained a substantive answer to the question that belonged to a theme not covered by the listed themes.

Measuring differences in the 'None of the above' rate

To investigate whether the proportion of responses which were mapped to 'None of the above' was associated with disability status, we calculated a contingency table and performed a Chi-squared (χ^2) test of association. This test compares the observed frequencies in our data with the frequencies that would be expected if disability status and the probability of a response being mapped to 'None of the above' were truly independent.

We calculated a p-value for the test. This tells us the probability of observing the data that we obtained under the assumption of independence. If we observe a low p value, then we might choose to make the inference that the variables are unlikely to be independent. However, with such a large sample size we are very likely to obtain such 'statistical significance'. As a consequence, our primary metric of interest for this portion of the analysis was instead an effect size (Cramer's V). This gives us an indication of the potential practical significance of any differences that we observe (irrespective of their 'statistical significance').

Findings

'None of the above' was used at a similar rate for those who report having a disability and those who do not.

The rate at which 'None of the above' was assigned by reviewers is almost identical for respondents who answered "Yes" and "No" to the disability status question, whilst the rate is slightly higher for those who answered "Prefer not to say". As expected, our Chi-squared test of association yields a very small p-value. However, we can see that this statistical significance does not translate into practical significance. This is evident from the value of 0.02 for our effect size which is considerably below the 0.1 threshold that is conventionally held to indicate a 'small' effect[7].

[7] Cohen, J. (1998) Statistical Power Analysis for the Behavioural Sciences. Chapter 7.

Table 6. 'None of the above' rate by disability status and the results of Chi-squared test.

Disability status	% of responses 'None of the above'
Yes	4.41
No	4.39
Prefer not to say	5.80

χ^2 test statistic	p value	Cramer's V [95% confidence interval]
19.2	0.00007	0.02 [0.01, 0.03]

Consult's theme mapping performance does not differ with disability status.

The rate at which reviewers made changes to Consult's mapped themes is very similar for respondents who answered "Yes" and "No" to the disability status question. It is slightly different for those who answered "Prefer not to say" but the confidence interval for this group still overlaps with the confidence intervals for the other two groups. Additionally, the rate is slightly lower for the "Prefer not to say" group which is intuitively less concerning as it implies better performance. Similarly, the overlap observed in the confidence intervals for the mean F1 scores for each of the three disability status groups reassures us that Consult's theme mapping performance is broadly equivalent.

Table 7. Performance metrics by disability status.

Disability status	% of responses reviewers changed [95% confidence interval]	Mean F1 score [95% confidence interval]
Yes	27.0 [26.6, 27.4]	0.819 [0.815, 0.822]
No	27.6 [26.7, 28.5]	0.815 [0.806, 0.822]
Prefer not to say	25.8 [24.5, 27.0]	0.808 [0.796, 0.819]

Conclusion

Consult is an effective tool for mapping themes to responses.

- Consult's theme mappings were consistent with people. Its mapping was more similar to reviewers than reviewers' mappings were to each other.
- During the single review phase of the evaluation, the reviewer made no changes to Consult's mapped themes for 73% of responses. When two reviewers were compared to each other, they gave identical theme mappings for 65% of responses.
- The overall performance (F1) score was 0.816 (out of 1) when comparing Consult to a single reviewer. When the reviewer groups were compared to each other, then the equivalent score was 0.710.
- Differences between Consult and reviewer-mapped themes had minimal impact on the overall theme rankings, implying similar qualitative interpretations.

Consult's theme mapping performance does not differ with disability status.

- The rate at which 'None of the above' was assigned by reviewers is almost identical for respondents who answered "Yes" and "No" to the disability status question at 4.41% and 4.39% respectively.
- The overlap in the confidence intervals for the mean F1 scores across the three disability status groups indicates that Consult's theme mapping performance is generally comparable for respondents with and without disabilities.

Annex

Supplementary figures

Figure 3: Mean F1 scores per comparison with bootstrapped sampling distributions.

