Noise Infusion at BEA

Authors John Bockrath and Dan Yorgason, U.S. Bureau of Economic Analysis

Contact John.Bockrath@bea.gov

Date April 2026

Abstract This paper provides an overview of BEA's plans to transition from cell suppression to noise infusion (multiplicatively perturbing reported survey values) as its primary method of statistical disclosure limitation (SDL). BEA intends to begin using noise infusion for SDL in summer 2026 for statistics based on its surveys of new foreign direct investment in the United States and transitioning in the coming years to using noise infusion for statistical products based on other surveys. In this paper we discuss both the planned transition for the new foreign direct investment statistics and some broader issues and opportunities resulting from our exploration of noise infusion. Based on our testing, transitioning the new foreign direct investment statistics to noise infusion will offer enhanced utility to data users in the form of publishing substantially more data while still providing confidentiality to BEA's survey respondents.

Keywords Disclosure limitation, noise infusion, official statistics

JEL Code C82, F21, F23

Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

1. Introduction

The U.S. Bureau of Economic Analysis (BEA) has commenced a process to switch its primary statistical disclosure limitation (SDL) method from cell suppression (not publishing cells that reveal too much information) to "EZS" noise infusion (multiplicatively perturbing reported survey values).[1] At BEA, SDL is mainly applied to statistics that are based on surveys conducted by BEA itself.[2] These surveys provide the source data for many of BEA's international economic accounts statistics, some of which feed into gross domestic product (GDP) and other National Economic Accounts statistics.

The purpose of SDL at BEA is to limit the degree to which information about individual survey respondents—primarily firms—can be inferred from BEA statistics. In other words, BEA seeks to provide confidentiality for survey respondents and their data. BEA statistics are intended to provide information on activities, trends, and patterns that apply to groups of firms, not those that apply to any particular (identifiable) firm. The use of SDL for BEA survey data is necessitated by legislation that authorizes the surveys. Most notably, the International Investment and Trade in Services Survey Act requires that BEA cannot provide information "…in a manner that the person who furnished the information can be specifically identified…" and that information collected under the act "…may be used only… for analytical or statistical purposes with in the United States Government." On a practical level, BEA's use of SDL provides survey respondents confidence that the information they provide won't be used to their disadvantage by competitors, regulation, enforcement, or policymaking. This confidence ultimately results in better statistics by limiting any hesitation on the part of firms to respond accurately (or at all) to the surveys.

The switch from cell suppression to EZS noise infusion represents a switch from one "legacy" SDL method to another. Both approaches have been used to provide confidentiality for decades, and neither provides any formal privacy guarantee (e.g., differential privacy) of the sort associated with more modern SDL approaches. BEA has elected to continue to use legacy methods in part because the more modern approaches have proven challenging to apply to skewed data of the sort collected on BEA's firm-level surveys.

BEA intends to switch to EZS noise infusion on a rolling basis. The first survey BEA intends to apply transition to noise infusion is the BE–13, Survey of New Foreign Direct Investment (NFDI) in the United States, which feeds into annual statistics on NFDI. The upcoming June 10, 2026 release

1. Other terms, including disclosure avoidance, are used to express the same idea as SDL. We use SDL throughout this paper to emphasize that although the disclosure risk of confidential survey data can be largely attenuated it cannot be entirely eliminated—short of not publishing any statistics at all.

2. BEA also applies SDL to a portion of its Regional Economic Accounts statistics that is based on Quarterly Census of Employment and Wages data from the U.S. Bureau of Labor Statistics. The changes in BEA's SDL approach discussed in this paper do not apply to such statistics.

of preliminary 2025 and revised 2022–2024 NFDI statistics will represent the first use by BEA of EZS noise infusion in an official statistical product.[3] This paper focuses in large part on the application of EZS noise infusion to the NFDI statistics—including the benefits and costs to data users, some of the details of the specific approach adopted, and the likely effects on the statistics—based on our in-depth internal evaluation of the impact of the switch. In broad terms, we find that "protecting" the NFDI statistics with EZS noise infusion can be expected to improve data utility without sacrificing current protection levels—i.e., those protection levels provided by cell suppression.

A second focus of the paper is the effects of EZS noise infusion in a more general sense.[4] These effects vary by the composition of contributions to the statistic in question and by the specific distributions selected to generate noise. A third focus of the paper is the opportunities and challenges for BEA from expanding the use of noise infusion to other BEA surveys.

1.1 SDL Environment at BEA

To serve as a basis for many of its International Economic Accounts statistics, BEA surveys firms that are engaged internationally, either through trade in services or direct investment. Separate surveys are used for separate types of activities within these two broad categories of engagement. Firms provide a variety of information on the surveys, much of which is considered confidential by the firms themselves and all of which is presumed by BEA to be confidential. Indeed, BEA considers even the fact of filing on a given survey to be confidential.

The information provided by firms is the source for many of the statistics BEA publishes on the international engagement of the U.S. economy. SDL at BEA can be viewed as the process of going from confidential microdata to tables of statistics that are "safe" enough to publish. The major source of SDL is aggregation; BEA does not publish the microdata it collects, only aggregates of that microdata. For instance, in the trade in services statistics the reported values of every firm exporting service type X to country Y would be summed and only that total would be published as a cell in a table. The reported values of these firms would also typically contribute to other more aggregated cells in the table—for instance, total services exports in world region Z. For several reasons, though, relying solely on aggregation is rarely a viable approach to SDL.

To assist aggregation in protecting the confidentiality of data provided by surveyed firms, BEA has historically used cell suppression. The primary shortcoming of using cell suppression is that certain table cells are not published. Which cells are unpublished tends to vary from period to period

3. Statistics for 2022 and 2023 will be revised only to incorporate cost updates—the replacement of expected costs with actual costs—on establishments or expansions initiated in those years but lasting multiple years.

4. This focus is provided, in part, in lieu of a more detailed discussion of the impact of noise infusion on the NFDI statistics; such a discussion could compromise confidentiality by providing too much information about the specifics of BEA's noise infusion scheme.

as the composition of the cell changes, making it difficult for data users to know which statistics will be available. Moreover, as cells become increasingly disaggregated (as BEA endeavors to provide data users with more granular statistics), a larger percentage of table cells are suppressed, which limits the utility of the statistics.

The ultimate cause of suppressed cells is aggregations that are dominated by large contributors. This occurs at various rates in each of BEA's survey-based tables, but it is particularly pronounced in the NFDI statistics. These statistics cover investment activities that are lumpy—that is, activities that don't occur every period but are large in scale when they do occur. Furthermore, even when a reasonably large collection of such activities is represented in an aggregate, the underlying microdata are often skewed so that one or two firms dominate the aggregate.

EZS noise infusion is an alternative SDL method in which no cells are suppressed but every microdata contribution to a cell is "perturbed" by a positive or negative factor. BEA is adopting this approach to SDL for the NFDI data because it expects that full publication of table cells will substantially increase the utility of the tables. This increase in utility is expected to more than offset the downside of EZS noise infusion—the fact that every published cell value is slightly perturbed (by virtue of being the sum of perturbed contributions), more so for cells dominated by large contributions and less so for other cells. BEA also expects that switching to EZS noise infusion for SDL will improve the utility of other survey-based statistics, but it is switching first for the NFDI statistics for the following reasons:

1. The pronounced lumpiness of the NFDI data leads to many suppressions.
2. The lumpy nature of the NFDI microdata also implies that any data user attempting to impute the values of suppressed cells in these statistics (which is far from a trivial undertaking) will likely make relatively poor imputations compared to imputations that could be made for suppressed cells in other BEA statistics. This means that the utility of the NFDI statistics can potentially be raised considerably by using a method other than cell suppression.
3. Other considerations discussed in more detail later in the paper, such as the limited interaction of NFDI statistics with other BEA statistics and the few variable relations across data series and over time, make this method especially straightforward to implement with the NFDI statistics.

Aside from improving the utility of its existing statistics, BEA anticipates the switch to EZS noise infusion will facilitate the introduction over time of more detailed versions of some of these statistics. The elimination of cell suppressions loosens one of the major constraints to increasing the level of data granularity BEA can provide, meaning that data quality and relevance will be the primary considerations associated with such decisions once noise infusion has been adopted.

1.2 Organization of the Paper

Following this introduction, the paper is organized as follows: First, Section 2.1 provides an overview of BEA's current SDL method, cell suppression. Under cell suppression, confidential data responses are protected by suppressing—that is, withdrawing from publication—any cell value that provides "too much" information about one or more of the individual survey responses underlying the cell value.[5] Cell values that are withdrawn are replaced with a text value, usually a "(D)", in the statistical publication. A significant complication, and irritant, with cell suppression is that often cells that are not themselves "vulnerable"[6] must also be suppressed to ensure that the values of suppressed cells that actually are vulnerable cannot be inferred (or "uncovered") via arithmetic manipulation.

Next, Section 2.2 provides an overview of BEA's new SDL method for the NFDI data, EZS noise infusion. This overview consists of two parts: a description of the mechanics of this method and simulation-based general results on the effects of the method on statistics for a variety of configurations of noise and underlying reported data. Under EZS noise infusion, all cells are published; none are suppressed. However, instead of publishing cell values that exactly reflect the underlying survey responses, "perturbed" cell values reflect unperturbed cell values plus the application of random noise. The noise is applied at the record level (i.e., to the values in each survey form), not the cell level, thereby providing protection to each record. The cell-level perturbation is thus the sum of record-level perturbations. The main tradeoff in moving from cell suppression to EZS noise infusion is being able to publish more cells at the cost of distortion of cell values generally. EZS noise infusion seeks to limit the size of distortions, particularly for cells that aren't vulnerable, but the random nature of the noise infusion means that two otherwise identical cells will generally experience distortions of different sizes.

Section 3 examines the specific impact of EZS noise infusion on the NFDI statistics. In this section, we describe the NFDI tables, note current levels of suppression, assess options for how to apply EZS noise infusion, and present some simulation-based outcomes for how past releases of NFDI statistics would have been affected had noise infusion been used instead of cell suppression.[7]

5. The method is called cell suppression, rather than just suppression, in recognition of the fact that statistics are mostly published as table cells. The tables these cells are a part are often two dimensional, but they can be one dimensional or three (or more) dimensional.

6. Under cell suppression, a cell is viewed is vulnerable if publishing it would provide too much information about an individual survey response.

7. Note that the application of noise infusion for 2025 preliminary and 2022–2024 revised statistics will not affect the previously published statistics they supersede—namely, 2024 preliminary statistics and earlier vintages of 2022 and 2023 statistics for some tables—or statistics for reference periods prior to 2022. Cell suppression will continue to be the SDL method protecting those statistics.

Section 4 summarizes potential applications of noise infusion to other BEA products. Finally, Section 5 summarizes the information presented in this paper and the reasons BEA expects that noise infusion will enhance the utility of its statistics.

2. Approaches to SDL

2.1 Cell Suppression

BEA currently uses cell suppression as its primary SDL method for the NFDI tables and other published statistics that are based on BEA-administered surveys. Under this method, if publishing a table cell would lead to excessive disclosure of the survey data of individual respondents, that cell's value is removed from the table and replaced with a suppression indicator, usually a value of "(D)".

There are two types of suppressions—primary suppressions and complementary suppressions—though published tables do not distinguish between the two. Primary suppressions are cells that are vulnerable, meaning that the cells themselves include contributions that would be at risk if the cell were to be published. Complementary suppressions are cells being used to "cover" primary suppressions—that is, to prevent the value of a primary suppression being unwound through arithmetic manipulation of related table cells. Cells used as complementary suppressions are typically not vulnerable themselves, though a primary suppression can also serve as a complementary suppression if it covers another primary suppression.

BEA uses the p-percent rule to identify cells requiring primary suppression.[8] Under this rule, BEA calculates the internal (to the cell) coverage available for the largest contribution and compares it to that contribution.[9] To calculate the available coverage, BEA focuses on a scenario in which the second-largest contributor in a cell attempts to estimate the largest contribution. Under the p-percent method's assumptions about prior knowledge,[10] the second-largest contributor is the most informed potential "attacker" of the largest contribution, as it knows and can account for its own contribution.[11] Therefore, coverage for the largest contribution must come from all other contributions to the cell—that is, from every contribution but the two largest.

8. See pages 61–62 in Statistical Policy Working Paper 22 (Second version, 2005) Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology for an explication of the p-percent rule. Note that our description is based on assumption that the coalition size equals one. Working Paper 22 also discusses some of the complexities of complementary suppression.

9. Under BEA's assumptions, if the cell's largest contribution is adequately covered, the cell's other contributions will also be adequately covered.

10. It is assumed that potential attackers can initially estimate each contribution (other than their own) to within 100 percent (plus or minus) of the real contribution. This means that the prior "range of uncertainty" about the contribution spans the interval from 0 to 200 percent of the actual value. A contributor is assumed to know the value of its own contribution with certainty.

11. An "attacker" refers to a data user whose focus is using the data product to infer confidential information about one or more survey respondents.

