Annex A. Transparency of SDC methods and parameters
(Post-meeting EAP paper for SDC for Census August 2021)

1. Introduction

Maintaining confidentiality of respondents to the 2021 UK Census is paramount for the Census Offices. Statistical disclosure control (SDC) methods are employed to achieve this and the basis of the 2021 methodology is described in papers from two previous visits to the Methodology External Assurance Panel, the latter of these being https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP125-Statistical-Disclosure-Control-SDC-for-2021-UK-Census.docx.

The purpose of this paper is to discuss (i) the default parameters to be used in 2021 live processing and (ii) the level to which these parameters are to be made public. The paper only considers the frequency tables from the standard variables and does not cover other outputs such as origin-destination tables and microdata. The methods were approved at UKCC in October 2020 and are harmonised across ONS, NRS and NISRA. The parameterisation and transparency described in this paper are specific to England and Wales. It is expected that there will be differences between the Offices in application of the methods, including the variables and classifications for targeting, the perturbation rates and perturbation matrix, and the rules that comprise the disclosure checks. Likewise, the considerations over transparency are likely to be different and therefore the conclusions may also differ.

2. The SDC methods and their roles

The methods to be employed in the protection of Census data have been approved by the UK Census Committee (UKCC) in October 2020, subsequent to the recommendation from the Methodology External Assurance Panel. Each of the three methods offers a complementary form of protection. These are:

Record Swapping: Every individual and household is assessed for uniqueness or rarity of a number of characteristics. Records that are unique or rare on one or more characteristics are highlighted as 'risky records' and all of these will be swapped. Similar households that match on some basic characteristics are sought from other areas to be used as swaps, in order to preserve data quality. These characteristics included household size, so that the numbers of persons and numbers of households in each area are preserved. Depending on availability of good matches, the numbers of different types of households are also preserved as much as possible. Households are only swapped within local authorities (LADs) or, in rare cases of households with very unusual characteristics, with matches in nearby authorities. To reduce the swap rate and maintain better data utility, other risky records are prioritised for use as 'matches' to be swapped with another risky household. However, it is not always possible to find a good match so some 'non-risky' records are used as matches, meaning that every household has a chance of being swapped.

Record swapping is also used for communal establishment data. In this case, individuals are swapped between communal establishments in different areas. The matching criteria are similar but with additions tailored to their position and the type of establishment they are in.

Cell Key Perturbation: The method is based on an algorithm which applies a pre-defined level of perturbation to cells in each table. The same perturbation is applied to every instance of that cell.

Firstly, a record key (a random number within a pre-defined range) is applied to every record in the microdata. This is done once and once only, so an individual's record key never changes. When frequency tables are constructed, each cell is a count of the number of respondents, and the cell key is calculated by summing their record keys. The combination of cell value and cell key is then read from a previously constructed look-up table (often termed as the ptable) to decide the amount of perturbation that should be used. Where the same cell (or same combination of respondents) appears in different tables, both instances will have the same cell value and cell key, and so receive the same perturbation. This also ensures that repeated requests of the same table, will have the same perturbation applied consistently.

For the 2021 Census, there is also some perturbation of cells where the counts are zero. A random number is assigned to each category of each variable and used to produce a random and uniformly distributed category cell key, in a very similar way to the cell key. This category cell key can be used to make a random selection of cells to perturb. Applying a category cell key in this way ensures zero cells are perturbed more consistently across tables the same way the cell key method ensures consistency when the same cell appears in different tables. As part of the zero perturbation, zero cells are chosen to be perturbed by, say +1 or +2. The same number of small cells are chosen based on category keys to be perturbed by -1 or -2. The zero perturbation method does not lead to any increase or decrease in overall population totals.

Note that the choice of which zeros to perturb is also based on whether the combination has appeared at a higher geography, in order to avoid perturbation of structural zeros.

The method offers protection against disclosure by differencing, and in instances where a number of tables can be constructed and could otherwise be linked together to reconstruct records from the microdata.

