

Improving the data HMRC collects from its customers – Submission

Arun Advani
Associate Professor of Economics
University of Warwick

8th October 2022

Executive summary

- HMRC statistics are immensely valuable for policy analysis. Expanding their usefulness and relevance is vital, and the areas put forward for expansion would all substantially improve our understanding of the UK economy and support the ability to develop economic policy (including, but not limited to, tax policy).
- I **support in the strongest possible terms** each of the proposed improvements that have been suggested, as each of these has numerous use-cases, just a small subset of which I touch on below.
- One higher level comment is that the framing of questions in the consultation regularly raises questions targeted specifically at businesses and agents, as both providers and users of the data. Another obvious set of potential users are people in the academic and policy research communities. It would make sense both to proactively send this consultation to such users, based on any existing mailing lists that are held, and to ensure that in future there are also questions directed to these users.

About me

Arun Advani is Associate Professor of Economics at the University of Warwick. He is also a Research Fellow at the Institute for Fiscal Studies, and a Visiting Fellow at the LSE International Inequalities Institute. He was a member of the Department for Education's Skills and Productivity Board (SPB) for the duration of its existence. He is also co-leading a project, in collaboration with the Department for Education, enhancing and extending the Longitudinal Educational Outcomes (LEO) data, which make use of HMRC data. He studies issues of tax compliance and tax design, with a particular focus on those with high incomes or wealth; and also issues of education and skills development in the labour market.

My research

Through my tax-related research I am a frequent user of many of the datasets discussed in the consultation, included particularly heavily both the Real Time Information (RTI) data on PAYE income tax payers, and the Income Tax Self Assessment data (ITSA).

Through my education and labour market research, both as an academic and as a former member of the SPB, I also have a particular interest in expanding our knowledge around how individuals move through the labour market, across occupations and locations, and the role this plays in improving skills and supporting growth through better allocation of talent. Many of the proposals included in this consultation would support our ability to understand these issues.

Detailed responses

The business sector of the self-employed

As the consultation document lays out, there are a number of reasons why having business sector data collected consistently for the self-employed would be of value.

1. In the context of COVID-19, policy targeting could potentially have been done better, reducing the fiscal cost, if policies could have been made sector specific, and thus better accounted for the impact of COVID on different industries.
2. From a compliance standpoint, there is good evidence that there is variation across industries. For example, the construction and transport industries have very high shares of individuals who are non-compliant, while the amount of revenue owed among the non-compliant is highest in the hospitality and legal industries.¹ Having better, more consistent data on industry among the self-employed would therefore support HMRC functions in “collect[ing] the right tax” and making it “hard to bend or break the rules.”
3. More broadly, opportunities for tax planning vary across industry
4. There is also value for wider policymaking, where – especially given the growth in self-employment – sectoral information would help better understand economic trends in the UK. This is important for numerous users seeking to understand economic developments, for example both those entering into education and educational facilities themselves, who would value the information about what to study/what courses to provide. Of course there are limitations to using such data, but they are likely to be smaller than ‘flying blind’ as currently.

In collecting data, it would be helpful to use a procedure that quickly links the entry made to a SIC code, rather than relying on manual coding or *ex post* automated matching to SIC codes. Manual coding is not effective at scale. And *ex post* matching is more likely to be subject to errors. Since many of the collection processes (e.g. income tax self-assessment) are typically filed online, using a dropdown box system for individuals to narrow to increasingly closer industries might be the most effective way to achieve this.

Relating to the note in the consultation document about ‘sector-specific taxes’, I add the obvious word of warning that the assignment of sector-specific taxes must necessarily be based on something other than this self-declared industry, else there is likely to be a large response in terms of what industries are declared.

In summary, I **strongly support** the proposals for improvement relating to business sector data for the self-employed: both making business description a compulsory field, and changing “nature of business field” to SIC.