Adequate coverage is available if these other contributions are at least p percent as large as the largest contribution,

$$p\% = \frac{p}{100} \leq \frac{X - x_1 - x_2}{x_1} \qquad (1)$$

where X, x_1, x_2 are, respectively, the total cell value, the largest contribution, and the second-largest contribution.[12] Under the p-percent method's assumptions about prior knowledge, satisfying this inequality results in a range of uncertainty about the actual value of the largest contribution that, on either side of the contribution, is at least p percent of the contribution. That is, it ensures that an attacker knows, at most, that the largest contribution is in a range that starts at $(100-p)$ percent of its actual value and ends at $(100+p)$ percent.

Identifying primary suppressions is thus a straightforward, mechanical exercise of identifying any cells that fail Equation 1.

Once primary suppressions are identified, BEA then determines which cells to use as complementary suppressions. This process is more complicated, in part because several different cells can often be used as complementary suppressions. When selecting complementary suppressions, BEA aspires to maximize the utility of the remaining table.[13] In the simplest case, if the primary suppression is a subaggregate of a published aggregate, another subaggregate (of sufficient size) must be used as a complementary suppression. In the more typical situation, a primary suppression is a subaggregate of two or more different aggregates in two or more different dimensions. Then, at least one complementary suppression must be used in each dimension. Moreover, these complementary suppressions must themselves be covered in each dimension to adequately protect the primary suppression.[14]

Table 1 below provides an example of primary suppression for a table with two dimensions. In the table, a lone primary suppression with a value of 200 is shown with purple shading and a potential set of complementary suppressions is shown with yellow shading. Note that there are many other possible sets of valid complementary suppressions.

In this example, the primary suppression is covered in the row dimension 1 with the cell with value 220 and covered in the column dimension with the cell with value 110. The final complementary suppression (330) covers the first two complementary suppressions; together with the primary, the set of complementary suppressions forms a closed "box."[15]

12. Later in the paper, "concentration ratio" will be used to denote the term on the right-hand side of the inequality.

13. In this context, utility is an ill-defined term that can mean different things to different data users.

14. Here the terminology is slightly misleading. The final complementary suppression still actually provides protection for the primary suppression, but it is in neither the same row nor the same column as the primary.

15. The suppression pattern shown is a canonical, but simple, case. In practice, patterns actually used often include more circuitous "routes"; more than one complementary suppression being used to cover the primary in a given row or column; additional suppressions arising from a third (or higher) dimension, often only apparent when considering multiple tables as a system; the suppression of marginal totals; and/or other primary suppressions being used as complementary suppressions; and other things.

Table 1: Example Table for Selection of Suppressions

Row	Column			
	Total	C1	C2	C3
Total	3,210	1,040	1,440	730
R1	990	500	430	60
R2	680	110	240	330
R3	680	230	330	120
R4	860	200	440	220

When published, the values of the suppressed cells in the table would be replaced with “(D)”, as shown in Table 2:

Table 2: Suppressed Example Table

Row	Column			
	Total	C1	C2	C3
Total	3,210	1,040	1,440	730
R1	990	500	430	60
R2	680	(D)	240	(D)
R3	680	230	330	120
R4	860	(D)	440	(D)

To qualify as a complementary suppression (by itself), a cell must be large enough to provide, in combination with the existing coverage from contributions in the cell with the primary suppression, p-percent coverage to the largest contribution in the primary suppression; that is:

$$p\% \leq \frac{X - x_1 - x_2 + X_c}{x_1} \qquad (2)$$

where X_C is the total value of the complementary suppression.[16] If a given cell does not provide sufficient coverage, BEA could select additional complementary suppressions (in the same dimension) until this coverage is obtained.

As previously noted, selecting the appropriate complementary suppressions—so that all primary suppressions receive sufficient coverage while retaining as much utility in the table for data users as possible—is a challenging task. Different statistical agencies use different approaches. BEA identifies

16. This applies to the simple case in which the largest contributor of the primary suppression is not also a contributor to the potential complementary suppression.

complementary suppressions using a mixture of automated and manual approaches that are resource intensive and not immune from error. Moving away from using suppression for SDL would increase BEA's efficiency in publishing statistics and remove a potential source of error.

Moving away from suppression would also permit a full set of statistics to be published each release—one without holes. In addition, it would result in complete time series for individual cells (and columns, rows, tables, etc.).[17]

2.2 EZS Noise Infusion

BEA intends to transition away from cell suppression to EZS noise infusion as its primary SDL method, starting with the June 2026 release of the NFDI statistics and proceeding over the next few years to cover all other statistics it publishes that are based on surveys it conducts. These include statistics based on BEA's surveys of international trade in services, direct investment transactions and positions, and activities of multinational enterprises.[18]

This section discusses both the mechanics of EZS noise infusion and the relationship between noise infusion and the characteristics of table cells (i.e., how the average impact of noise infusion varies for different types of published cells).

2.2.1 Mechanics of EZS Noise Infusion

EZS noise infusion was first proposed in 1996 at the U.S. Census Bureau by Timothy Evans, Laura Zayatz, and John Slanta.[19] It is one of several SDL methods currently in widespread use at the Census Bureau.[20] EZS noise infusion, like cell suppression, is a legacy SDL method. Legacy SDL methods have the virtue of a long track record in use by statistical agencies, and they are typically relatively easy for data users to understand. Legacy SDL methods, though, are at a disadvantage in certain respects vis-à-vis more modern SDL methods. In particular, unlike legacy methods, the

17. Suppose that a cell has a 70 percent probability of not being suppressed in each statistical release. If the probability of suppression is independent from release to release (or even just weakly correlated), the probability of a time series of values from that cell being complete (i.e., containing no suppressed values) can be much lower than 70 percent. Thus, analyses that require complete time series are particularly disadvantaged when suppression is used.

18. EZS noise infusion was used in 2025 by BEA for experimental direct investment statistics on Direct Investment equity by ultimate host economy. Its use for the NFDI statistics will be the first use for official statistics.

19. The original working paper, "Using Noise for Disclosure Limitation of Establishment Tabular Data," is available on the Census Bureau's website. Also see Timothy Evans, Laura Zayatz, John Slanta (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data," Journal of Official Statistics, 14(4), 537–551; Federal Committee on Statistical Methodology, "EZS-Noise Addition," In Data Protection Toolkit: Protecting Data While Increasing Access; and Laura McKenna and Matthew Haubach (2019), "Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau," Research and Methodology Directorate, U.S. Census Bureau, Washington, DC.

20. This SDL method is used for statistical products such as the Quarterly Financial Report, Commodity Flow Survey, County Business Patterns, Integrated Longitudinal Database, Job-to-Job Flows, Quarterly Workforce Indicators, Nonemployer Statistics, OnTheMap, and Survey of Business Owners.

modern methods provide provable privacy guarantees (e.g., differential privacy) that are independent of the prior knowledge of the data user. Modern methods can also typically be more fully specified publicly than many of the legacy methods.[21]

Under EZS noise infusion, each record underlying a cell value is assigned a noise factor. The noise factors are constructed as the product of two random draws: a non-negative noise magnitude, measured as a percentage, and a noise sign, either positive one or negative one with equal probability:

$$NF_r = M_r S_r \qquad (3)$$

where *NF* represents the noise factor, *M* represents the magnitude draw, *S* represents the sign draw, and *r* indexes records. For instance, a noise magnitude draw of 8 percent combined with a negative sign draw yields a noise factor of –8 percent.[22]

Next, one plus the noise factor is multiplied by the reported record value to yield a perturbed record:[23]

$$y_r^p = y_r\,(1 + NF_r) \qquad (4)$$

where *y* represents the record value and *p* denotes a perturbed value. Applying the –8 percent noise factor to a reported record of 500 gives a perturbed record of 460.

The sum of the perturbed record values for a cell equals the perturbed cell value, which differs from the unperturbed cell value by the sum of the record-level perturbations,

$$y_c^p = \sum_{r \epsilon c} y_r^p = \sum_{r \epsilon c} y_r(1 + NF_r) = y_c + \sum y_r NF_r \qquad (5)$$

where *c* indexes cells.

Note that the application of EZS noise has several key characteristics:

- Noise is applied at the record level, not at the cell, or some other, level.
- Noise is applied to every record, not just some subset of records.

21. Modern methods are not always well understood by data users, a significant limitation. Moreover, modern methods still confront the tradeoff between privacy and data utility, which is made explicit by use of a privacy budget. For certain types of data, including skewed data such as company-level data, maintaining satisfactory data utility only comes with relatively weak privacy guarantees.

22. Note that our terminology differs slightly from that used in the original EZS literature. Specifically, a noise factor in the original literature is equal to our noise factor plus one. Our –8 percent noise factor would therefore be a noise factor of 0.92 in the original literature. The use of different terminology has no substantive consequence on how the method is applied aside from the fact the noise factor is directly multiplied by the reported record value in the original literature instead of first adding one.

23. Note that BEA does not use sampling weights in constructing its statistics. Cell values are the sum of reported values and imputed values (for non-responding or non-sampled). Imputed values are treated the same as reported values for the purposes of noise infusion.

- Each published cell represents an aggregation of perturbed records.
- If a record contributes to more than one cell, the same perturbed record value is used for each cell. The record is not assigned different noise factors for different cells.
- Noise factors are zero in expectation. Negative and positive factors offset on average. Consequently, the expected perturbation (distortion) at the cell level is zero, though most perturbations will be nonzero.
- Roughly, the more contributions in a cell, the more likely the cell perturbation will be close to zero (as a percentage).
- SDL is provided in an expectations sense. Not every vulnerable cell is materially perturbed, but vulnerable cells tend to be more perturbed than nonvulnerable cells.
- In practice, it is convenient to use a distribution for noise magnitude draws that is bounded from above at one or lower and whose lower bound is nonzero.[24] For instance, the Census Bureau uses a split triangular distribution in noise infusing its Quarterly Financial Report statistics.[25]
    - If noise factor distributions are bounded from above at X percent, relative (absolute) distortions of cells composed exclusively of non-negative records are also bounded at X percent. There is no bound on the relative distortion of cells that include both positive and negative records.
- Magnitude draws can be independent of each other or not, as can sign draws. For instance, all records from a given survey respondent might be assigned the same sign and/or the same magnitude. Magnitude draws and sign draws are independent of each other.

2.2.2 Impacts of Noise Infusion by Cell Features

The distortion of a given cell resulting from using noise infusion is random, but its expectation depends on several elements including the distribution from which cell-level perturbations (noise factors) are drawn and the characteristics of the cell itself, such as how many contributions the cell includes and how those contributions are distributed. Some of these elements (e.g., the noise factor distribution) are under the control of the entity applying noise infusion, but others are not (e.g. the relative distribution of contributions within a given cell). In this section, we use simulated data to examine the effects of various elements on the expected magnitude of distortions. First, we examine distortions for a set of 45 specific combinations of noise factor distribution and

24. Bounding from above at one or below ensures that a perturbed record with a negative sign draw has the same sign as the unperturbed record. Bounding from below at a positive value ensures that the perturbed record does not equal the unperturbed record.

25. Using our terminology, this distribution encompasses both the magnitude and sign draws; see the Census Bureau's FAQ on QFR disclosure avoidance methodology for details on this distribution.

cell characteristics.[26] These are only illustrative, because the number of possible combinations is limitless, but they are intended to span a wide range of possible outcomes for cells with small numbers of contributions. Then, we select one specific noise factor distribution to demonstrate the impact of different combinations of cell vulnerability and different numbers of cell contributions. Throughout the analysis, we utilize a simple version of EZS noise infusion in which every contribution is independent of every other contribution's noise factor. The noise factor distributions (and their ranges) used for the examples in this section are only illustrative. To promote integrity of the noise infusion system and confidentiality of survey responses, BEA will not reveal the parameters BEA will use for noise infusion. The parameters it will use do not necessarily match the ones used for these examples.

Distortions for select cells and noise factor distributions

Comparing bounded symmetric distributions of noise factors, those distributions with more weight close to zero will generally see smaller distortions. For set bounds, this means that the mean absolute percent distortions from noise infusion should be smaller for distributions such as triangular or PERT distributions than for more diffuse distributions such as the uniform distribution.[27] It also means that, all else equal, distortions will be smaller for distributions that are "split" further away from zero.

Perhaps more important, though, than the specifics of the distribution of noise factors in determining distortion magnitudes is the distribution of contributions within a cell. Cells that have large numbers of contributions, or cells that have a fair number of relatively equal-size contributions, will see smaller distortions than cells dominated by one or two contributions.

In this section, we illustrate these relationships using a simple simulation exercise for hypothetical noise factor and cell distributions. For this exercise, we use three types of distributions—PERT, triangular, and uniform—and three different lower bounds for the absolute percentage noise factor: 0, 3, and 6.[28] For each distribution, the (percentage) upper bound of the absolute noise factor is set as the lower bound plus 12.[29] Figure 1 illustrates probability distribution functions (PDFs) of noise factor for a lower bound of 3 percent.