Disclosure Rules: The production of tables from a table builder could allow one to build an extremely large bank of tables. One could have access to large and extremely sparse tables which could allow identification of individuals and disclosure of information, notwithstanding the protection of 'risky records' (record swapping) and the protection against disclosure by differencing (cell key perturbation). Hence some limitations are advisable on the detail available.

Albeit that the resultant microdata would be post-swapping, the perception would (correctly) be that we would be providing personal information for every census respondent, though some might not be quite in the right geographic area. Certainly, it would be straightforward to identify individuals from knowing a few of their details and roughly where they live, and thus discover the remainder of their information.

The disclosure checks are the rules by which decisions can be made as to whether to allow release of outputs pertaining to specific combinations of variables. In previous censuses, the policy has always been to assess whether the release of tables is acceptable for all areas, and so every table that was passed was available for every area. That did mean that tables that might have been acceptable for some areas were not released because the corresponding table was not acceptable for other areas. This was particularly the case for some tables with ethnic group or country of birth, where such minority populations might be clustered in a small number of metropolitan city areas. Our aim for 2021 is to make tables available for those areas where the disclosure risk would be sufficiently low, rather than reject for all areas because some would incur higher risk. We refer to the two approaches as the '*blanket*' approach – where tables are produced for all areas, and the '*patchwork*' approach – where tables are produced for the subset of areas where the risk is low.

3. Parameter Setting

3.1 Record Swapping parameters

The parameters we are concerned with here are the 'target classifications', the breakdowns of the variables against which we assess uniqueness of individuals and households in the data. In the main, they reflect the most detailed classifications that are proposed for inclusion in the table builder.

The classifications were chosen on the basis of likely demand for the level of detail at low geographies. It means that if a user selects these classification at OA geography, we can be sure that there has been some protection applied to the resulting figures. Because we prioritise 'risky to risky' matches, the percentage of records being swapped is at a lower level to maintain more data utility. The classifications were employed for risk assessment in the SDC Processing Rehearsal in July 2021, using data from 2011 Census adapted to the structure of 2021 Census. The resulting overall swap rate for England and Wales for the Rehearsal was within the range of swap rates used for Delivery Groups in live processing for 2011 UK Census, agreed by UKCC at that time.

Of course, the swap rate is data driven and so in 2021 live running is not guaranteed to be similar to that in our Rehearsal. We will address specific needs as we see the live data and, indeed, we will have early sight of 'tuning data' coming through earlier methodological processes prior to the formal period of SDC processing. Some of the target classifications may thus be adjusted in the light of this – though they will remain constant across the country - and if there are changes in the details of table builder variables, that may arise from the current outputs consultation.

Within the Census data there will also be responses on sexual orientation and gender identity, and we are adopting a cautious approach to release of information relating to these. There is no intention to include these within the table builder at low geographies. The disclosure risk will be analysed at local authority level and there may be some information released for those areas where the disclosure risk is assessed to be negligible.

3.2 Cell key perturbation parameters

As with any form of disclosure control, the cell-key method has the potential to affect analysis and reduce the usefulness/utility of census data. As a light touch in addition to swapping, the perturbation rate (proportion of cells that receive non-zero noise) should be minimised to reduce the impact on utility. Conversely, the perturbation rate must be high enough to introduce sufficient uncertainty to disclosure by differencing. It must also be sufficiently high to prevent unpicking of the noise added. When noise is added to frequency tables, this causes inconsistencies in totals where the breakdowns of the same variables are different. These inconsistencies can be observed by users, and conversely indicate where perturbation has taken place.

The risk of differencing attacks is hard to quantify, especially with some elements of the dissemination system being unknown at this stage. The core protection of cell key perturbation is that estimates obtained by differencing have uncertainty introduced by the noise. This could be measured by the proportions of cells affected by noise with higher proportions reducing confidence in information gained by differencing attacks.