I would also go further, and propose to collect this information for partners as well as the self-employed, while these changes are being made, else some partners involved in small partnerships will remain excluded and be the subject of a future consultation to rectify the issue later.

¹ Advani, A. (2022) “Who does and doesn't pay taxes?”. *Fiscal Studies*, 43(1). DOI: <https://doi.org/10.1111/1475-5890.12257>.

The occupations of employees and the self-employed

Information on occupation would be incredibly useful to a number of users. Some of the use-cases for such data include:

- As previously, from a policy standpoint it becomes possible to target policies like ‘furlough’ and SEISS based on occupations that were actually affected by the need to stay at home
- For educational institutions, trends in occupation are again useful information in understanding on which courses they may need to increase or decrease capacity, given changes in the occupational structure. Of course there are limitations to using such data, but they are likely to be smaller than ‘flying blind’ as currently.
- For both organisations such as the Office for Students, and for students themselves, there is an interest in better understanding how students progress post-study. Having occupational data would allow clearer analysis of where different education and training courses lead students, and hence support informed decision-making around these courses.
- For policy, understanding transitions between occupations is important as part of the growth agenda, helping to recognise which transitions are and aren’t happening, and where policy interventions could help individuals make transitions to higher paying occupations. This is something that was considered by the Skills and Productivity Board, and is likely to be of value to its successor, the Unit for Future Skills.
- For researchers of inequality, these data would make it possible to answer questions like whether individuals living in different regions progress through occupations at different rates (see also the next question on location of work place), and whether post-maternity women rates of job transition are slower (see also question about hours worked).
- For inequality reporting, such as mandatory gender pay gap disclosure, if businesses routinely collected occupational data in a structured way, this would also allow some understanding of whether some occupations had higher levels of inequality than others. This would support more targeted interventions by managers in the business to tackle such issues, were they so inclined.

In summary, I **strongly support** the proposals for collecting data on occupations for employees as part of RTI and via ITSA for the self-employed, and the approaches to doing this.

Specifically, I agree with the proposed two-stage approach to collection for employees, which would require initial “baseline” collection, and then simply updates when employees change occupation. I also agree with the proposal to make data collection compulsory, to ensure representativity of responses. Currently these data are collected for a 1% sample in ASHE, and recent work has shown that in practice compliance is both imperfect and non-random, so that some types of employers are less likely to send in information.² Existing SOC codes are certainly the right approach to collecting these data in a way that is useable by the maximum number of users, due to the consistency with the main existing occupational taxonomy in the UK.

Similar to SIC code collection, dropdown boxes narrowing down to the correct occupation may be more reliable than relying on automatic matching.

Also similar to with SIC codes, it would be appropriate to ensure partners are also captured in the same way as self-employed individual, to understand better the nature of what they actually do.

² Stokes, L., Forth, J., Ritchie, F., Singleton, C., Phan, V., Bryson, A., Whittard, D. and McKenzie, A. (in press). “Using ASHE to examine trends in low pay: Initial exploration of the data”. Bristol: Low Pay Commission URL: <https://www.wagedynamics.com/wp-content/uploads/2022/06/Note-on-ASHE-cross-sectional-weights-May2022-final2.pdf>.

The location(s) of an employment or a business

The consultation document highlights a number of reasons why this would be useful in better understanding the spatial location of activity within the UK.

I agree with these, and would **support the collection of these data**.

Of all the suggestions, this is probably the one I would least prioritise, if not everything could be done, since in many cases home address is not a terrible proxy for activity, within large geographic regions (e.g. travel to work areas). And, again as highlighted by the document, in some of the other cases workers are travelling to multiple locations, so a single “place of work” may not be appropriate.

Despite this, it would still improve the accuracy of what can currently be done, and would also allow better study of commuting decisions at a large scale, which would be highly valuable.

If these data are being collected, the proposal to do so via RTI, and the fields proposed are appropriate. As described in the consultation document, the problem of unknown actual place of work is more acute for employees than for the self-employed.