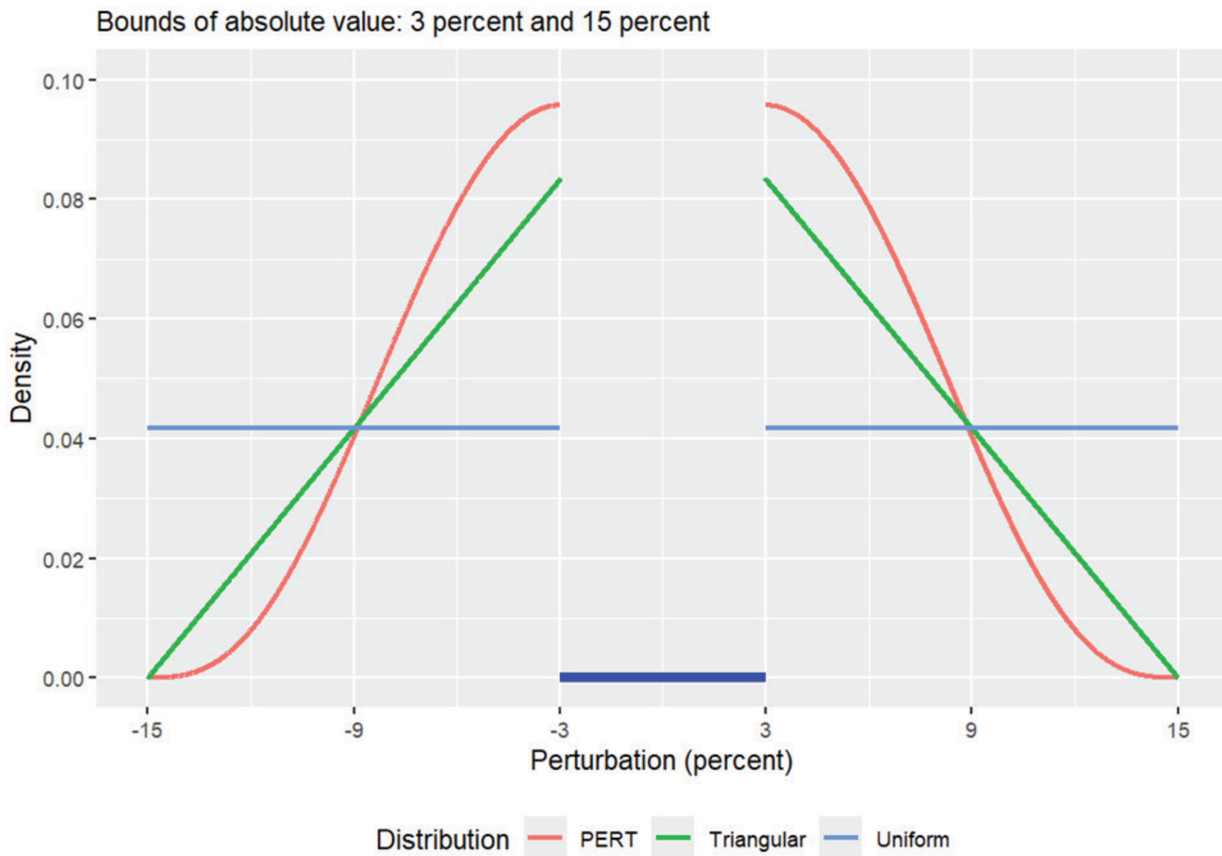
26. In line with the discussion in the previous section, under EZS noise infusion the noise factor distribution reflects the distribution of the product of a magnitude draw and a sign draw. One way to think about the magnitude draw is as the absolute value of the draw from a distribution—call it distribution D—that is symmetric about zero. Then, in somewhat circular fashion, the (combined) noise factor reflects a draw from distribution D.

27. The PERT distribution is more bell shaped than the triangular distribution, but it can be described using the same parameters (min, max, mode).

28. The lower bound of the distribution of absolute values corresponds to the distance the center of the underlying distribution is split away from zero. For example, a lower bound of 3 percent means that all noise factors have a magnitude of at least 3 percent, though some are positive and some are negative.

29. The PDF of a uniform distribution with a lower bound of 3 (percent) would consist of two nonzero segments, the first ranging from –15 to –3 and the second ranging from 3 to 15. The height of the segments over each range would be constant and equal to 1/24.

Figure 1: PDFs of Uniform, Triangular, and PERT Noise Factors

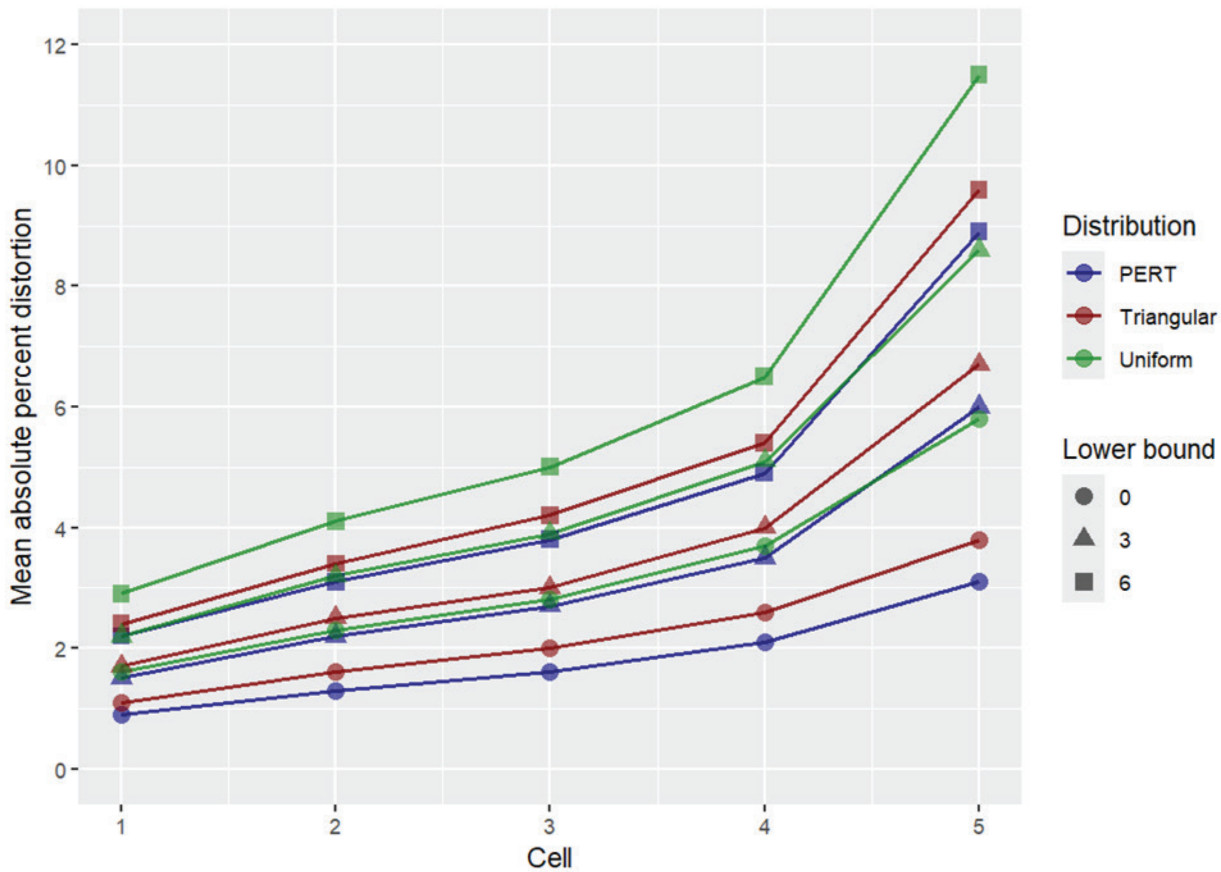


The impact of noise infusion depends on both the distribution of the noise factors and the distribution of contributions within the cell. In addition to the 9 combinations of noise factor distributions (3 distribution types, 3 split parameters), we investigate 5 different cell contribution distributions, generating 45 distinct combinations in total. Specifically, we show results for each of the following five cell contribution distributions (in combination with the nine noise factor distributions):

- Cell 1: 12 contributions of 200 each.
- Cell 2: contributions of 600, 500, 400, 300, 200, 100, 100, 100, and 100.
- Cell 3: contributions of 600, 600, 600, and 600.
- Cell 4: contributions of 1,100, 1,100, 100, and 100.
- Cell 5: contributions of 2,300 and 100.

Figure 2 shows the mean absolute percent distortion for each combination of the 45 combinations of cell, distribution type, and split parameters.

Figure 2: Mean Absolute Percent Distortion by Cells, Distributions, and Lower Bounds



Generally, distortions are smaller for smaller lower bounds of the absolute percent noise factor and for more diffuse cell contributions. Other things equal, the PERT noise factor distributions result in smaller distortions than triangular distributions, and both of these result in smaller distortions than uniform distributions.

Table 3 provides additional summary statistics of the impact of noise infusion on each combination, sorted by mean absolute percent distortion.[30] Generally, mean distortions are slightly larger than median distortions, suggesting a bit of right skew in the distributions of distortions. Standard deviations of distortions tend to rise with means, but not uniformly; in particular, standard deviations of distortions are relatively small when the PERT distribution is used as a basis for noise factors.

Table 3: Summary of Percent Distortions for 45 Cases

Distribution	Lower bound of abs. % noise factor	Cell	Absolute Percent Distortion			Mean Absolute Percent Noise Factor
			Mean	Median	SD	
PERT	**0**	**1**	0.9	0.8	0.7	3.3
Triangular	**0**	**1**	1.1	1.0	0.9	4.0
PERT	**0**	**2**	1.3	1.1	1.0	3.3
PERT	**3**	**1**	1.5	1.3	1.2	6.3
Triangular	**0**	**2**	1.6	1.4	1.2	4.0
Uniform	**0**	**1**	1.6	1.4	1.2	6.0
PERT	**0**	**3**	1.6	1.4	1.2	3.3
Triangular	**3**	**1**	1.7	1.5	1.3	7.0
Triangular	**0**	**3**	2.0	1.7	1.5	4.0
PERT	**0**	**4**	2.1	1.8	1.5	3.3
PERT	**3**	**2**	2.2	1.9	1.6	6.3
PERT	**6**	**1**	2.2	1.9	1.6	9.3
Uniform	**3**	**1**	2.2	1.9	1.7	9.0
Uniform	**0**	**2**	2.3	2.0	1.6	6.0
Triangular	**6**	**1**	2.4	2.1	1.8	10.0
Triangular	**3**	**2**	2.5	2.1	1.8	7.0
Triangular	**0**	**4**	2.6	2.2	1.9	4.0
PERT	**3**	**3**	2.7	2.4	2.0	6.3

table continues on next page

30. For the three absolute percent distortions columns of Table 3, two of the corresponding *signed* summary percent distortion measures are trivial; the mean and median signed percent distortions are, by construction, expected to be zero under the EZS approach.

The standard deviation of signed percent distortions can be calculated as the square root of the following sum: the variance (i.e., squared standard deviation) of the absolute percent distortion plus the square of the mean absolute percent distortion. For example, in the uniform/0/3 case, the standard deviation of signed percent distortions approximately equals $\sqrt{2.04^2 + 2.80^2} = 3.46$

Table 3: Summary of Percent Distortions for 45 Cases—*continued*

Distribution	Lower bound of abs. % noise factor	Cell	Absolute Percent Distortion			Mean Absolute Percent Noise Factor
			Mean	Median	SD	
Uniform	**0**	**3**	2.8	2.4	2.0	6.0
Uniform	**6**	**1**	2.9	2.5	2.2	12.0
Triangular	**3**	**3**	3.0	2.7	2.2	7.0
PERT	**6**	**2**	3.1	2.8	2.2	9.3
PERT	**0**	**5**	3.1	2.8	2.2	3.3
Uniform	**3**	**2**	3.2	2.8	2.3	9.0
Triangular	**6**	**2**	3.4	3.0	2.4	10.0
PERT	**3**	**4**	3.5	3.3	2.6	6.3
Uniform	**0**	**4**	3.7	3.2	2.6	6.0
PERT	**6**	**3**	3.8	3.8	2.9	9.3
Triangular	**0**	**5**	3.8	3.4	2.7	4.0
Uniform	**3**	**3**	3.9	3.5	2.8	9.0
Triangular	**3**	**4**	4.0	3.6	2.9	7.0
Uniform	**6**	**2**	4.1	3.6	2.9	12.0
Triangular	**6**	**3**	4.2	4.0	3.1	10.0
PERT	**6**	**4**	4.9	4.8	3.9	9.3
Uniform	**6**	**3**	5.0	4.8	3.7	12.0
Uniform	**3**	**4**	5.1	4.4	3.7	9.0
Triangular	**6**	**4**	5.4	5.1	4.1	10.0
Uniform	**0**	**5**	5.8	5.8	3.3	6.0
PERT	**3**	**5**	6.0	5.7	2.2	6.3
Uniform	**6**	**4**	6.5	5.7	4.9	12.0
Triangular	**3**	**5**	6.7	6.2	2.7	7.0
Uniform	**3**	**5**	8.6	8.6	3.3	9.0
PERT	**6**	**5**	8.9	8.5	2.2	9.3
Triangular	**6**	**5**	9.6	9.1	2.7	10.0
Uniform	**6**	**5**	11.5	11.5	3.4	12.0

Although larger distortion magnitudes are associated with larger magnitudes of noise factors, this relationship is quite rough in this exercise. For instance, the largest average noise factor magnitude, 12 percent, is associated with both a 2.9 percent absolute distortion and an 11.5 percent absolute distortion, as well as several other distortions between these two values. Rather than the distortion being driven primarily by the noise factor magnitude, the most determinate factor in Table 3 is the type (composition) of cell. This level of distortion varies considerably across the five types of cells examined, but as a fraction of the mean absolute noise factor, the distortion level is consistent enough within each type that there is no overlap for the specifications considered.