We assume here (see Table 2) that two tables are differenced from each other to give the new, smaller, cell values ('estimates from differencing') and that the perturbation is applied independently to each table. We also consider the estimates obtained by differencing to be equally protected if they have been changed by any value of noise (±1, ±2, ±3 etc.).

Table 2. Rates of perturbation required to affect a given proportion of estimates obtained by differencing.

To affect this proportion of cells in a differenced table:	Each table needs to receive this rate of perturbation:
5%	2.5%
10%	5.1%
16.67%	8.71%
20%	10.56%
25%	13.40%
33%	18.35%
50%	29.3%

Given that the differenced cells are affected by two independent sets of perturbation, the rates of perturbation required to affect given proportions of differenced cells are shown (using $P(A \cup B)=P(A)+P(B)-P(A \cap B)$, accounting for cells being affected by perturbation in both tables simultaneously).

3.2.1 Unpicking

To follow on from this, although a light-touch of perturbation will be applied, it must be sufficient to prevent unpicking. If too few perturbations are applied, it is possible to deduce what noise that has been added (and hence remove it). If the perturbation can be unpicked, it will not provide sufficient

protection against differencing. This issue is most pronounced in small tables containing very few cells, which will unavoidably contain few perturbations.

Unpicking is possible in certain circumstances when only a small number of perturbations are present in the table but gets more difficult as the number of perturbations increases.
Tables 3a and 3b show the expected number of perturbations in smaller output tables, for 2 variables and 3 variables (later referred to as 2-dimensional and 3-dimensional). Cells are highlighted based on how many perturbations are expected, with 2-3 perturbations highlighted in yellow, 3-5 perturbations in light green, and 5+ perturbations highlighted in blue. These calculations assume a single rate of perturbation for cells of any size, with every cell being equally likely to be perturbed. In this case, the expected number of perturbations is simply the number of cells multiplied by the probability that each cell is perturbed. Independent marginal totals are counted here. The 2*2 + T label signifies a table with two variables, each with two categories (four internal cells). When including marginal total cells, two cells for independent totals in variable one, two cells for independent totals of variable two, and one overall total, nine cells have the potential to be perturbed. 3*3 + T represents a table of two variables, each with three categories (nine cells). There are seven cells for marginal totals that could also be perturbed, so 16 cells are potentially perturbed.

Table 3a. Expected number of perturbations within a table with two variables, by number of categories and perturbation rate

	+ marginal totals + total						
Table size	2*2 + T	3*3 +T	4*4 +T	5*5 +T	6*6 +T	8*8 +T	10*10 +T
Number of cells, including totals	9	16	25	36	49	81	121
3.0%	0.3	0.5	0.8	1.1	1.5	2.4	3.6
4.0%	0.4	0.6	1.0	1.4	2.0	3.2	4.8
5.1%	0.5	0.8	1.3	1.8	2.5	4.1	6.2
8.0%	0.7	1.3	2.0	2.9	3.9	6.5	9.7
8.6%	0.8	1.4	2.2	3.1	4.2	7.0	10.4
10.0%	0.9	1.6	2.5	3.6	4.9	8.1	12.1
12.5%	1.1	2.0	3.1	4.5	6.1	10.1	15.1
15.0%	1.4	2.4	3.8	5.4	7.4	12.2	18.2
20.0%	1.8	3.2	5.0	7.2	9.8	16.2	24.2
25.0%	2.3	4.0	6.3	9.0	12.3	20.3	30.3
30.0%	2.7	4.8	7.5	10.8	14.7	24.3	36.3

Table 3b. Expected number of perturbations within a table with three variables, by number of categories and perturbation rate

	+ marginal totals + total			
Table size	2*3*4	2*4*4	3*4*4	4*4*4
Number of cells including totals	59	74	99	124
3.0%	1.8	2.2	3.0	3.7
4.0%	2.4	3.0	4.0	5.0
5.0%	3.0	3.7	5.0	6.2

