Crowdsourcing measurement of e-government usability

Tom Nicholls[*]and Simon Gray[†]

Paper for presentation at IPP 2014

1 Introduction

In an age of austerity, the cost-effectiveness of government is of increasing concern to policy-makers. The government is once again pushing the adoption of effective digital service delivery as one method of achieving cost savings while preserving the quality of public services. As a result, the quality and ease of use of governments' online offerings has become increasingly important in recent years.

However, this drive to put services online has not been matched by a corresponding effort to develop the measures needed to assess whether these websites are doing a good job. This problem is particularly acute in local government, where the number of discrete services delivered by a single organisation is high, and the websites are correspondingly complex and difficult to navigate.

In parallel, there has been an increase in interest in recent years in crowdsourcing – achieving difficult goals by mobilising the contributions of many individuals rather than by a team of professionals. Crowdsourcing has been used in recent years by groups

[*] Oxford Internet Institute, University of Oxford: tom.nicholls@oii.ox.ac.uk (corresponding author)

[†] Birmingham City Council: simon.gray@birmingham.gov.uk

as diverse as scientists classifying galaxies (Clery, 2011), civic activists collecting data on how scenic different parts of the country are (MySociety, 2014), and Google correcting transcription errors in scanned books (Ahn et al., 2008).

Against this background, we describe an alternative approach to measuring local government website quality based upon crowdsourcing: the council website usability dashboard ("Dashboard").

> "The purpose of this site is to build up a picture of the usability of different council websites in England, Scotland, and Wales – but doing so by letting users carry out a series of task-based tests, to gradually build up a comprehensive objective picture of the usability of the sites across a much broader range of tasks than other methodologies do, by crowdsourcing the data." (LocalGov Digital, 2014a)

By conducting a quantitative analysis of the full set of ratings provided by users on the Dashboard, we examine the nature of the Dashboard as a crowdsourcing tool, and ask empirical questions about local government web provision and its measurement.

We find that an analysis of Dashboard test results supports the previous experimental finding that Google is a more effective navigation method for government websites than internal navigation (National Audit Office, 2007), but that the Dashboard data as it currently stands is not yet sufficiently complete for its primary purpose of measuring local government web quality.

We suggest that policymakers should further explore the possibilities of crowdsourcing and the greater use of Google as a preferred access method for local websites.

2 Background

Although the notion of crowdsourcing work via the Internet is a relatively new one, the idea of citizens and consumers co-producing services goes back a long way in the public policy and public administration literatures. Parks et al. (1981) drew attention to consumer co-production in areas such as education, Neighbourhood Watch provision, and the curbside collection of refuse. Subsequent literature has particularly focused on childrens' and adults' social services literature (e.g. Prentice, 2006).

Relating more strongly to the Internet, there is a substantial literature flowing from Benkler (2002) looking at what he refers to as "commons-based peer production", characterised by decentralised information-gathering and the breaking of work into fine-grained tasks maybe taking only a few minutes each.

The Dashboard draws from both these traditions. The creators are local government web specialists, familiar with both the Internet practice of gaining insights from "found" user data and focused A/B testing, and also the local government tradition of local citizen / customer interaction. In this case, the aim is to co-produce service evaluation, using crowdsourced methods.

Where the present case is perhaps different from the body of literature on service user co-creation is that Dashboard participants are not necessarily a user of services from the local authority being rated: the scope is Internet-wide rather than geographically restricted.

Looking at local government web measurement more broadly, the leading method in the United Kingdom is Socitm's annual Better Connected (Socitm, 2013) set of surveys, which use expert assessors to judge the quality of particular transactions.

However, relying on Better Connected has some limitations: the size of the problem and the availability of reviewers means that only a small number of "top task" services are judged in any one year[1]; as a result of the cost of the exercise the detailed results are limited to paid subscribers to Socitm's *Insight* service; and the judgements of the assessors are necessarily those of interested experts rather than the typical users of council websites who are local citizens with wildly varying levels of online skill.

[1] A further issue with the top tasks approach is that it ignores the long tail of pageviews on council websites. Whilst top task service pages are accessed an order of magnitude more often than other individual service pages, the cumulative number of non-top task pageviews is still very high.

It would be useful to have a rating scheme which in addition to being reliable is broad-based, open, and timely. That is the context within which the Dashboard was created.

3 Dashboard development, motivation, and data

Dashboard volunteers can, after registering, rate any one of 70 tasks, at any one of 176 local authorities[2]. This is an ambitious undertaking. Even allowing that not all councils deliver all services[3], it implies a desire to collect data on each of 10,958 council / task pairs. As the creators note above, this is a much wider range of tasks than other methodologies, and hence a much larger number of tasks to rate. That the crowdsourcing approach normally relies on having multiple raters for each item in order to ensure reliability makes the volume of data sought even more impressive.