Specifically, the ratio between mean absolute percent distortion and mean absolute noise factor is:

- Cell 1: 24–28 percent.
- Cell 2: 34–40 percent.
- Cell 3: 41–49 percent.
- Cell 4: 53–64 percent.
- Cell 5: 96–96 percent.

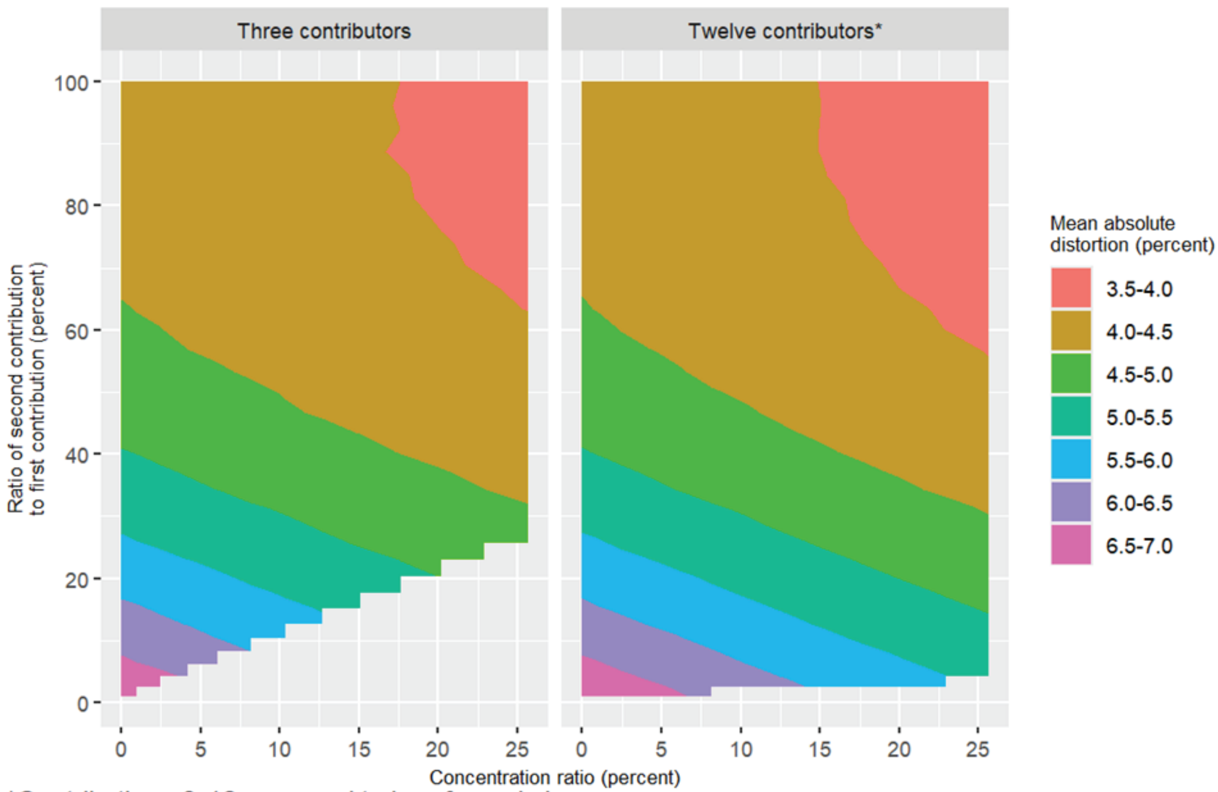
For example, for Cell 1 with a uniform distribution and a lower bound of 6, the ratio is 24.1 percent (2.89 percent / 12.00 percent). As cells grow more concentrated (moving toward Cell 5), the average noise infusion distortion moves steadily closer to the midpoint of the noise factor range.

Vulnerability and distortions

In assessing distortions, it is often useful to distinguish between distortions to vulnerable and nonvulnerable cells. In light of BEA's extant use of p-percent-based suppression as the primary SDL technique, it is natural to define a vulnerable cell as one that would be a primary suppression under BEA's current application of the p-percent rule. In moving from suppression to noise infusion, it is generally desirable that nonvulnerable cells are distorted as little as possible. For vulnerable cells, the calculus is more complicated. On one hand, too little distortion can reveal information that should remain confidential. On the other hand, too much distortion mitigates the benefits of moving from suppression to noise infusion.

Vulnerable cells (and, indeed, nonvulnerable cells), can vary in terms of vulnerability. Some vulnerable cells need more protection than others to achieve a given level of safety. Under EZS noise infusion, it is often the case that more vulnerable cells receive larger average (absolute percent) distortions than less vulnerable cells. This is not always true, however. Sometimes a more vulnerable cell, as measured for the application of the p-percent rule, will have a smaller expected percent distortion than a less vulnerable cell.[31] This is illustrated, using simulated data and an arbitrarily selected approach to noise infusion, for a specific example in Figure 3.

Figure 3: Mean Absolute Distortion by Concentration Ratio and Second-largest Contribution



*Contributions 3-12 assumed to be of equal size
Note: Based on triangular distribution of absolute noise factors with lower bound of 3 percent and upper bound of 15 percent.

This example uses a triangular distribution with a 3 percent lower bound and a 15 percent upper bound for absolute noise factors. Two cases are considered: cells with 3 contributors and cells with 12 contributors. For either case, the "concentration ratio" is measured as the sum of all but the two largest contributions divided by the largest contribution (expressed as a percentage).[32] This is the percentage that is compared with the "p" threshold in applying the p-percent rule. For a cell with 3 contributors, a concentration ratio of 15 percent implies that if the largest contribution is 600, the

31. In the analysis that follows, vulnerability is measured as a matter of degree rather than a binary question. Under the p-percent rule, by contrast, a cell is either vulnerable (and requires a primary suppression) or not vulnerable (and does not require a primary suppression). Part of the reason for examining the degree of vulnerability here is that BEA does not disclose the threshold value used in its application of the p-percent rule. Another reason is that actual vulnerability is probably best viewed as a matter of degree. The selection of any particular p-percent threshold is necessarily arbitrary, and some disclosure risk always arises from publishing any survey-based statistic.

32. The "concentration ratio" terminology is borrowed from the Census Bureau's Federal Statistical Research Data Center Disclosure Avoidance Methods: A Handbook for Researchers. The terminology is somewhat misleading because more concentrated cells generally have lower ratios. As defined here, the concentration ratio is more a measure of dispersion than concentration.

smallest contribution is 90. For a cell with 12 contributors, the smallest 10 contributions sum to 90. In this example, each of these 10 contributions are assumed to be of equal size, so each equals 9.

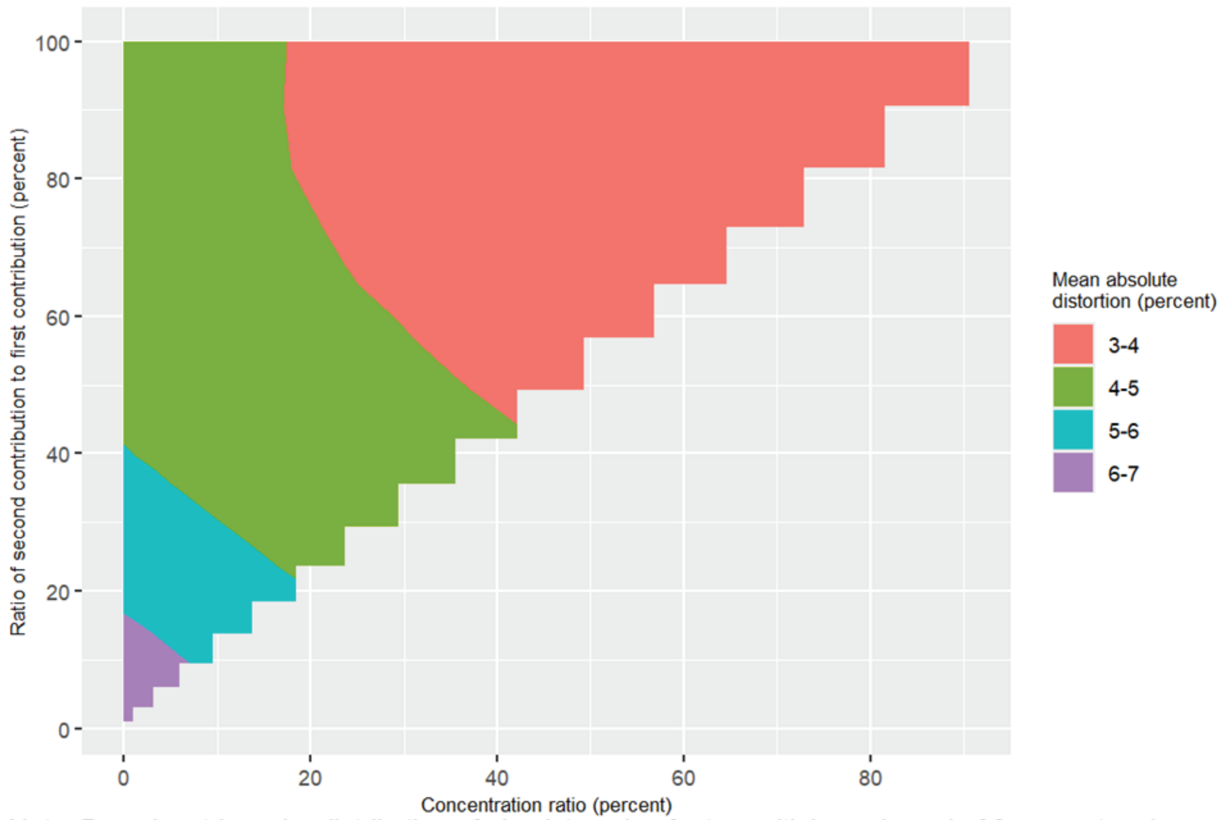
Under the p-percent rule, vulnerability is assessed solely by the concentration ratio—the smaller the concentration ratio, the more vulnerable the cell. The size (or lack thereof) of the second-largest contribution is immaterial.[33] Under noise infusion, by contrast, both the concentration ratio and the size of the second-largest contribution (relative to the largest contribution) affect the average magnitude of distortions. As shown in Figure 3, distortions are largest when both of these ratios are small.[34] Holding the size of the second-largest contribution constant, a lower concentration ratio results in higher mean percent distortions. However, a lower concentration ratio can result in smaller distortions if the second-largest contribution is large enough. For example, a 10 percent concentration ratio produces an average distortion of about 4.1 percent when the second-largest contribution is 85 percent of the largest. This is smaller than the roughly 5 percent average distortion that results when a 15 percent concentration ratio is paired with a second-largest contribution that is only 25 percent of the largest.

Figure 4 and Figure 5 provide the same analysis over a wider range of concentration ratios. With larger concentration ratios, average distortions are smaller, but the basic patterns are similar to those seen in Figure 3. For the three-contribution case, the largest possible concentration ratio is 100 percent, corresponding to a situation where the smallest contribution is the same size as the largest contribution because all three contributions are equal in size.

33. Nonetheless, under the p-percent rule, the size of the second-largest contribution in a primary suppression typically affects how much protection, relative to the size of the cell, is needed from complementary suppressions.

34. In the figure, uncolored portions of the grid reflect infeasible combinations of the x-axis and y-axis variables. Namely, the second-largest contribution cannot be smaller than any of the contributions other than the largest contribution. Relatedly, the y-axis stops at 100 percent because the second-largest contribution cannot be larger than the largest contribution. The x-axis in this example stops at about 25 percent—not because of any feasibility constraint, but to focus the analysis on relatively vulnerable cells. The grid spacing underlying the graphics is not fine enough to give the impression of continuity; nonetheless the main results are clearly apparent.