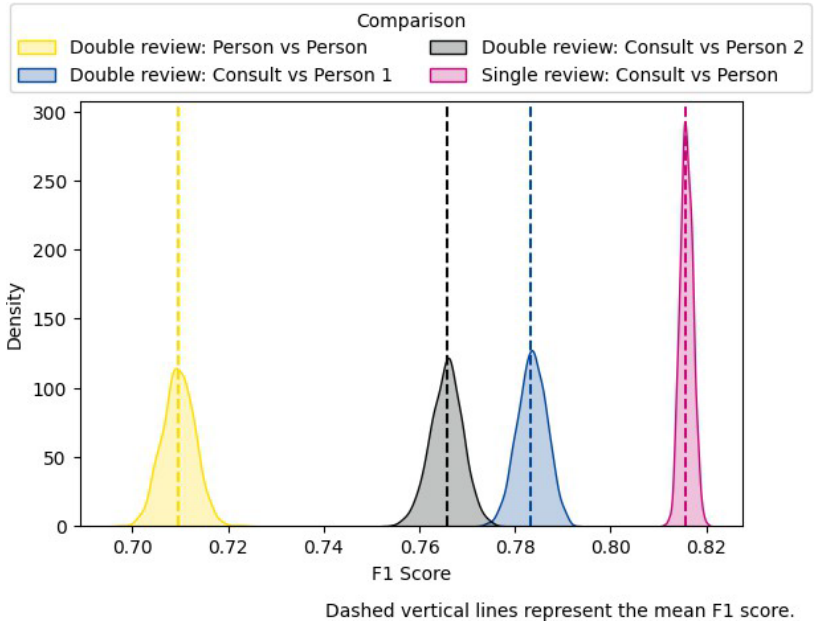
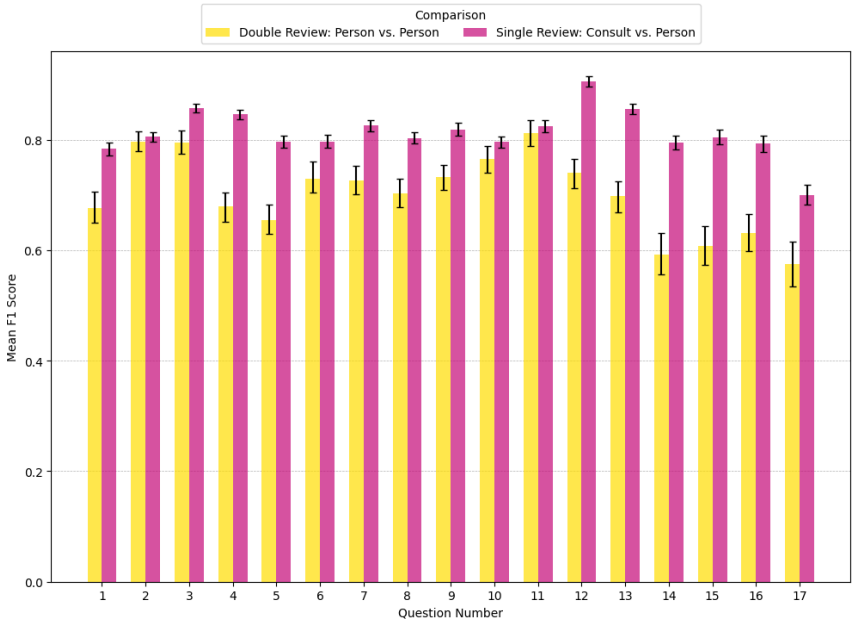


Figure 4: Mean F1 scores per question and comparison with 95% confidence intervals.



Supplementary tables

Table 8: Number responses reviewed by question.

Question Number	Double review phase	Single review phase
1	1000	4992
2	1000	5000
3	1000	5000
4	1000	4651
5	1000	4885
6	1000	4355
7	1000	4089
8	1000	4097
9	1000	3402
10	1000	4578
11	1000	3532
12	1000	3513
13	1000	4529
14	1000	3925
15	1000	3753
16	1000	3324
17	1000	3113
Total	**17000**	**70738**

Theme stability check (A.k.a. 'Dip Test')

Due to the practical constraints of the consultation period, it was necessary for DWP to begin using Consult to generate and map themes (for the purpose of evaluation) before the consultation had closed. This is no different than the way in which themes are generated in the current, manual consultation analysis process (in fact, using Consult allowed DWP to start generating themes later than usual). However, it still created a risk that the themes that were generated (and used to evaluate Consult's mapping) may not be representative of the full set of responses that Consult would later be used to analyse[8]. To alleviate this risk we conducted a small 'Dip Test' evaluation after the consultation had closed, in which Consult's theme mappings for 200 responses to each question were assessed by a single reviewer.