Each task is tested and rated "according to the four different methods of navigating to complete the task – navigation from the home page, navigation using the A–Z directory, searching on Google, and searching using the site's internal search" (LocalGov Digital, 2014b). The ratings that can be given are ordinal: 'Very easy', 'Easy', 'Difficult', and 'Gave up'. It is also possible to skip assessing one or more navigation methods and give no rating.

The Dashboard is still in development; although it is open for public beta testing, its creators are clear that it is not to be considered finished. Although no particular audience is identified on the publicly-facing parts of the website, the principal users at present consist of members of council web teams. That is to say that in this particular case, the crowd which is being sourced is not the general public but largely a

[2] Testers are required to register with the site in order to ensure commitment to the work of testing. Ultimately, the intention of the Dashboard developers is to enable web teams to embed test forms onto the relevant pages of their own sites in order to better solicit real user test results, but early testing showed users were more inclined to use the form to comment on the quality of service delivery rather than the usability of the website.

[3] Some of the authorities are in two-tier areas, in which case separate councils deliver different parts of the service mix

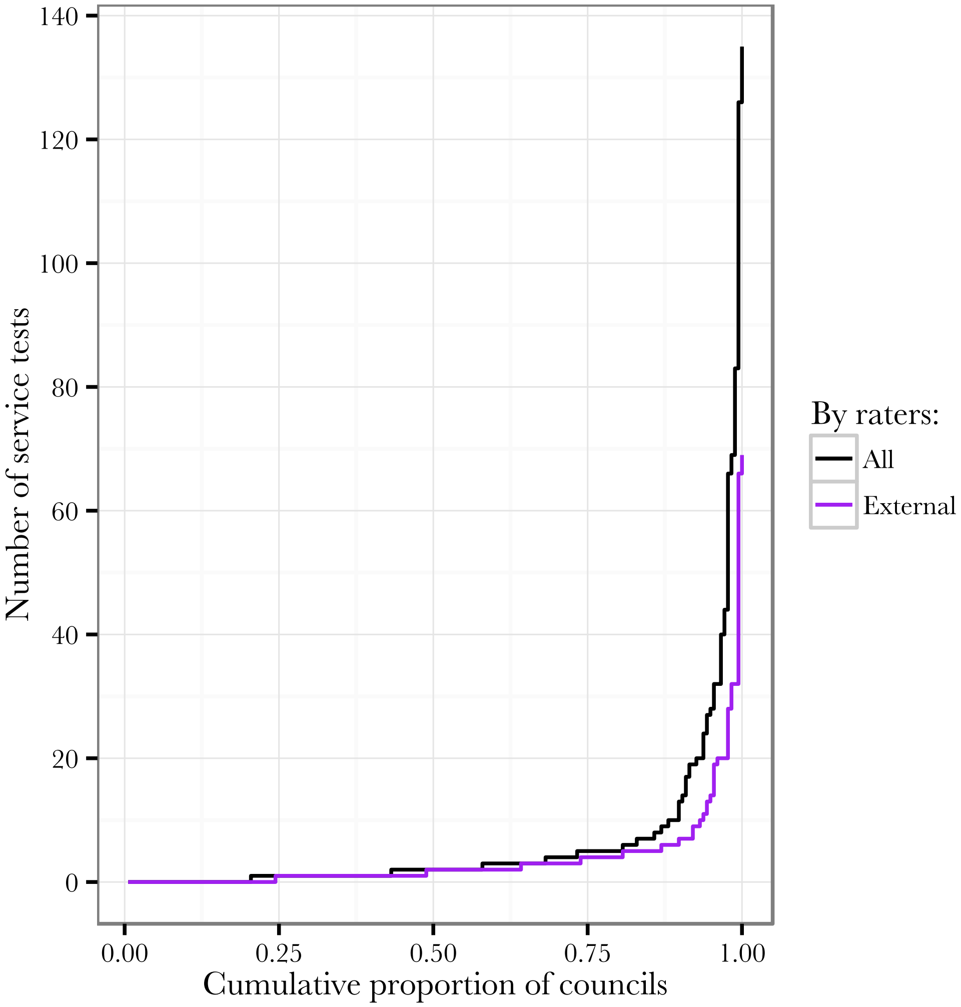

Figure 1: Number of Dashboard service tests for each council

disciplinary group of colleagues. Of the data collected to date, 999 of the 1198 service assessments (83%) have been carried out by people with .gov.uk email addresses[4]. Furthermore, 517 of the 1198 assessments (43%) were carried out by raters with an email address that matched the domain being rated – that is to say that they were internal reviewers rather than external peers or members of the public. The distribution of assessments between councils is shown in Figure 1.