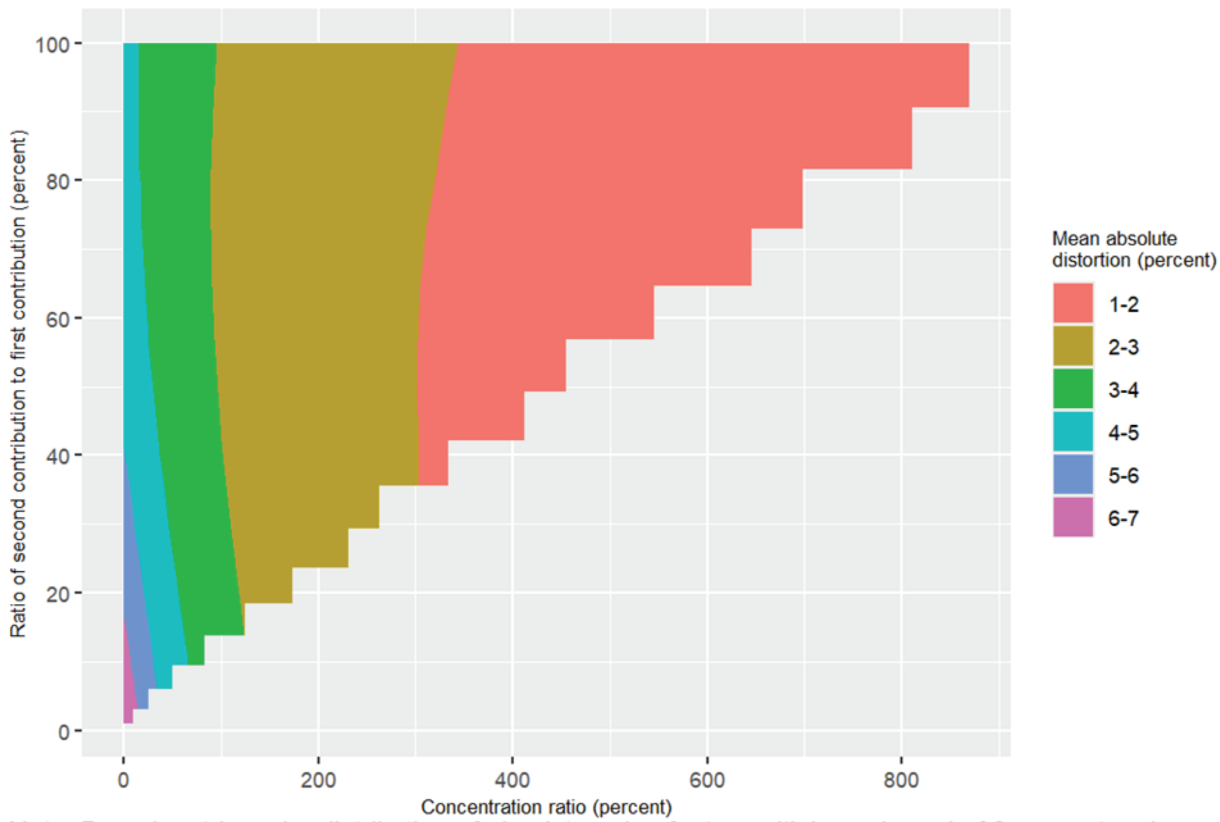
Figure 4: Mean Absolute Distortion by Concentration Ratio and Second-Largest Contribution (Three Contributors)



Note: Based on triangular distribution of absolute noise factors with lower bound of 3 percent and upper bound of 15 percent.

For the 12-contribution case, the largest possible concentration ratio is 1,000 percent, corresponding to a situation where the smallest 10 contributions are each as large as the largest contribution because all twelve contributions are equal in size.

Figure 5: Mean Absolute Distortion by Concentration Ratio and Second-largest Contribution (Twelve Contributors)



Note: Based on triangular distribution of absolute noise factors with lower bound of 3 percent and upper bound of 15 percent.

Cell size (number of contributions) and distortions

As suggested by the results in Table 3, cells with more contributors generally experience smaller average relative distortions as positive and negative contribution-level perturbations partly or fully offset each other. All else equal, the more contributions, the more likely it is that such offsetting will occur.

Figure 6 illustrates this result, based on simulated data and an arbitrarily selected approach to noise infusion, for three situations where the number of contributors in a cell is increasing. Each of the three panels includes a baseline case where contributions equal in size to existing contributions are added as the cell increases in size. The first panel also includes two cases where each subsequent contribution is larger, by a fixed multiple, than the largest existing contribution as the cell increases in size. The second panel includes two cases where contributions are added in triplets that replicate the triplets already in the cell. The distribution of contribution in these triplets differs across the two cases. The third panel includes cases where the size of each contribution is a function of the order in which it is added to the cell.

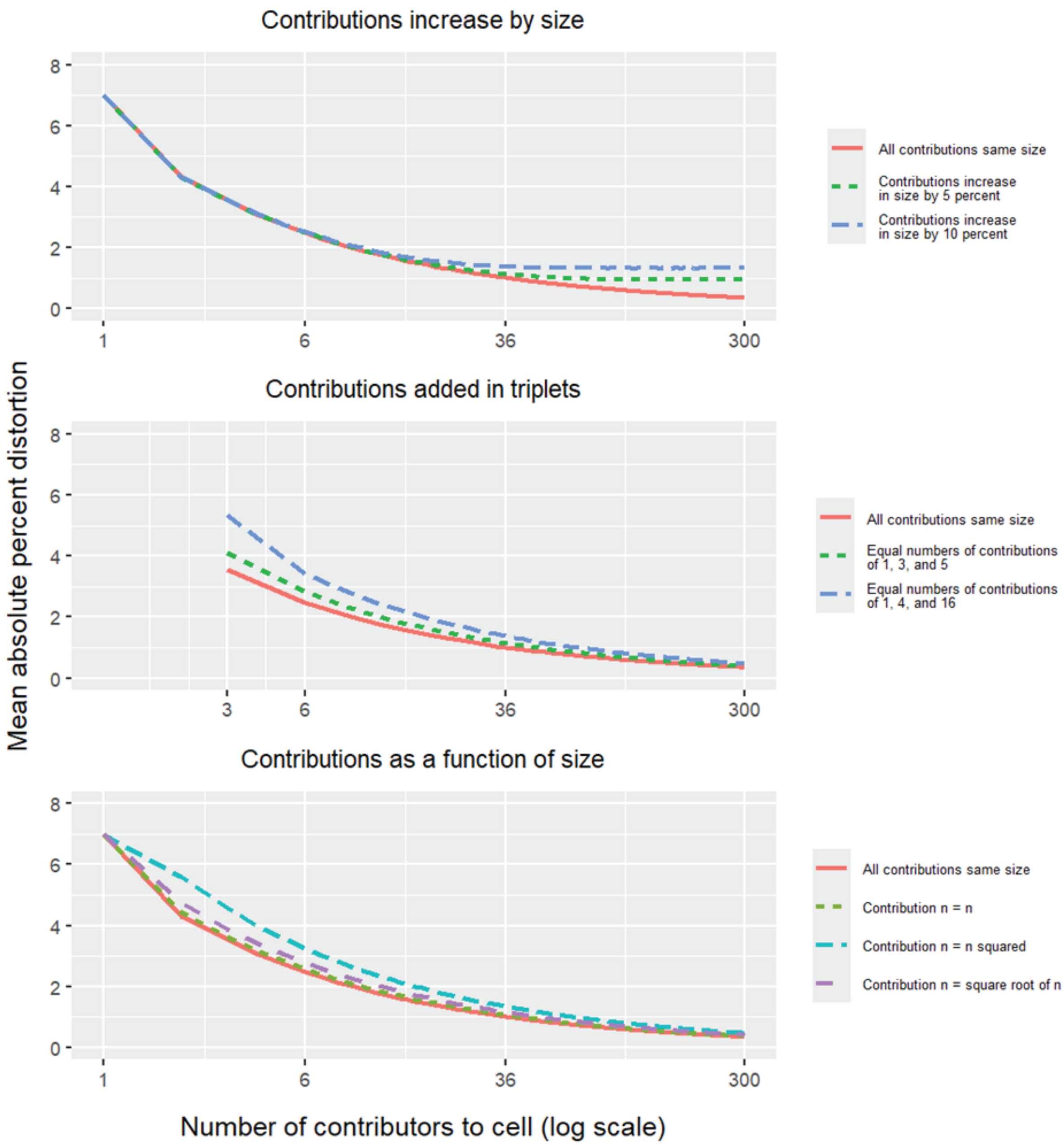
Table 4 gives contributions for cells of size 3, 6, and 9 for each of the cases, with different colors illustrating the portion of each set of contributions covered by each cell size, illustrating the pattern by which the composition of contributions within a cell having a given contribution count is determined. For consistency, the first contribution to a cell is assumed to be one in every case.

Table 4: Composition of Cells for Cases in Figure 6

Case	Contributions
All contributions same size	**1, 1, 1,** **1, 1, 1,** **1, 1, 1**
Contributions increase in size by 5 percent	**1, 1.05, 1.1,** **1.16, 1.22, 1.28,** **1.34, 1.41, 1.48**
Contributions increase in size by 10 percent	**1, 1.1, 1.21,** **1.33, 1.46, 1.61,** **1.77, 1.95, 2.14**
Equal numbers of contributions of 1, 3, and 5	**1, 3, 5,** **1, 3, 5,** **1, 3, 5**
Equal numbers of contributions of 1, 4, and 16	**1, 4, 16,** **1, 4, 16,** **1, 4, 16**
Contribution n = square root of n	**1, 1.41, 1.73,** **2, 2.24, 2.45,** **2.65, 2.83, 3**
Contribution n = n	**1, 2, 3,** **4, 5, 6,** **7, 8, 9**
Contribution n = n squared	**1, 4, 9,** **16, 25, 36,** **49, 64, 81**

*The three-contribution case is represented by **green** entries.*
*The six-contribution case is represented by **green** and **blue** entries.*
*The nine-contribution case is represented by **green**, **blue**, and **red** entries.*

Figure 6: Mean Absolute Percent Distortions by Number of Cell Contributions



Note: Based on triangular distribution of absolute noise factors with lower bound of 3 percent and upper bound of 15 percent.
See Table 4 for information on the composition of cells for the various cases.

In the first panel, the three cases yield similar average distortion magnitudes for small cells but diverging average distortions for larger cells. For the two cases where contributions increase in size, there is not much difference between large and small contributions for smaller cells, but the difference grows geometrically and eventually becomes very large. For these cells, the mean absolute percent distortion decreases toward a lower limit that is higher than zero. By contrast, mean absolute distortions of the cell with contributions of equal size head to zero.[35] The convergence to nonzero limits of the two cells with contributions that increase geometrically in size is related to the fact that, for these cells, the ratio of the largest contribution to the sum of the contributions converges to a nonzero limit as cell size gets large.[36] This implies that the perturbation of the largest contribution never gets fully offset, on average, by the perturbations of the remaining contributions as the contribution count increases.

In the second panel, mean absolute percent distortions are initially larger for cells with more diffuse triplets. As the cell size increases, however, they appear to converge to similar values as for the cell with equal-size contributions. A similar pattern emerges in the third panel (except when the number of contributions is one); in particular, even when the value of each contribution is the square of its "number," the mean absolute percent distortion appears to converge to the same value as for the cell with equal-size contributions. Unlike the cases shown in the first panel, each value in the sequence of cell members increases slower than geometrically (though more quickly than arithmetically), and the ratio of the largest contribution to the sum of contributions converges to zero.