Table 9: Summary of theme mapping performance (including the 'Dip Test').

Evaluation Phase	# of responses reviewed	Comparison	Identical	Mean F1 Score	95% CI F1 Score
Double Review	17,000	Reviewer 1 vs. Reviewer 2	65%	0.710	[0.703, 0.717]
Single Review	70,738	Consult vs. Reviewer	73%	0.816	[0.813, 0.818]
Dip Test	3,400	Consult vs. Reviewer	73%	0.849	[0.839, 0.860]

The results of the 'Dip Test', with a mean F1 score of 0.849, indicate that Consult's theme mappings remain highly consistent with human labellers. This suggests that, despite starting theme generation before the consultation closed, Consult's outputs are reliably representative of the entire set of responses. The high level of agreement provides reassurance that any potential impact of data drift on the analysis is minimal and supports the decision to use Consult to analyse the full consultation.

[8] This is an example of the phenomenon known as 'data drift' in machine learning engineering where the data used for model training and evaluation are not representative of the future data the model will be expected to handle when deployed.

Figure 5: Mean F1 scores per comparison with bootstrapped sampling distributions (including the 'Dip Test' results).

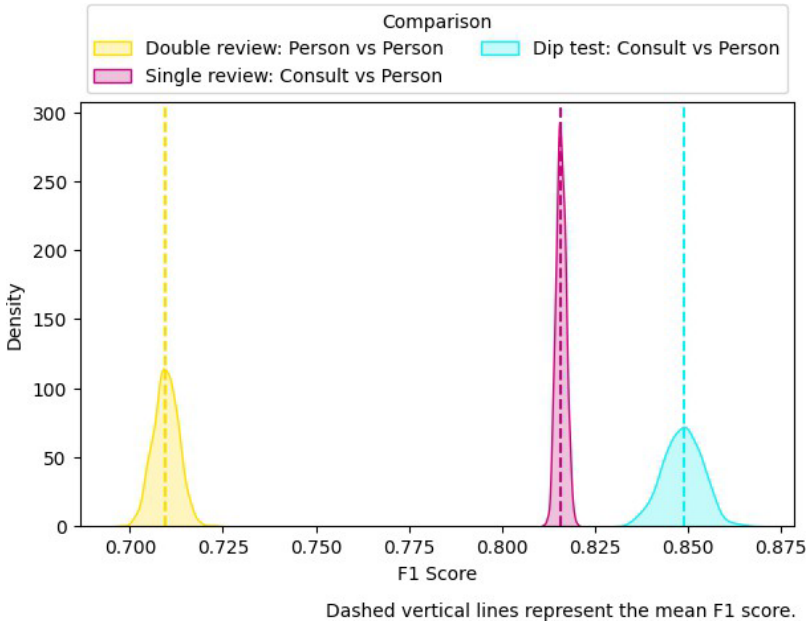
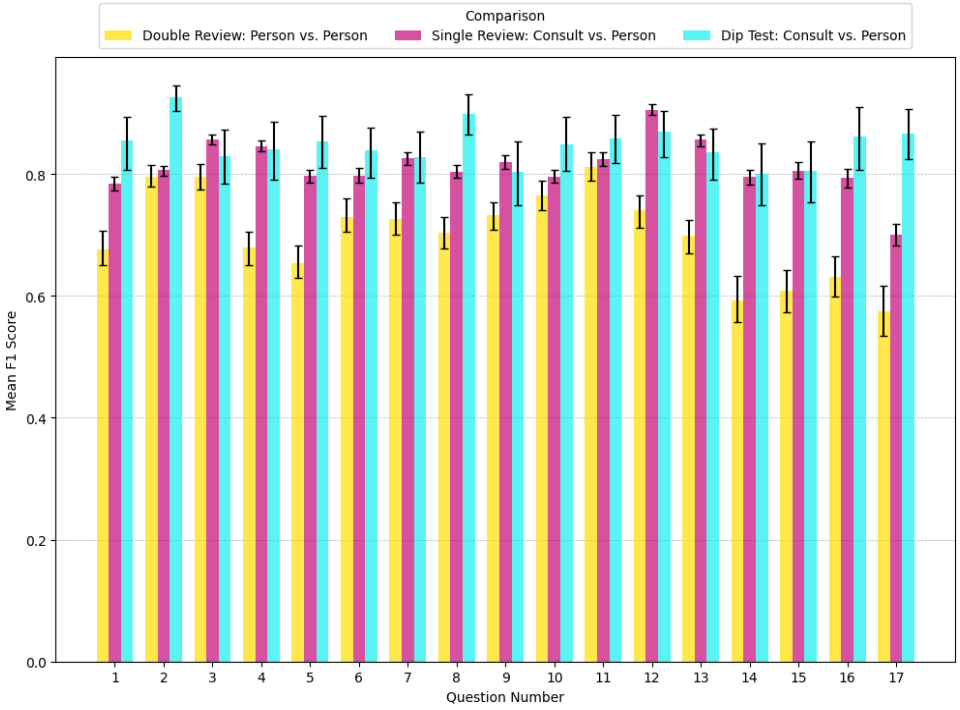