[4] It's also possible that some of the raters not coded as having .gov.uk email addresses are local government officers as there is no requirement to give a work email address or institutional affiliation when registering.

As a public dataset, a summary of the ratings are published (with identifying information removed) including free format comments made by testers. This is intended to enable webteams to more clearly see and share examples of good and bad practice to aim for and avoid. The Dashboard's creators intend in the future to publish the the raw data in XML format to further enable interested parties to carry out more detailed analyses.

Because testers identify themselves and their level of expertise before completing a series of tests, additional relevant demographic information is gleaned such as age, computer literacy, disabilities, and levels of expertise: eventually it will be possible to drill down into the results to see what effects these have on the ease of carrying out the tests. In principle, action can be taken to ensure that making it easy for one demographic does not make it difficult for other demographics. At present, data is too sparse for such analyses to be useful.

4 Research questions and method

- How can data from the dashboard develop our understanding of the present usability of local authority web provision?

- Do the data from the Dashboard support the experimental finding that Google is a more effective navigation method for government websites than the websites' own navigation tools (National Audit Office, 2007)?

- Are the Dashboard overall council scores usable as a substitute for Better Connected results (Socitm, 2013) for measuring the overall performance of council web services?

- Does this example of crowdsourcing provide insights about or suggest other opportunities for crowdsourcing the evaluation of policy and service delivery in other areas of government?

The key methodological issue with analysing the Dashboard's performance is a relative shortage of data. Although there are 1198 individual service ratings have been made on the Dashboard, there are also 10,958 possible council / task pairs, only 953 of which have at least one rating in the database. The majority of these have only one judgement recorded (see Figure 2, noting the logarithmic scale). This number may grow in time, as the site comes out of beta and is more widely publicised, but for present purposes of evaluation there are difficulties in selecting appropriate statistical tests which are robust to the large areas of missing data.

The overall distribution of task ratings across councils is summarised in Table 1 and graphed in Figure 3. One key observation is the relatively high level of "gave up" responses for navigation methods which are actually under the control of the local authority, as opposed to the Google approach.

	A–Z	Google	Internal Search	Navigation	Sum
Gave up	203	106	212	224	745
Difficult	121	97	141	221	580
Easy	197	226	228	273	924
Very Easy	348	588	532	452	1920
Sum	869	1017	1113	1170	4169

Table 1: Distribution of task ratings

5 Comparative channel usability

As each rater on the Dashboard is invited to attempt four different ways of completing the task – via Google, the site's own search, navigation from the site home page, and the use of a council A–Z of services if present – the Dashboard presents an interesting dataset for looking at the relative merits of search and navigation strategies as ways of presenting government information. This has previously been explored using experimental methods (National Audit Office, 2007, §D), with a finding that the "open search"