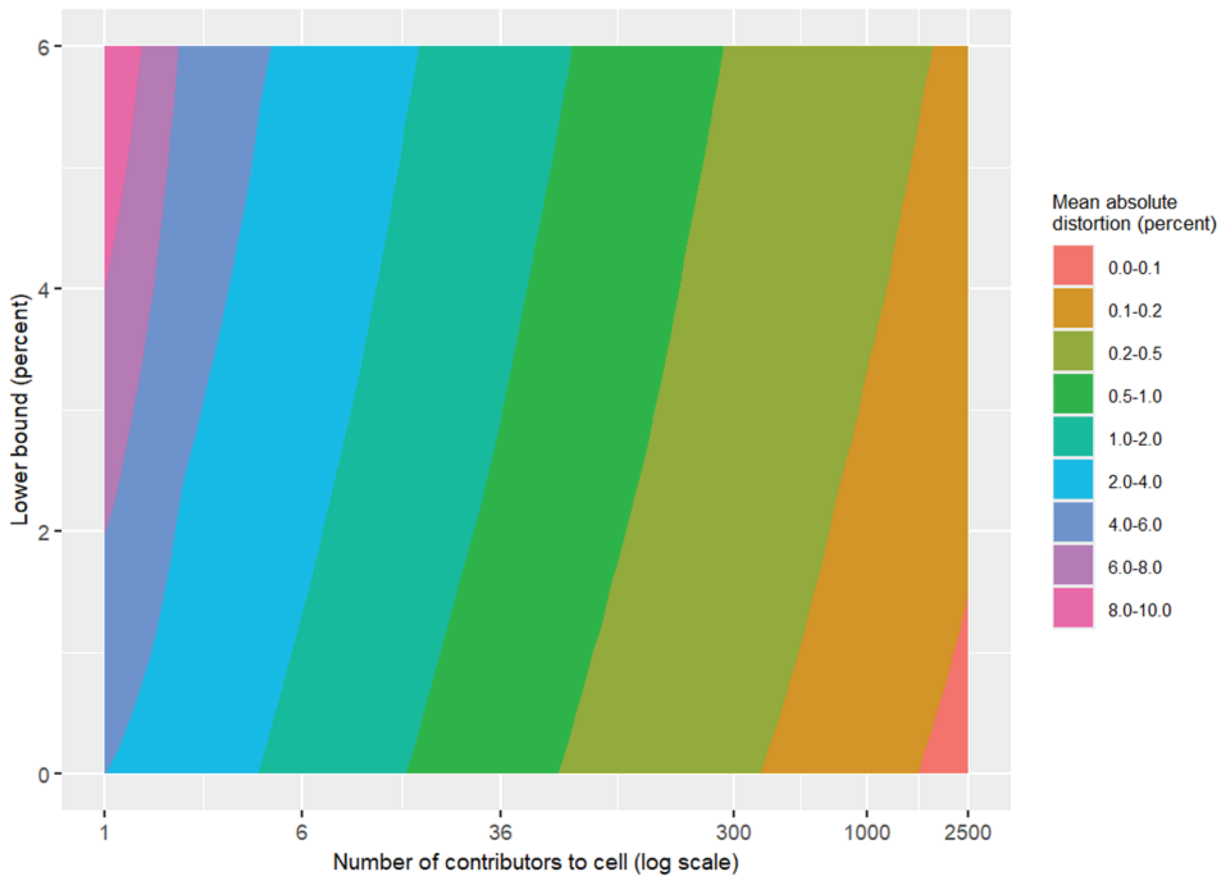
The results in Figure 6 cover the joint impact of cell size and cell composition for a triangular noise factor distribution with a range of 12 percentage points on either side of the mode and split away from zero on either side by 3 percentage points. The next two figures show the impact of varying the size of, respectively, the split away from zero or the range of the noise factor distribution. Both figures assume all contributions to the cell are equal.

35. For a cell size of 300 contributors, the mean absolute percent distortion is approximately 0.35 percent. This decreases to approximately 0.19 percent, 0.11 percent, and 0.06 percent, respectively, when the cell size increases to 1,000, 3,000, and 10,000.

36. For the cell where the largest contribution increases in size by 10 percent, the limit of this ratio as cell size approaches infinity is 1/11. For the cell where the largest contribution increases in size by 5 percent, the limit is 1/21.

Figure 7 holds the range of the distribution of absolute noise factors at 12 percentage points but allows the split away from 0 to vary from 0 (no split) to 6 percentage points. For small cells, the size of the split (i.e., the lower bound of the distribution of absolute noise factors) has considerable impact on the average magnitude of distortions. For larger cells, the impact is considerably diminished.[37]

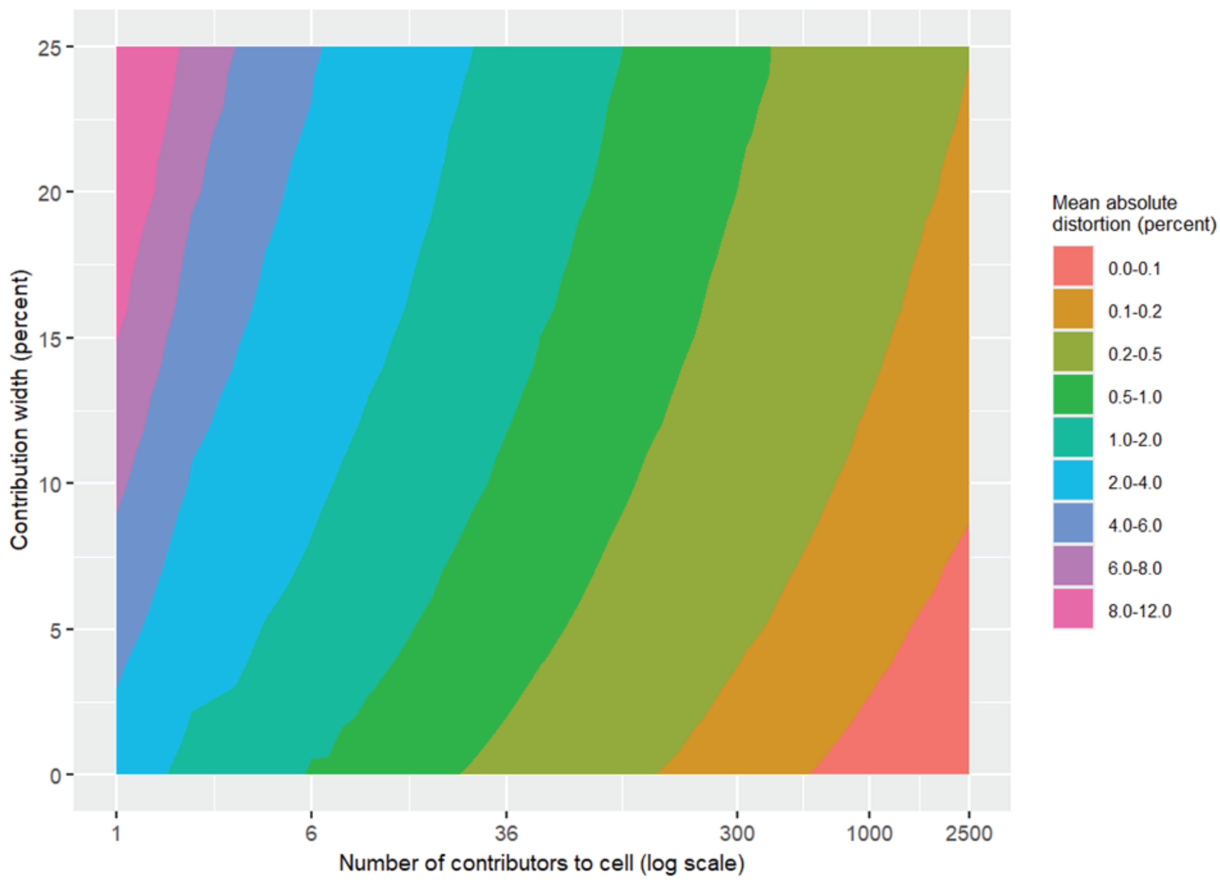
Figure 7: Absolute Distortion by Number of Cell Contributors and Lower Bound of Absolute Noise Factor



37. Note the that the colored regions represent distortion ranges of different widths. For ranges above 2 percent, each region's range width is 2 percentage points. However, below 2 percent each region represents progressively smaller distortion ranges.

Figure 8 holds the split away from zero at 3 percentage points but allows the range from the lower to upper bound of the distribution of absolute noise factors to vary from 0 to 25 percent. Allowing the range to vary has similar effects on magnitude of distortions as allowing the split to vary. In particular, the marginal impacts of varying the range are much larger for small cells than for large cells.

Figure 8: Absolute Distortion by Number of Cell Contributors and Range of Noise Factor Distribution



3. Noise Infusion in the NFDI Tables

3.1 NFDI Data and Tables

The NFDI tables are based on data from the BE–13 survey. A U.S. firm is "required to report if (1) it is acquired or established by a foreign person or entity resulting in the creation of a foreign direct investment relationship or (2) it is an existing U.S. affiliate of a foreign parent and establishes a new U.S. legal entity, expands its U.S. operations, or acquires a U.S. business enterprise." As part of these surveys the U.S. firm discloses information on the transaction (i.e., the cost of the investment), information on the U.S. and foreign firm(s) (each firm's industry, information on the foreign firm's country of ownership, etc.), and certain operating information about the U.S. business.

BEA uses these survey data to create tables showing two types of information: investment expenditures and employment.[38] Some of these tables include information on multiple survey items (for example, tables on employment separately show first-year employment and planned future employment). In this paper we refer to these individual survey items as "series" and the two broad categories as "series types."

Employment data are published in three tables that show data by investment type (whether the investment is establishing a new firm, expanding an existing foreign affiliate, or acquiring an existing U.S. firm) and either U.S. state, the U.S. affiliate's industry, or the country of the ultimate beneficial owner of the U.S. affiliate.

Investment expenditures are published for these three combinations of dimensions as well as for dimensions that are not used for employment series. This includes supplemental tables on "greenfield" investments, investments that involve the creation of new business operations (either via establishment or expansion), rather than the acquisition of an existing U.S. business.[39]

Table 5 summarizes the NFDI tables, showing the series included within each series type and the available combinations of dimensions for each series.

38. This description of the NFDI tables covers the table structure used for the July 2025 NFDI release. In previous periods the NFDI tables also included tables with information on firms' sales, net income, and balance sheets. However, in July 2025 BEA stopped producing these series.

39. The NFDI tables include one additional table showing the *number* of investments in each investment size class. This table is not affected by the change to BEA's approach to SDL.

Table 5 : NFDI Series Types, Series, and Dimensions

Series Type	Series	Dimension Combinations
Investment expenditures	First-year expenditures Greenfield planned expenditures	Industry of Affiliate by Type of Investment
		Country of UBO by Type of Investment
		State by Type of Investment
		Industry of UBO by Type of Investment
		Country of UBO (all countries)
		Industry of Affiliate (all industries)
		Country of UBO by Industry of Affiliate (first-year expenditures only)
		Country of Foreign Parent (first-year expenditures only)
		Type of Expenditure by Type of Investment (Greenfield planned expenditures only)
	Initial planned expenditures Greenfield actual expenditures	Type of Investment by Year of Expenditure
		Industry of Affiliate by Year of Expenditure
		Country of UBO by Year of Expenditure
		State by Year of Expenditure
		Year of Expenditure by Year of Investment
Employment	Current employment Projected future employment	Industry of Affiliate by Type of Investment
		Country of UBO by Type of Investment
		State by Type of Investment

UBO Ultimate beneficial owner

To give some insight into the composition of cells in the NFDI tables, we calculated the following for each NFDI cell: the number of contributors, the ratio of the second-largest to largest firm, and the concentration ratio (i.e., the ratio of the largest firm to all contributions). Table 6 shows the breakpoints between the first and second terciles of all NFDI cells for each of these measures, along with the lower and upper bounds, by series type.

Table 6: Distribution of Cell Characteristics by Series Type

Series type	Variable	Minimum	Breakpoint (1st to 2nd tercile)	Breakpoint (2nd to 3rd tercile)	Maximum
Investment expenditures	Contributor count	1	4	20	8,182
	2nd largest to largest (percent)	0.00	13.10	60.00	100.00
	Concentration ratio (percent)	0.00	3.30	90.80	(D)
Employment	Contributor count	1	5	20	4,470
	2nd largest to largest (percent)	0.00	15.00	54.50	100.00
	Concentration ratio (percent)	0.00	5.00	78.80	(D)

(D) Suppressed to avoid the disclosure of data of individual companies.

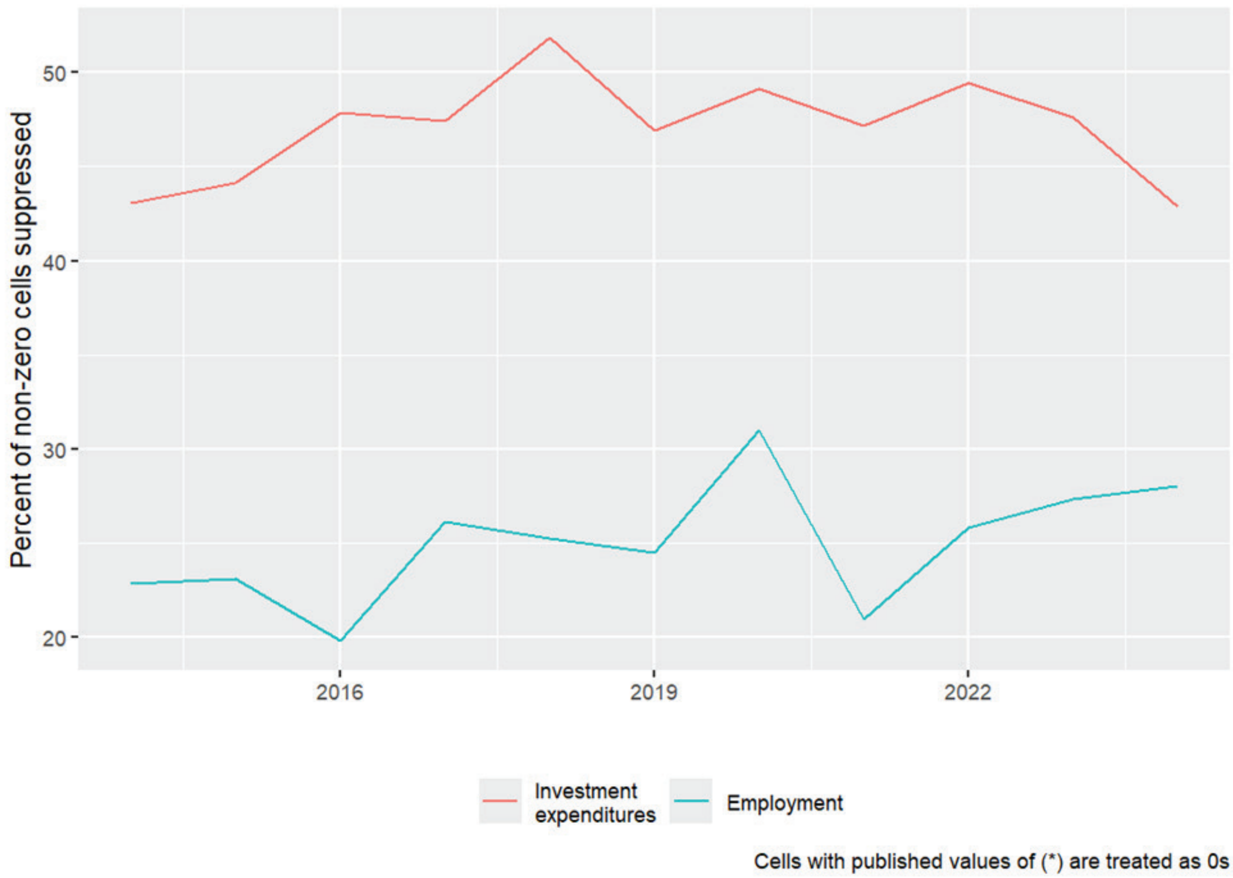
Concentration ratio equals ratio of sum of all but the two largest contributions to a cell to the largest contribution to a cell.