8.0%	4.7	5.9	7.9	9.9
8.6%	5.1	6.4	8.5	10.7
10.5%	6.2	7.8	10.4	13.0
12.5%	7.4	9.3	12.4	15.5
15.0%	8.9	11.1	14.9	18.6
20.0%	11.8	14.8	19.8	24.8
25.0%	14.8	18.5	24.8	31.0
30.0%	17.7	22.2	29.7	37.2

The number of expected perturbations increases rapidly with table size. Unpicking is expected to be possible in smaller tables only. We consider that preventing unpicking in the smallest tables (2 categories by 2 categories) would require a very high perturbation rate, which would unduly affect the utility of the outputs.

Although they are most likely to be unpicked, in many ways, the smallest tables remain the lowest in terms of disclosure risk since they contain much less detail, and larger counts which are unlikely to be useful in differencing attacks.

The level to which a table should be unpickable should be low but need not be zero. An intruder will not know for certain that the unpicking is correct, and the inconsistencies with other tables will throw doubt on the certainty with which an intruder might feel they have succeeded with apparent unpicking. Any data that an intruder may have apparently ‘unpicked’ will still have protection from targeted record swapping.

We also note that even simple unpicking requires some effort, knowledge of how the perturbation works, and coding skills. The unpicking applied in testing required a sustained effort from an ONS SDC research officer, and the level of attacker sophistication and commitment of time and resources required to systematically implement these attacks is very high. Given that unpicked tables would be small and limited in detail for differencing attacks, and underlying data would still be swapped, there should be limited motivation to carry out such an unpicking attack.

3.3 Disclosure check parameters

To support record swapping and cell key perturbation methods, the detail within the resultant tables need to be restricted, both to prevent identification of unusual records, albeit in an area that is imprecise locally but likely still broadly correct (e.g. in the right local authority or nearby). We can set parameters that define this risk, and so these disclosure checks can be applied within the online table builder, rather than manually, which is a major benefit allowing outputs to be published much faster this time.

We will still be making some frequently requested tables available on a blanket basis through the Table Builder, if available in previous censuses, due to the greater protection of swapping and the cell key method in 2021. The rationale is that if they were sufficiently protected in 2011, they will be at least as well protected in 2021. However, each of these will still receive an individual assessment, looking at whether the 'worst' areas were sufficiently low risk.

The patchwork approach mostly covers combinations of variables and categories that were not published in 2011 and will now be provided subject to one or more rules. The parameters are intended to prevent tables that are too large, sparse, or give the perception of disclosure, taking into account the uncertainty introduced by imputation, swapping, and cell key perturbation. These parameters are subject to further testing including intruder testing on the 2021 data. Some of the rules discriminates between tables with/without variables that have been targeted through record swapping.

4. Transparency

Previous meetings with the Methodology External Assurance Panel have highlighted the question of transparency. Parameters of census disclosure control have historically been kept private, though we recognise there has been a recent trend towards greater discussion of transparency. This section seeks to outline which aspects of the disclosure methodology and parameters could be made public.

Disclosure control and other statistical processes have a responsibility to communicate their methods effectively and honestly. This will help facilitate better understanding and use of outputs, allow method selection to be challenged and improved upon, and build trust in the ONS and the statistics it produces. In the case of disclosure control, we also need to consider in which circumstances increased transparency may indirectly aid disclosure of personal information, or more likely, reduce the protection that the methods provide. Our aim is therefore to provide as much detail possible as possible to users, unless doing so is likely to reveal private information or significantly reduce the disclosure protection.

4.1 Transparency in 2011 UK Census

In 2011 targeted record swapping was applied to protect unique households deemed at risk of disclosure. Output checking was also carried out before a table was published, with those deemed too disclosive being subject to redesign.