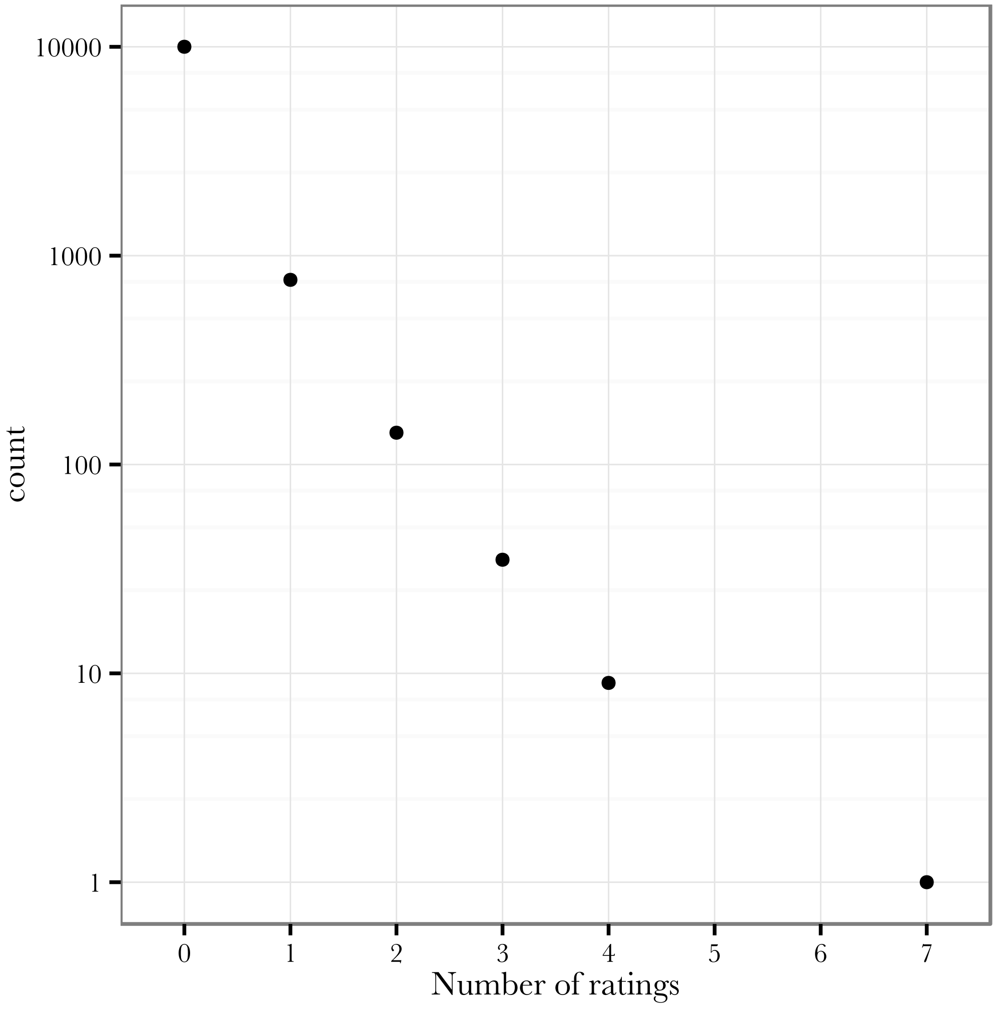

Figure 2: Number of ratings for each council/task pair

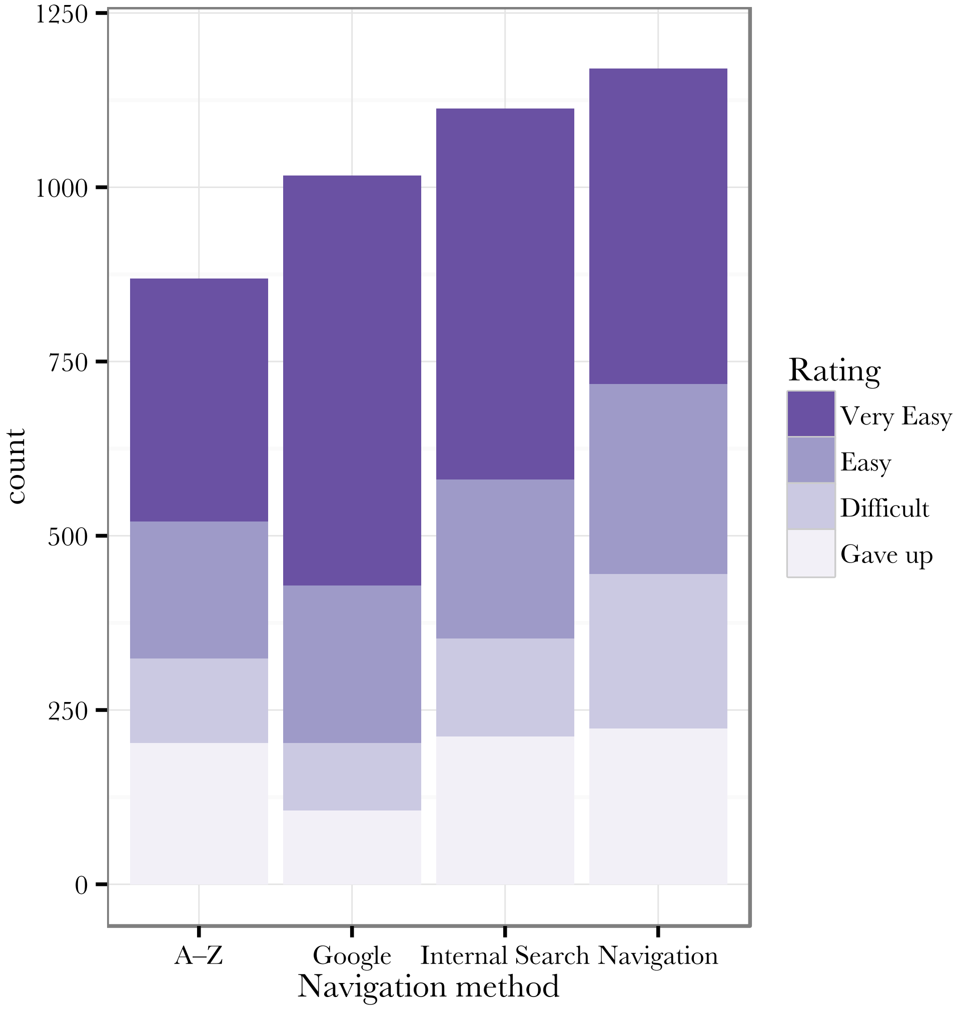

Figure 3: Distribution of ratings for each council / task pair, by access technique

(Google) treatment was faster than internal navigation for experimental subjects using DirectGov, a previous UK central government portal.