3.2 Suppression Levels in NFDI Tables

BEA's current approach to protecting survey confidentiality (detailed in Section 2.1) has led to relatively high levels of suppression in the published NFDI tables. This is largely a consequence of the relatively small number of transactions underlying the NFDI statistics and their skewed nature; a few large transactions can account for large amounts of the aggregate data for a particular dimension combination (i.e. many of the resulting cells have few respondents).

To demonstrate the level of suppression in these data, Figure 9 shows the percent of nonzero published cells that are suppressed by year for each series type. The denominator of this calculation only includes cells that are eligible for suppression (thus, it excludes cells with a value of 0).

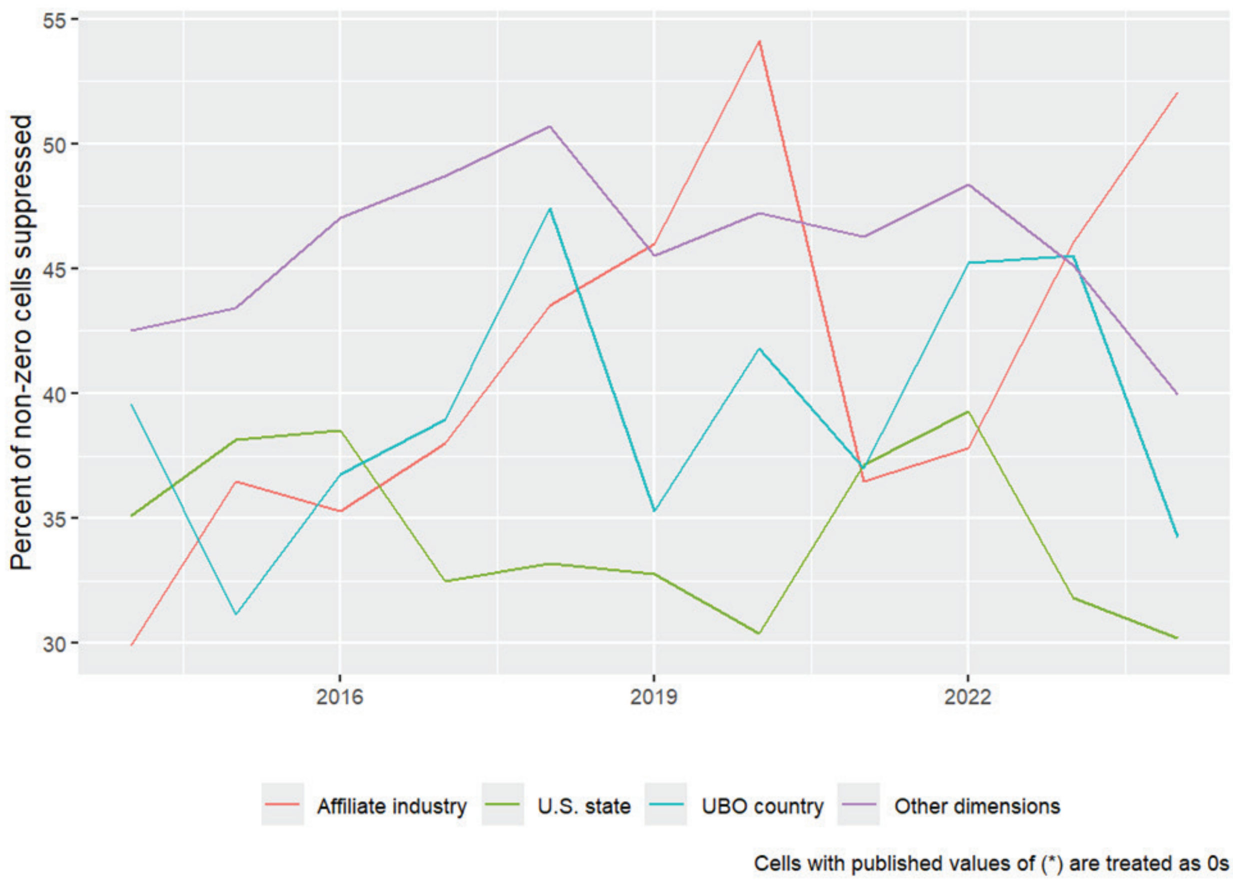
Figure 9: Percent of NFDI Cells Suppressed by Series Type



Although there are noticeably lower levels of suppression for employment, roughly 25 percent of employment cells are suppressed, with roughly 4 percent of cells as primary suppressions. For investment expenditures, closer to 45 or 50 percent of all cells are suppressed, with roughly 14 percent of cells as primary suppressions.

Moving from the frequency of suppressions by series type to the frequency of suppressions by dimension, Figure 10 shows the percent of published cells that are suppressed in each year for the three key dimensions (investment type by either affiliate industry, country of ultimate beneficial owner (UBO), or U.S. state) and all other dimension combinations.

Figure 10: Percent of NFDI Cells Suppressed by Dimension Combinations



With some variance, roughly 40 percent of all potential cells are suppressed overall. In the "other" dimension combination tables, which are often more granular, there are years where more than half of nonzero cells are suppressed.

3.3 Noise Infusion Options

The EZS approach to noise infusion is flexible in several dimensions, providing substantial discretion for a statistical agency to tailor the exact implementation to the specific needs of the agency. This section discusses the available options and the decisions BEA made in adapting the EZS approach for the NFDI data. First, we address two overriding questions that arise in considering a switch from cell suppression to noise infusion:

1. How does the protection conferred by noise infusion compare to the protection conferred by cell suppression?

2. How does the utility of the statistics protected by noise infusion compare with the utility of the statistics protected by cell suppression?

3.3.1 Protection: Noise Infusion vs. Suppression

In broad and somewhat imprecise terms, we aim to provide at least as much protection to respondents under noise infusion as is provided under cell suppression. There are a couple of problems with this statement. First, noise infusion and cell suppression provide protection in different ways, so it is difficult to directly compare protection levels. Second, because of the vagaries of complementary suppression, a cell suppression scheme often overprotects—that is, it generates uncertainty intervals larger than p-percent—and we do not aim to match such overprotection.[40] Thus, a somewhat more precise statement of our goal is that we aim to provide sufficient protection under noise infusion such that a p-percent (or larger) uncertainty interval is obtained for each respondent's contribution to a particular cell.

Noise infusion and cell suppression both create uncertainty intervals in the minds of users. However, both types of uncertainty intervals are incompletely specified, but for different reasons. For noise infusion, the incompleteness arises from the fact that statistical agencies typically do not reveal the specifics of the distributions used to draw noise factors. For cell suppression, the incompleteness arises from ignorance about prior subjective distributions about a given respondent's contribution to a cell.[41]

40. Aggregation also generally leads to overprotection, under both cell suppression and noise infusion. Because we are not proposing changes in the NFDI aggregation structure as a result of the switch from suppression to noise infusion, overprotection resulting from aggregation is ignored in this analysis.

41. Moreover, because each data user comes with a different background, each user is likely to come with a different prior distribution about a given contribution. By Bayes theorem, the posterior probability (the final uncertainty interval in this application) depends on both the prior probability (the initial uncertainty interval) and the experimental realization (the feasible interval for a suppressed cell derived from relationships in a published table).

Thus, to compare protection under the two methods, some additional assumptions are needed. To ensure that noise infusion provides sufficient protection, we use conservative assumptions. Specifically, for noise infusion, we assume that data users know the actual distribution from which noise factors are drawn. We further focus on protection for the most vulnerable respondents: contributors to one-contributor cells or contributors to two-contributor cells in which the other contributor has correctly inferred their own noise factor. For cell suppression, we assume users have a very diffuse prior distribution, namely a uniform distribution over the p-percent uncertainty interval.[42] To the extent that our assumptions are too conservative, the approach to noise infusion we propose will provide more protection than required under the p-percent rule.

In the context of noise infusion, we interpret "protection" as attackers being unable to confidently determine if the true contribution from a firm is close to the published or inferred point estimate for that contribution. For suppression, we interpret protection as the absence of attacker confidence that the true value is close to the midpoint of the p-percent uncertainty interval.[43]

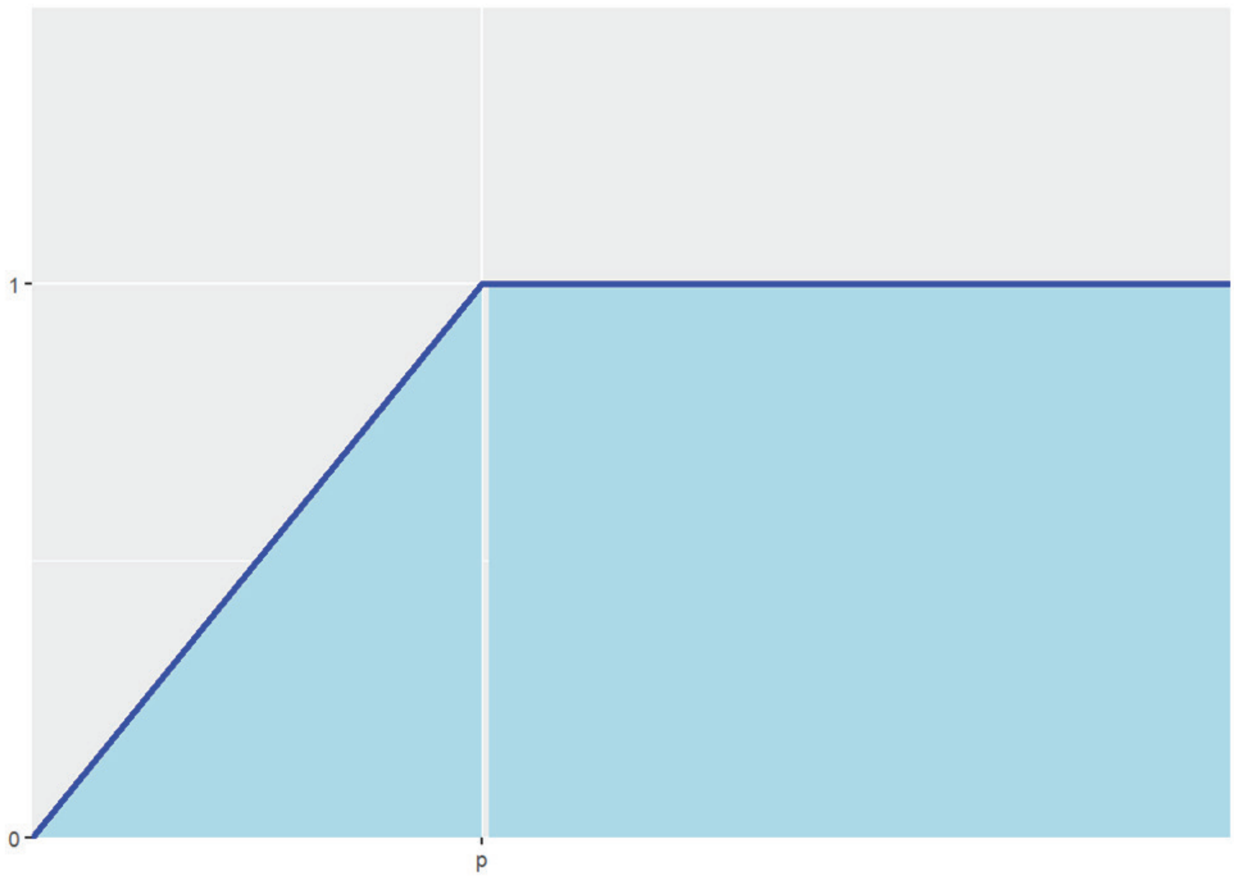
Exploiting the (near) symmetry of all the posterior distributions considered in this analysis, closeness or "concentration" can be measured by the cumulative distribution function (CDF) of the absolute distance between a given threshold point on the distribution and the reference point (published/inferred estimate, or midpoint). For instance, for a threshold of 2 percent, we would compare the CDF of the absolute value of noise factor draws at 2 percent with the CDF of the absolute value of uniform draws at 2 percent.[44] If the CDF of the absolute value of noise factor draws is lower than that of the absolute value of uniform draws, the posterior distribution under noise infusion has less mass within 2 percent of the point estimate and is therefore less concentrated within that range. Put another way, noise infusion provides an attacker less confidence than cell suppression that the true contribution is within 2 percent of the point estimate. Thus, we conclude that noise infusion provides more protection than cell suppression against an attacker being able to estimate a value to within 2 percent.