The general swapping method was communicated to users long in advance of Census, though the swap rate and targeting criteria have not been made public. It is also worth noting that the swap rate was dependent on the level of protection from imputation of non-responses, such that some areas had a lower swap rate than others. The matching criteria used were not described in detail, though it was made clear that preserving population totals of local areas, and totals of age by sex were the priority. It was agreed that as part of the output checking, tables would contain a minimum (but unpublished) proportion of real attribute disclosure (AD) cases that imputation and swapping have protected, and of apparent AD cases (i.e. in the swapped data) that are not real. The table re-design process was at times slow and frustrating for users, though as a result of the slow manual checking, and multi-staged re-design process, rather than the disclosure measures applied or any lack of transparency around them.

4.2 Transparency in 2021 UK Census

In 2021, as well as targeted record swapping, cell-key perturbation and automated ‘disclosure checks’ form part of the protection. Aspects of these three methods have been listed in Table 5, with our intention to make this information public, highlighted green, or keep it private, highlighted in red.

Several features of the SDC methodology are intended to be kept private. These are selected on the assessment that publication of these features may increase the likelihood of disclosure, or significantly reduce the protection applied such that it outweighs the benefits of transparency (of this feature). On this note, it is worth considering that transparency of some features would be more beneficial to users and the public than others.

Table 5. Proposed Transparency for SDC approaches for 2021 UK Census

Topic	Feature	Status	Visible	Comments
Swapping	Method	Public	Yes - Already publicised	
Swapping	Risk criteria	Private	No	Describing the criteria on which households are targeted for swapping implicitly describes scenarios where particular households will not be targeted for swapping. If it was known with high confidence that particular households were not swapped, this would significantly reduce the uncertainty associated with disclosures on households of these types.
Swapping	Matching criteria	Public	No	Broad detail: "Match similar households", "on things like household size, age, sex"
Swapping	Rate	Private	Not to any great precision	Knowledge of the exact swapping rate may increase confidence in apparent attribute disclosures. For example, if it were known that x% of households were swapped, apparent attribute disclosures *could* be viewed by an intruder as having (100-x)% chance of being correct. However, we would be willing to consider giving a broad range of the swap rate that would give researchers some reassurance. The swap rate is data-driven so we cannot give an estimate as yet but – based on comparisons between 2011 and 2021 methods (used on similar data) the swap rate is likely to be not too different.
Perturbation	Cell-key method	Public	Yes - Already publicised	
Perturbation	'Overall' rate	Public	Yes	Can be calculated using prevalence of inconsistencies
Perturbation	Rate (small counts)	Private	No	Small counts represent the majority of risk in a table, either through identity or attribute disclosures, and

				we intend to perturb small cells more frequently than larger cells. Unlike the 'overall' rate of perturbation which can be accurately estimated, the rate applied to small cells cannot easily be discovered by a user. Since the small cells represent a higher risk, and small overall proportion of table cells, we intend to keep the perturbation rate private for those small cell counts. It will be stated that small cells are perturbed at a higher rate than other cells. This will maintain a higher level of uncertainty in differencing using small cells.
Perturbation	Rate (medium counts)	Private	Somewhat	May be possible to get tables with only medium counts and look for inconsistencies
Perturbation	Rate (large counts)	Private	Somewhat	Possible to get tables with only large counts and look for inconsistencies
Perturbation	Distribution of noise	Public	No	Laplace would be assumed (good statistical properties)
Perturbation	Perturbation bounds (+-3, +- 5)	Private	No	Knowledge of the maximum size of perturbation would reduce the disclosure protection. For example the maximum possible difference caused by perturbation being (say) +-3, can be helpful for unpicking perturbation, including through "bounding attacks" - comparing upper and lower potential bounds for true cell values to isolate perturbation.
Perturbation zeros	General method (catkeys)	Public	Yes - Already publicised	