The data from the Dashboard can be used to further test this hypothesis, focusing particularly on the subjective experience of participants rather than the objective measure of time used in the earlier study. Both are relevant to an understanding of comparative channel usability: if the time of the interaction is extended, customers are not being treated efficiently. If the subjective experiences of customers are bad, however, they may choose another delivery channel.

Hypothesis 1 *Dashboard users found Google more usable for navigation on council websites than the websites' own navigation structures.*

The data are summarised by Table 1 and Figure 3. For further analysis of this question, only those volunteers who provided a matched pair of ratings (by rating both Google searchability and internal navigation for the same task) are included, so N = 1004. As the data are in matched groups, a one-tailed Wilcoxon signed rank test was selected to assess the two approaches.

As is evident from Figure 3, there is indeed a statistically-significant (and sizeable) difference between the Google and internal navigation ratings. The test statistic is 92644, with p < 0.001. As a result, we can reject the null hypothesis and accept H1a: Google is more usable for navigation than websites' own navigation structures.

This has implications for the development of council websites, as well as their future evaluation. If Google is a more efficient tool for finding local government web content than internal methods, it is arguable that more effort should be made to encourage citizens to access their services via a search engine, that councils should de-emphasise internal search and navigation and that the Better Connected series should explicitly focus on Google navigability[5]. It would be interesting to back up this analysis with

[5] Better Connected does consider Google, but only does so in the context of selecting the one route for each service that it thinks citizens are most likely to use: "For each task (except for task 7) we determined

an insight into what proportion of page views on council websites are generated via Google, what proportion are via front-page navigation, and whether the relative usability of each route has an impact on individual council proportions.

If desired, a similar analysis would be possible for the other two access methods. Looking at Figure 3 again, it seems that sites' internal search engines are less usable than Google and the A–Z approach is less usable than both. This is slightly surprising, given that the existence of an A–Z of services is held to be an important part of website usability (Socitm, 2013, p.38).

6 Council Performance

This section presents a comparison of the Dashboard's ratings against those of the Better Connected survey. Comparison is complicated by the sparseness of the Dashboard's data, which arises in part from its ambitious scope. Nevertheless, we present a comparison of website ratings using the Dashboard and Better Connected 2013 (Socitm, 2013), alongside an inter-coder reliability analysis to measure the consistency of ratings amongst members of the Dashboard's user community.

6.1 Inter-coder reliability

To conduct an inter-coder reliability trial, a request was made by the Dashboard operators to registered users of the site to rate three specific councils on four specific tasks[6].

a primary route, eg a Google search, that in our opinion, was most likely to be used by a member of the public, and asked questions about each step in the journey from that point" (Socitm, 2013, p.40).

[6] The councils selected were Birmingham City Council, Manchester City Council, and Copeland Borough Council; the tasks "Local services and information on a map – look at", "Trading Standards – complain about a dodgy item from a shop", "Missed bin collection – complain about one", and "Planning permission – apply for it". The names of the tasks highlight the use of everyday expressions rather than local government jargon. Words such as "bin" rather than "refuse" have been increasingly used online by councils in recent years.

Hypothesis 2 *The inter-coder reliability between raters of services on the Dashboard is sufficiently high that raters are freely interchangeable and consequently a single rating on a service can be treated as reliable.*

Krippendorff's alpha was selected as the test statistic, with the liberal threshold of 0.667 considered to be the "lowest conceivable limit" for tentative conclusions (Krippendorff, 2004, p.241).

Despite the number of users approached, only a single rating was acquired for one of the specified council / task pairs as a result of the call. This precluded the planned analysis based upon a compact set of solicited responses.

As an alternative, the same inter-coder reliability statistic was computed for the council / task pairs which had been serendipitiously rated at least twice during the normal operation of the site to date. This has two major disadvantages compared to the planned trial:

1. The resulting data matrix is very sparse, with many raters participating a handful of times rather than a few raters rating all items.

2. The trial conditions are not well controlled. In particular, the versions of the sites rated will be different because raters accessed the sites at different times. The instructions given to raters have also evolved over the lifetime of the Dashboard.