Figure 6: Mean F1 scores per question and comparison with bootstrapped confidence intervals (including the 'Dip Test' results).



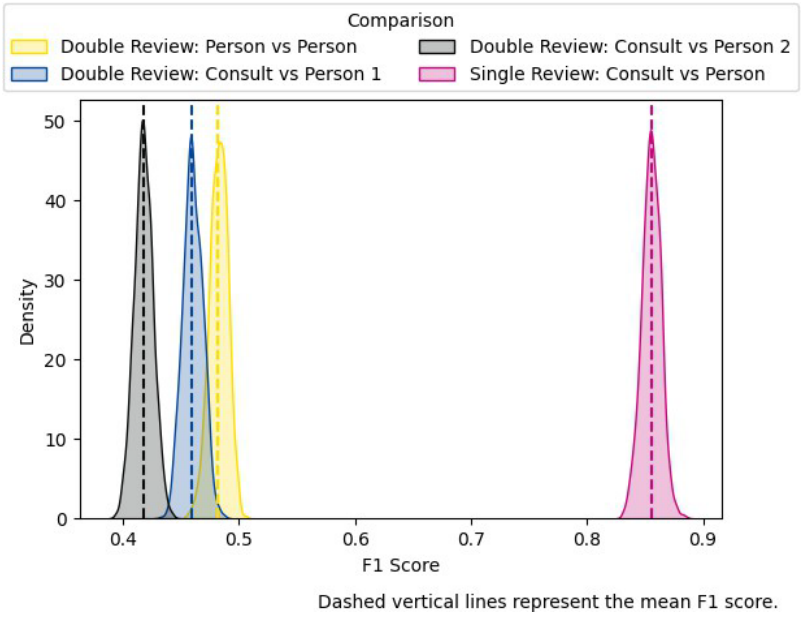
Campaign results

In addition to the analysis of the main consultation responses, Consult was also used to map the responses submitted to a third-party website-related campaign. This involved using Consult to map responses to the following three free text questions:

- Question 3. "What impact would losing PIP have on your life, or on someone you care for?"
- Question 7. "Is there anything else you'd like to say about these changes or how the government can better support sick and disabled people?"
- Question 9. "Is there anything else you'd like to say about PIP or how the government can better support sick and disabled people?"

Our evaluation of Consult's performance on these responses followed the same structure as the evaluation we conducted on the main consultation responses. That is, we again completed an initial 'double review' phase (of 1000 responses per question) followed by a subsequent 'single review' phase (of 500 responses per question).