Disclosure checks	Rule A (Marginal minimum)	Public	Somewhat	Being public may encourage doubt over marginal 1s, 2s, rather than assuming they are real
Disclosure checks	Parameter A (marginal minimum)	Private	Exceptions to the rule should be confusing	The disclosure checks will disallow tables with small counts in any category (0 is allowed). An exemption

				is given for variables that have been targeted for swapping, since the households in those small counts will be likely have been swapped. If the parameter and logic for allowing exemptions is publicised, it could be determined the criteria and variables that were used for targeting of swapping (and which households are unlikely to be swapped).
Disclosure checks	Rule C (Marginal dominance)	Public	Yes (if rule known)	
Disclosure checks	Parameter C	Public	Yes (if rule known)	
Disclosure checks	Rule E (0s)	Public	Yes	
Disclosure checks	Parameter E	Public	Yes	
Disclosure checks	Rule H (ADs)	Public	Yes	
Disclosure checks	Parameter H	Private	No	The attribute disclosure limit is closely related to the marginal minimum rule, in that in some cases this only applies to variables that have been targeted for swapping. In a similar way knowledge of this parameter may reveal information about the targeting.
Disclosure checks	Rule J (1s + 0s)	Public	Yes (if rule known)	
Disclosure checks	Parameter J	Public	Yes (if rule known)	
Disclosure checks	Maximum number of cells	Public	Yes	
Disclosure checks	Maximum number of variables	Public	Yes	
Disclosure checks	Reason for failure	Private	No	When data for certain areas in a table are not released due to failure of the disclosure checks, the reason for failure will not be given. This would amount to describing the data deemed unsafe for release and could indirectly help reveal disclosive information.

Note that the transparency outlined here is only applicable to 2021 UK Census. It does not apply to previous censuses where some of the protection was assumed to be from the hidden nature of some of the details and parameters. Previous censuses, for example 2011, had different methods of protection and the risk assessments considered different ranges of tables and outputs.

4.2.1 Visible or Discoverable Features

Disclosure Rules: We note that several features could be discovered by close analysis of the tables available through the dissemination system, or that have simply already been made public in communication or consultation with users. In Table 5, the column 'visible' notes whether a feature has already been publicised or could be learned by users. In such cases, it is better to be transparent up-front, rather than give an intruder the impression that the protection has been 'broken' or was insufficient.

A large variety of tables will be available via the flexible dissemination system. By observing the range of data allowed through the table builder, in some cases it is possible to infer what is not allowed through the disclosure checks. The most obvious example is the proportion of cells that are zero. By requesting a large number of tables, it could be seen that the maximum proportion of zeros must be close to some threshold (x%), as no tables with rates above this threshold are seen by a user. The same logic can be applied to the majority of the disclosure checks, if it is known which checks are being applied. We thus propose to provide details of the rules used, though no details of which rule(s) any specific table failed. Knowledge of the rules may even discourage users from attempting to generate sparse tables.

'Overall' Perturbation rate: The application of cell key perturbation can cause inconsistencies between totals appearing in different tables. As a result, inconsistencies between totals indicate that perturbation has taken place and, conversely, a lack of inconsistency indicates that (almost certainly) no perturbation has taken place. By observing how often inconsistencies appear between tables with a known number of cells, it is possible to accurately estimate the perturbation rate. We intend to perturb smaller cells more frequently, and 'larger' cells less frequently, to preserve utility where possible and reflect the relative associated risk. A result of this is that tables containing proportionately more small cells will have a higher rate of perturbation than tables containing proportionately more large cells. It is suggested an expected 'overall' rate of perturbation be publicised, with the note the rate of perturbation of individual tables will depend on the size of the cells they contain.

5. Summary and Conclusion

This paper has reprised the 2021 UK Census disclosure control methods that have previously been approved by UK Census Committee. We have also outlined the considerations and proposals for the default parameters in using those methods, and whether or not to make these details available publicly. We have restricted details within this paper where we assess that they should not be released publicly, but the Methodology External Assurance Panel were invited to consider whether to discuss them specifically during the Panel meeting.

The Panel were asked for their comments on this paper, to support the approaches being taken and to make recommendations for the SDC team to take to UK Census Committee. These have been incorporated into this later version.