Although the latter in particular will affect the hypothesis results, it will do so by lowering the alpha statistic and so is a conservative bias.

The internal navigation scores were selected for the trial, being the navigation method with the largest number of results, as well as one where we might expect to find inter-coder variation as volunteers adopted different routes through the site. As some raters have rated the same council / task pair more than once, only the first such rating has been taken into account for the inter-coder reliability study.

With 221 raters and 953 council / task pairs, the final alpha statistic is 0.0653. This is far below the 0.667 specified. As a result, we must fail to reject $H2_0$: there is insufficient evidence that raters can be treated as freely-interchangeable.

The difficulty in producing a successful inter-coder reliability trial does provide one interesting learning point as to the community of raters behind the Dashboard. On the evidence of the request to the userbase, it doesn't seem like the Dashboard is currently usable to apply the concentrated attention of a crowd to a particular task of interest to an authority: even with the benefits of a centrally-distributed email to all registered users requesting assistance, only a single response was received. The consequences of this on the utility of the Dashboard as a service for targeted test-driven development of council websites are obvious.

6.2 Comparison with Better Connected

Hypothesis 3 *The Dashboard's overall ratings of councils are statistically distinct from the Better Connected results (Socitm, 2013) as a way of measuring the overall performance of council web services.*

The sparsity of data is a current problem with using the Dashboard as a general purpose performance measurement tool. A major advantage of the crowdsourcing approach is to bring the "wisdom of crowds" to bear on questions of interest. With the majority of completed tasks having only one datapoint, and 80% of the councils having fewer than five service scores (see Figure 1) this advantage is lost as scoring is based on one user's opinion only.

If the results of the inter-coder reliability trial had indicated that the coders were interchangeable, it would have been possible to treat the Dashboard results as a fixed set of accurate data, at which point the shortage of multiple scores would not have been an issue. With this not possible, the only statistically defensible approach is to treat the Dashboard scores as individual samples, with inherently wide confidence intervals due to the small numbers for each council / task pair.

To test H3, then, we can conduct a Wilcoxon signed rank test, using the council aggregated score for the Dashboard rating, and the 1*–4* categories from Better Connected. If the test result is statistically significant at the 0.05 confidence level, it supports the claim that the ratings are drawn from different populations and therefore measuring distinct things. If the test result is not significant, then it is possible either that the Dashboard and Better Connected are measuring the same underlying set of qualities, or alternatively that there is insufficient statistical power in the test run to be able to differentiate the two. In either case, we would fail to reject $H3_0$.

For the overall council ratings, the Dashboard's own published data summarisation and aggregation method has been used. This has two advantages, one conceptual and one pragmatic. Conceptually, the raters may be influenced in their choice of Likert scale responses by a knowledge of how the site uses these to construct ratings; this is particularly true as the method is prominently displayed in an "About the scores" section at the bottom of the council summary tables. Pragmatically, a method of producing a composite measure needs to be chosen; there is no particular reason to deviate from that used by the site itself, given that it seems well thought-out[7].

> "Scores are assigned according to 5 = Very easy, 3 = Easy, 1 = Difficult, and −3 = Gave up. As more users complete any given test, the score for each aspect of each task becomes a mean average of all the individual testers scores. [...] The Usability Index is the sum of the scores of all the tests which have been carried out, divided by the number of tests which have been taken for this council, resulting in an index score between −12 and 20." (LocalGov Digital, 2014b)

[7] There is one difference between the numbers calculated for this study and those published on the Dashboard itself. The Dashboard's calculation method assumes that each rater provided a rating for all four navigation methods, penalising those councils where raters provided fewer than four ratings. This study divides proportionally to the number of scores actually input.