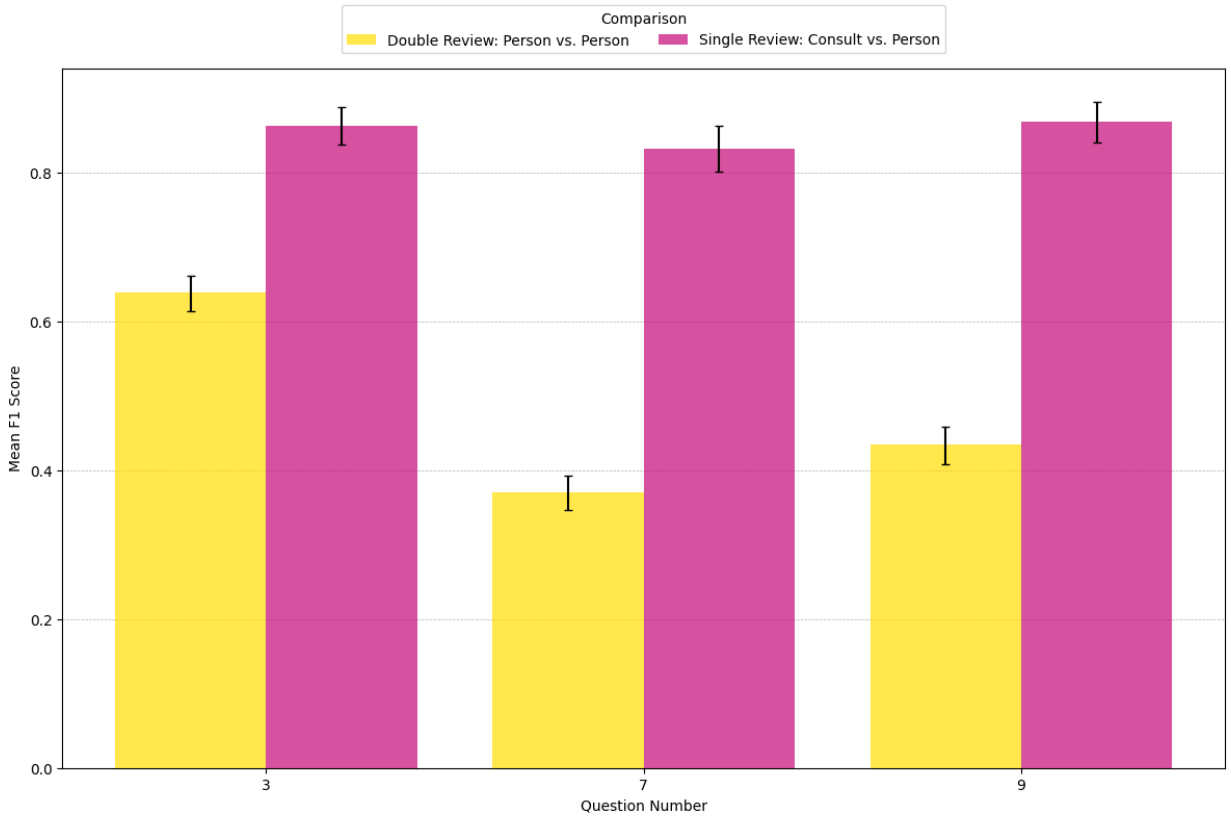
Figure 7: Mean F1 scores per comparison with bootstrapped sampling distributions for campaign responses.



The results from the double review phase indicated that Consult's performance, with F1 scores of approximately 0.46 and 0.42 when compared to human labellers, is slightly below the agreement observed between the human labellers themselves (F1 score of around 0.48). Though it is worth noting that we do see some overlap of confidence intervals. Notably, all three pairwise comparisons yielded relatively moderate F1 scores in absolute terms.

However, in the follow-up single review phase, Consult's performance improved markedly when compared to a single human labeller, achieving an F1 score of approximately 0.86 with a confidence interval from 0.84 to 0.87. Taken together, these findings are sufficiently encouraging, indicating that Consult's performance matches, and potentially surpasses, the typical level of agreement seen between human labellers.

Figure 8: Mean F1 scores per question and comparison with bootstrapped 95% confidence intervals for campaign responses.



Calculating the multi-label F_1 score

Defining common terms

This annex provides a more detailed explanation of the multi-label F_1 score that we used to assess the performance of Consult relative to the reviewers.

To clarify the definition, we can first define some common terms using the classic 2x2 confusion matrix setup.

- Note that we're only using the simpler 2x2 case to make our initial definitions clearer. In the actual analysis we also extended the metric to the multi-label case.