The hours employees work

This would be extremely valuable information, as currently it is not possible to measure in administrative data either the official or the effective wage that individuals are paid. There are many use cases, including but are not limited to:

- Knowing from population administrative data how much of income differences across groups (e.g. between women and men) measured in administrative data are driven by wage differences versus hours worked. Survey data lets some of these questions be answered, but with lower granularity and lower precision than if this could be studied in administrative data.
- Better understanding of productivity differences across regions of the country, occupations and industries. For example, incomes in finance are known to be high. But anecdotally at least, workers in finance also work long hours. The productivity differential between finance and other industries is therefore likely to be lower than the income differential, but currently the extent of this cannot be estimated.
- Recent popular and policy discussions about underemployment would be well-served by evidence on the extent to which people are working less than full-time hours, again with the granularity and precision that comes from using administrative data. Such data also have a longitudinal structure, so transitions in hours worked could also be better understood than in the short-panels that are often used in surveys.
- For policy enforcement, actual hours worked is needed alongside income to understand whether the National Living Wage/National Minimum Wage rules are being complied with.

I therefore **strongly support** the proposal to collect information on hours worked.

In terms of implementation, for those on hourly contracts, employers have this information and so it should not be a large ask for this information to be submitted as part of RTI. For those on fixed salaries, contractual hours could certainly be provided.

It would also be useful for employers to submit, for salaried workers, estimates of actual hours typically worked, which could be collected by an internal survey and be provided at the occupational level (rather than specific to the individual).

Dividends paid to shareholders in owner-managed businesses

Separation of dividends income into that received from an owner-managed business and that received from arms-length sources is very valuable, and not a particularly complex additional request for self-assessment filers. Some use cases include:

- For policy, in the COVID-19 pandemic some of the difficulties in providing support to the self-employed was around the lack of information on dividends paid to owner-managers separately from wider dividend information.
- For policy design and enforcement, it is also important to understand how much income received by owner-managers is taken in the form of dividends (which are taxed at a lower rate of income tax) versus as employment income. Current policy encourages owner-managers to take returns as dividends, and this is particularly common among those on high incomes, who therefore face lower effective tax rates than headline employment income tax rates would suggest.³ Better measurement of this would allow more effective policymaking, as well as allowing any anti-avoidance provisions to be more effectively enforced.

I **strongly support** the proposals to collect this information.

Specifically, I strongly support the proposals to make the company director and close company fields mandatory, to gather the value of dividends from a close company separately, and to gather the percentage shareholding. I also strongly support the proposal to link owner-managers to their (part-)owned company, which would complete the link between data sources, and provide data users with similar levels of information on owner-managed companies related to individuals that they currently have available for employees.

Implementation of this additional data collection is unlikely to be burdensome. Individuals filing self-assessment already need to collect together the information on all dividends received, and as owner-managers they already need to consult their company accounts, so separating out the dividends from the two types of sources should not be difficult.

The start and end dates of self-employment

This again seems a relatively trivial ask in terms of data collection, and would be of valuable both for policy and for understanding the nature of self-employment work.

On the policy side, as noted above a number of times, this would have been useful for the SEISS scheme in terms of targeting. If the data were collected even now, it would be useful in trying to analyse the effects of that scheme, albeit needing corrections for those individuals who have exited self-employment and who would therefore no longer need to report any more detail beyond what can already be observed (i.e. tax years in which they were engaged in some self-employment).

In terms of better understanding self-employment work, and how e.g. growth in self-employment incomes takes place, being able to accurately measure the time period over which the initial income was earned is important. Without this it is not possible to understand the growth between say the first and second year of self-employment unless additional assumptions are made on (the distribution of) start dates.

I therefore **strongly support** the proposals for collecting these data, and ensuring that completion of these tax boxes is mandatory.

³ Advani, A. and Summers, A. (2020) "How much tax do the rich really pay?". *CAGE Policy Brief 27*. URL: <https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/bn27.2020.pdf>.