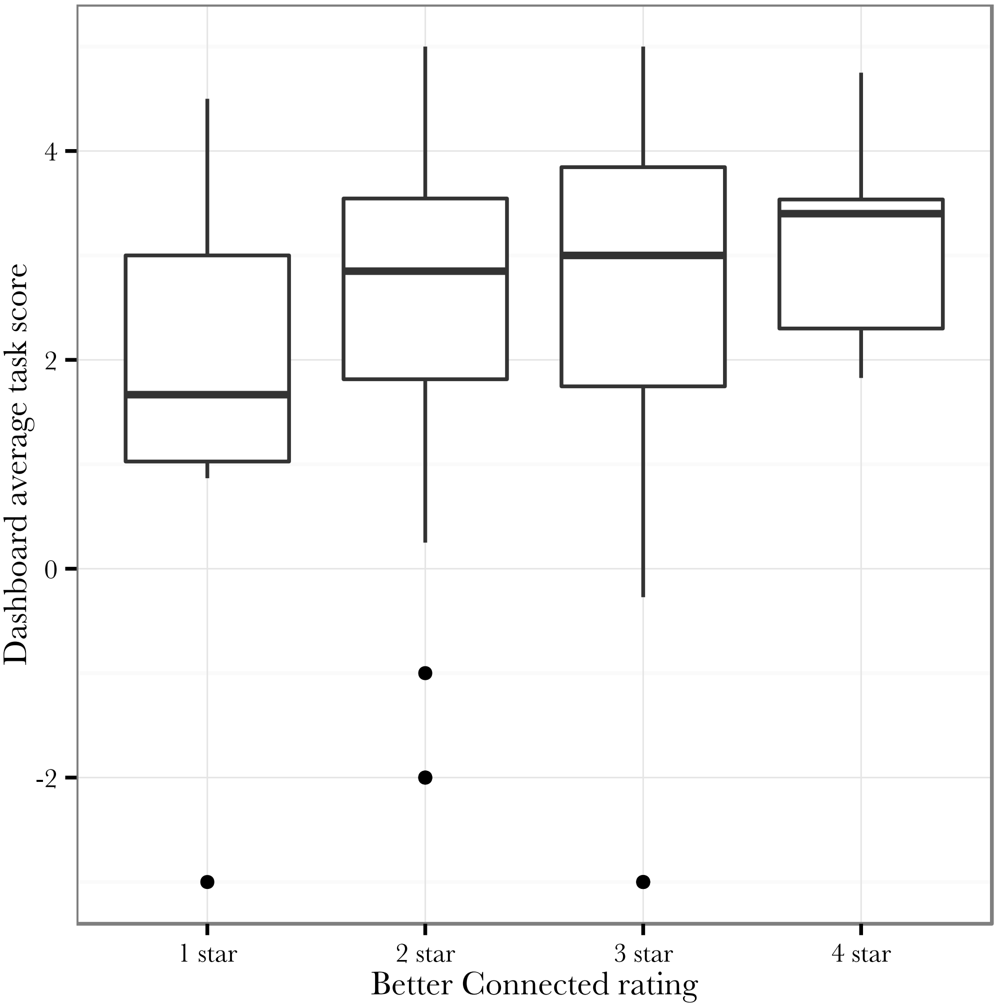

Figure 4: Better Connected scores and average Dashboard task ratings, by council

These Usability Index scores provide a single measure of council performance, which can be compared with the Better Connected ratings.

The Dashboard dataset contains at least one point of data on 144 councils. Although the Better Connected dataset is bigger, covering 433 authorities, only 316 of them are labelled by name[8]. The overlap, and therefore the sample size for the study, is 115. As the sample is non-random, p-values have little relevance.

[8] As Better Connected is part of Socitm's *Insight*, a subscription service, non-subscribers are not named in the report or underlying dataset, though they are assessed for comparison and listed as "County council 3" etc.

The distribution of Dashboard task ratings for authorities in each Better Connected rating is plotted in Figure 4. There does seem to be an positive association between Dashboard and Better Connected ratings, but at first glance it isn't a strong one, and the majority of the effect seems to be for authorities with a Better Connected rating of 1 star.

In the final analysis, the Wilcoxon signed rank test (two-tailed to reflect the "equal to" null hypothesis) results in a test statistic of 3418 and a p-value of 0.57. As a result we are unable to distinguish the results of the Dashboard and Better Connected results and must fail to reject $H3_0$.

This result in particular is clearly influenced by the relative lack of Dashboard data. With the volumes of Dashboard data currently available, coupled with the findings of the inter-coder reliability trial, it is difficult to draw firmer conclusions about the variation between the two scoring systems. As it is, the Dashboard data has several characteristics in common with the Better Connected approach: the majority of scores are single tests per task carried out by experts.

7 Conclusions and policy implications

It is clear that the Dashboard is not currently in a position to challenge Better Connected as a means of providing robust judgements about the whole of an authority's web provision. Although over a thousand service tests have been completed, the scale of the data required and the need for multiple judgements to average out the high inter-coder variation means that the per-council aggregated ratings must be treated with caution.

Ultimately, the primary audience for the Dashboard as currently constituted is not the public: it is peer professionals and bureaucrats. Although the data is publicly accessible, the thrust of the site is not towards increased accountability, but towards the producer-as-data-consumer: providing feedback and quality data to support service improvement.