		Predicted category		
		Positive	Negative	
Actual Category	Positive	True Positive □TP□	False Negative □FN□	Σ□Actual Positives □AP□
	Negative	False Positive □FP□	True Negative □TN□	Σ□Actual Negatives □AN□
		Σ□Predicted Positives □PP□	Σ□Predicted Negatives □PN□	

Building on the 2□2 matrix above, we can define eight ratios that help us characterise the performance of our classifiers (i.e. the output from the reviewers and Consult):

True Positive Rate $TPR = \frac{TP}{AP}$	False Negative Rate $FNR = \frac{FN}{AP}$
• The proportion of actual positives that are correctly predicted to be positive. • $TPR = 1 - FNR$ • A.k.a. Recall, Sensitivity	• The proportion of actual positives that are incorrectly predicted to be negative. • $FNR = 1 - TPR$ • A.k.a. Type II error
True Negative Rate $TNR = \frac{TN}{AN}$	False Positive Rate $FPR = \frac{FP}{AN}$
• The proportion of actual negatives that are correctly predicted to be negative. • $TNR = 1 - FPR$ • A.k.a. Specificity	• The proportion of actual negatives that are incorrectly predicted to be positive. • $FPR = 1 - TNR$ • A.k.a. Type I error
Positive Predictive Value $PPV = \frac{TP}{PP}$	False Discovery Rate $FDR = \frac{FP}{PP}$
• The proportion of predicted positives that are actually positive. • $PPV = 1 - FDR$ • A.k.a. Precision	• The proportion of predicted positives that are actually negative. • $FNR = 1 - TPR$
Negative Predictive Value $NPV = \frac{TN}{PN}$	False Omission Rate $FOR = \frac{FN}{PN}$
• The proportion of predicted negatives that are actually negative. • $NPV = 1 - FOR$	• The proportion of predicted negatives that are actually positive. • $FOR = 1 - NPV$

F_1 score

The F_1 score combines two of the eight ratios from the table above, True Positive Rate and Positive Predictive Value, to give an aggregated measure of performance in relation to the

'positive' class. Specifically, F_1 score is the harmonic mean of the True Positive Rate (TPR) and the Positive Predictive Value (PPV):

$$F_1 = \frac{2}{TPR^{-1} + PPV^{-1}}$$

During our evaluation, more than one theme could be mapped to a given response. This meant that we were evaluating a multi-label classification problem and needed to adjust our metric accordingly.

To do this, we converted the theme mappings into sparse matrices where each column represented one of the theme labels that could have been chosen and each row represented a response to the consultation. We then calculated the F_1 score for each individual response and averaged these scores to get the aggregate values reported. The benefit of this approach is that each response contributes equally to our assessment of performance.

This process is illustrated in the hypothetical below:

- Imagine that there were four possible themes (A, B, C, D), and for a particular response we received the following mappings from our labellers:
    - Reviewer = (A, C)
    - Consult = (A)
- We would convert the data into sparse matrices with a column per possible topic:
    - Reviewer = (1, 0, 1, 0)
    - Consult = (1, 0, 0, 0)
- We would calculate the True Positive Rate and Positive Predictive Value for this response:
    - TPR = 0.5
        - There are two 'actual positives' in the reviewer's answer.
        - Consult has correctly 'predicted' one of them.
    - PPV = 1
        - There is one 'predicted positive' in Consult's answer.
        - This predicted positive is 'correct' according to the reviewer.
- We would calculate the F_1 score for this response using the formula above:

- $F_1 = \frac{2}{TPR^{-1} + PPV^{-1}} = \frac{2}{\frac{1}{0.5} + \frac{1}{1}} = \frac{2}{3}$

- This process would be repeated for all the responses and the arithmetic mean of the response-level F_1 scores would be calculated.

We calculated this statistic using a function from scikit-learn's metrics module: f1_score().

Confidence intervals

In addition, we generated a confidence interval for the F_1 score using a process called bootstrapping. This involved taking a large number of random samples of responses from the data and calculating the F_1 score for each of these samples. This process gave us an approximate sampling distribution for the metric itself. Confidence intervals can be extracted from this sampling distribution by finding the relevant percentiles. We calculated 95% confidence intervals by finding the 2.5th and 97.5th percentiles. The confidence interval gives us a sense of the uncertainty we have in our metric.