In that sense, the Dashboard is not actually being used as a crowdsourcing or co-production effort in the sense of the literature discussed. The high number of council employees amongst the rater base and the low-level focus of the ratings suggests that the Dashboard is better seen as a practitioner-driven service improvement network, akin to a benchmarking club or knowledge-sharing network. The difference, in this case, is the use of the Internet as a platform to solicit and deliver peer feedback and have an instant pool of data from other authorities to compare performance against.

Greater responsiveness is possible by using a year-round crowdsourced system as a replacement or supplement for a once-a-year expert evaluation. It is possible that this may enable decision timescales for online services to be shortened, though this would depend on whether waiting for Better Connected data to be refreshed is currently a bottleneck in decision-making or policy implementation. Free access to the full dashboard dataset also potentially assists policymakers in having access to timely data and potentially opens possibilities for further crowdsourcing of analysis and recommendations for development.

We note the disappointing level of response from the registered users to the request for specific tasks to be carried out, and suggest that one lesson here is that for crowdsourcing to work as a methodology, the crowd itself has to give its committed support.

The findings of the comparative channel usability analysis also have policy implications. The content of local government web sites is diverse and its organisation complex; if Google is truly a more efficient way for citizens to find services and information about local government, the current navigation-heavy approach to web site design in local government may be in need of a rethink.

Furthermore, despite the previous local government reputation drive towards the publication of a (paper) A–Z of services (Local Government Association, 2006) and its electronic version's continued use as a measure of website usability (Socitm, 2013), the evidence from the Dashboard to date suggests that A–Z navigation is less usable than might be hoped. This finding should be treated with particular caution given that the Dashboard's audience is primarily made up of web developers at present and not end-users, but it is worthy of further investigation.

More broadly, the Dashboard provides a potential model for local authorities to explore crowdsourcing more broadly. Web site evaluation is in no way the only informational service which could benefit from the attention of the public. For this to be effective, though, a much broader-based approach to participant recruitment would be necessary. In this regard, the Dashboard provides an interesting software platform, but only begins to show the way towards a broader crowdsourcing approach.

References

Ahn, Luis von, Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum (2008). 'reCAPTCHA: Human-based character recognition via web security measures'. In: *Science* 321.5895, pp. 1465–1468. URL: http://dx.org/10.1126/science.1160379 (visited on 02/09/2014).

Benkler, Yochai (2002). 'Coase's penguin, or, Linux and the nature of the firm'. In: *Yale Law Journal* 112, pp. 369–446. URL: http://www.yalelawjournal.com/pdf/354_t5aih5i1.pdf (visited on 07/01/2011).

Clery, Daniel (2011). 'Galaxy Zoo volunteers share pain and glory of research'. In: *Science* 333.6039, pp. 173–175. URL: http://doi.org/10.1126/science.333.6039.173.

Krippendorff, Klaus (2004). 'Reliability in content analysis'. In: *Human Communication Research* 30.3, pp. 411–433. URL: http://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Local Government Association (2006). *Local government reputation campaign: Delivering for people and places*. London: Local Government Association. ISBN: 1-84049-522-7. URL: http://www.reputation.lga.gov.uk/lga/aio/251829 (visited on 30/06/2014).

LocalGov Digital (2014a). *Council website usability dashboard*. URL: http://council.usability-test.org.uk/ (visited on 30/07/2014).

LocalGov Digital (2014b). *Council website usability dashboard: Test results for Birmingham City Council.* URL: http://council.usability-test.org.uk/council-results?CouncilID=1 (visited on 13/08/2014).

MySociety (2014). *ScenicOrNot.* URL: http://scenic.mysociety.org/ (visited on 02/09/2014).

National Audit Office (2007). *Government on the Internet: Progress in delivering information and services online – research report.* London: National Audit Office, p. 89. URL: http://www.nao.org.uk/wp-content/uploads/2007/07/0607529_Research.pdf (visited on 11/07/2014).

Parks, Roger B., Paula C. Baker, Larry Kiser, Ronald Oakerson, Elinor Ostrom, Vincent Ostrom, Stephen L. Percy, Martha B. Vandivort, Gordon P. Whitaker and Rick Wilson (1981). 'Consumers as coproducers of public services: some economic and institutional considerations'. In: *Policy Studies Journal* 9.7, pp. 1001–1011. URL: http://doi.org/10.1111/j.1541-0072.1981.tb01208.x.

Prentice, Susan (2006). 'Childcare, co-production and the third sector in Canada'. In: *Public Management Review* 8.4, pp. 521–536. URL: http://doi.org/10.1080/14719030601022890.

Socitm (2013). *Better Connected 2013: A snapshot of all local authority websites.* Better Connected. Northampton, UK: Socitm. 248 pp. ISBN: 978-1-907608-28-5.