42. This comparison ignores the fact that the (percentage) posterior distribution about a given noise infused value is not exactly the same as the distribution from which the noise factors are drawn. Unlike the distribution of noise factors, the posterior distribution is slightly asymmetric about the published estimate. However, the differences are small.

43. For conservativism in finding that noise infusion provides as much protection as suppression, we use the midpoint as the point estimate. Given our assumption that the uncertainty interval can be represented with a uniform distribution, if any point other than the midpoint were selected as the point estimate/reference point, the confidence of an attacker in being close to the true estimate would be the same for some closeness thresholds and lower for other closeness thresholds.

44. The "CDF of the absolute value of uniform draws at 2 percent" starts with a uniform distribution ranging from -p percent to +p percent (away from the midpoint). The absolute value of a draw from this distribution is also uniformly distributed but it only ranges from 0 percent to +p percent. The value, at 2 percent, of the CDF of this distribution of absolute values equals the probability of a draw from the original distribution being between -2 and +2 (away from the midpoint).

We set as a standard that the noise factor distribution must be globally less concentrated than the uniform distribution assumed for cell suppression. That is, the CDF of the absolute value of noise factor draws must be everywhere equal or lower than the CDF of the absolute value of uniform draws for noise infusion to be deemed at least as protective as suppression.[45] Because the CDF of the absolute value of uniform draws is linear, the CDF of the absolute value of noise factor draws (expressed as a percentage) must lie completely within the shaded region below:



3.3.2 Data Utility: Noise Infusion vs. Suppression

Using noise infusion for SDL allows publication of cells that would otherwise be suppressed, but it also results in distorted cell estimates relative to "true" values. The former increases the utility of the statistics, while the latter decreases that utility. Assessing whether this tradeoff—more estimates published versus distortion—is justified is challenging for several reasons. Most fundamentally, it is difficult to compare the disutility of a distortion (or set of distortions) with the disutility of an estimate that is missing entirely. Relatedly, it is not clear whether disutility increases linearly with the size of a distortion or in some other manner. In addition, because neither suppressions nor distortions are distributed randomly across tables, it must be determined which cells, if any, are

45. This definition is the mirror image of first-order stochastic dominance.

more important than others and to what extent. Finally, it is not clear whether users view a missing (suppressed) estimate as entirely worthless or if they are willing or able to fill in (i.e., impute) missing values, either with point estimates or ranges, using information from elsewhere in the published table.

In practice, the calculus of disutility almost surely varies by user. Some users are more interested in aggregate cells; others focus on interior cells (either specific cells or the whole set). Some users put a premium on estimate precision; others are satisfied with a ballpark estimate. Some users are well-informed about the economic context reflected in the table and/or are technically savvy and can fill in suppressions with plausible values; others view suppressed values with near-complete uncertainty.

Despite these challenges, we attempt to quantify the tradeoffs arising from a switch from cell suppression to noise infusion for the NFDI data, recognizing our results will not be universally applicable. In this section we describe the exercise we use to quantify these tradeoffs but, to avoid revealing too much about the specific noise-factor parameters we will apply to the NFDI data, we do not show the quantitative results.

Our approach is simple in principle. We calculate the distortions from applying noise and compare them to distortions that users might obtain when imputing suppressed values on their own.[46] In practice, both the perturbed estimates (under noise infusion) and imputations of suppressions have a random element, so we compute 500 replicates of each and compare distortions over the whole set of replicates. This comparison requires specifying how to aggregate distortions across table cells and specifying how users might impute suppressed values.

For adding up distortions, we opt for a fairly conventional approach to proxy the disutility of distortions using a version of root mean square distortion (RMSD) over the replications of noise infusion and suppressed values. Specifically, for each NFDI table cell we first calculate the mean squared distortion (MSD) for each replicate as the average across replicates of the square of the difference between each cell's true value and distorted value, whether this difference results from noise infusion or imputation:

$$\text{MSD}_c = \sum_{r=1}^{500} \left(\frac{(\text{True Value}_{cr} - \text{Distorted Value}_{cr})^2}{500} \right) \qquad (6)$$

where c indexes cells in the table and r indexes replicates.

46. Having to impute a suppressed value creates disutility for a user separate from the disutility created by the distortion in the imputation. We ignore such disutility in this exercise, even though for some users the disutility is sufficiently great such that no effort to impute is made.

Next, we average the cell-level MSDs within each table, t, to get the table-level MSD:

$$\text{MSD}_t = \frac{\sum_{c=1}^{C_t} \text{MSD}_c}{C_t} \qquad (7)$$

where t indexes NFDI tables and where C_t is the total number of cells in table t.

Taking the square root yields the table-level RMSD.

$$\text{RMSD}_t = \sqrt{MSD_t} \qquad (8)$$

We convert this to percentage terms by calculating the relative root mean square distortion (RRMSD) for the table by multiplying the RMSD by the square root of the number of cells divided by the sum of squared true values:

$$\text{RRMSD}_t = \text{RMSD}_t \times \sqrt{\frac{C_t}{\sum_{c=1}^{C_t} \text{True Value}_{ct}^2}} \qquad (9)$$

Finally, to calculate the total distortion, we take the unweighted average of each method's RRMSD_t values across all NFDI tables:

$$\text{Total Distortion} = \frac{\sum_t^T \text{RRMSD}_t}{T} \qquad (10)$$

where T is the number of NFDI tables covered in this exercise. This produces a single value summarizing the distortion across the NFDI statistics from either noise infusion or suppression imputations.

How users might impute the value of a suppressed cell depends both on which values are feasible given the unsuppressed cells in the table and on their prior knowledge. For our application, we focus on a hypothetical well-informed, technically sophisticated user. In terms of prior knowledge, we assume that the user knows each suppressed value in a table to within 100 percent (plus or minus).[47] We also assume that the user initially assigns equal probability to any value within this range.

Based on these assumptions, we then assume imputation proceeds in two steps. The user first imputes a value for a small interior cell by drawing from the uniform distribution with lower bound zero and upper bound twice the true value. Sometimes row and/or column totals imply that certain

47. Note that this is the same assumption made about a potential attacker's prior information with respect to an individual contribution (or cell) in the standard derivation of the p-percent rule.

values in this distribution are infeasible; in such cases, the lower and/or upper bound is first adjusted to avoid the infeasible values. The user then moves sequentially to impute larger and more aggregate cells, updating their priors at each step to incorporate any constraints imposed by earlier draws.

Our focus on a well-informed, sophisticated user provides a relatively challenging standard for noise infusion. By comparison, users who have less prior knowledge about the underlying economic activities or who are less willing and/or able to systematically impute suppressed values will likely attach higher disutility to suppression, which would make infusion even more attractive.

In at least one respect, noise infusion seems well-suited for protecting the NFDI data. Namely, a significant percentage of cells in the NFDI statistical tables are suppressed, as noted in Section 3.2. Holding constant any utility specification, the more suppressions there are, the more likely noise-infused tables will represent an improvement. Another factor affecting the comparisons is how much oversuppression—that is, the number and width of uncertainty ranges wider than p percent—occurs in the tables.[48] It is more difficult to quantify oversuppression than the share of cells suppressed.

As noted above, we use the "Total Distortion" measure from Equation 10 to compare the use of noise infusion and suppression for SDL of the NFDI tables. Several parameterizations of noise factor distributions are available that provide both at least as strong respondent protection (see Section 3.3.1) and better data-user utility than suppression, including the one BEA intends to use. Other considerations underlying selection of this distribution, along with other aspects of the implementation of noise infusion, are discussed in the remainder of Section 3.3.

3.3.3 Distribution of Noise Factors

Taking as given that the distribution of noise factors is symmetric around zero, there are three decisions to be made about the distribution:

1. Which statistical distribution should be used (and, more broadly, whether to use a bounded or unbounded distribution).
2. How much variation the distribution should encompass.
3. Whether the distribution should be split.[49]

---

48. By this definition, complementary suppressions result in oversuppression even though they are necessary to protect primary suppressions.

49. The term "split distribution" could mean more than one thing. We use the phrase in the same manner as the Census Bureau does when describing the distributions they use for EZS noise infusion, i.e., a distribution in which the PDF is cut in half (at zero) with the positive half moved right and the negative half left by the same amount so that there is a range centered on zero containing no probability mass.

    In contrast to this usage, a split normal distribution is described on Wikipedia as "result[ing] from joining at the mode the corresponding halves of two normal distributions with the same mode but different variances."

We considered several distributions, formally testing some, and selected one to use for the NFDI data. In terms of boundedness, although certain parameterizations of certain unbounded distributions are almost certain to yield "reasonable" distortions, we prefer to use bounded distributions to be completely sure.

In order to avoid revealing information that could be used by attackers to infer more than we intend to disclose about cell values, we do not identify the specific distribution type or range of our chosen distribution in this paper.[50] We selected this distribution (and its parameters) based on a Monte Carlo comparison of the distortions (see Section 3.3.2) produced under various distributional specifications, under the constraint that protection of vulnerable contributions should be at least as strong as provided by the p-percent rule with suppression (see Section 3.3.1). Based on this analysis, we further chose to use a split distribution. While splitting the distribution is not the only way to increase the level of protection, it does guarantee that the only contributor to a single-contributor cell will not see its contribution published extremely close to the actual reported value.

There is a wide variety of bounded symmetric split distributions we could have selected, and the results from our selected approach could be closely approximated by using different distributions or parameterizations. The specifics of the selected distribution are not decisive. Nevertheless, our distribution has been selected to limit the disutility from noise infusion while preserving *at least as much* respondent protection as that provided by cell suppression.

3.3.4 Independent or Dependent Noise Factors

While the noise factors of unrelated contributions should be drawn independently, it may be beneficial to ensure that related contributions are assigned similar (or the same) noise factors. For instance, if a reporter provides values for items A and B (both assumed to be positive) and the compiling statistical agency adds those two values together to get composite item C, the probability of an implied noise factor close to zero for item C is higher than for the explicit noise factors of either item A or B if those explicit noise factors are drawn independently. On the other hand, if those two items are assigned factors of the same sign, the implied noise factor of C will be at least as far away from zero as the lower of the two assigned factors. Further, if composite item C is the difference between A and B, only using the same factor for both A and B guarantees that the perturbed value of C will take the same sign as the unperturbed value.

50. More information may be provided once noise infusion is used for the published statistics. For instance, the Census Bureau provides information about the specific distribution used (triangular) and whether it is split (yes) in its Quarterly Financial Report statistics, but it doesn't provide information on how much variation the distribution encompasses (i.e., the distributional maximum and minimum and how far apart the distribution is split away from zero).

For statistics classified by geography and/or industry (or other similar dimensions), like many BEA statistics, a single reporter may report different values for the same item in multiple categories.[51] Allowing for independent noise factor draws could result in less protection for the reporter at aggregate levels than at detailed levels.

Using similar factors for different items from the same reporter limits distortions of the ratio of two items. Likewise, using similar factors (e.g., same factor or same sign of factor) for a given item from the same reporter from one period to the next may limit growth rate distortions. Limiting distortions in cases such as these comes at the potential cost of reducing the amount of protection implicitly provided to ratios or growth rates. However, when factors are drawn fully independently, the protection provided to ratios or growth rates generally exceeds the protection provided to the underlying reported items.

Compared to other BEA surveys and statistics, the NFDI surveys and statistics feature few instances of the same reporter providing different values of the same item in multiple categories and, especially, few instances of similar distributions of contributions in a given cell from one period to the next.[52] Consequently, we found that many outcomes differ relatively little whether noise factors are drawn fully independently over time and over different reports from the same reporter. Thus, there was little cost in terms of foregone protection in constraining noise factors to be the same for all items from a given reporter. For example, if current-year greenfield investment of a given reporter were assigned a noise factor of $-X$ percent, then total planned greenfield investment for that reporter would also be assigned a noise factor of $-X$ percent.

3.3.5 Enhancements to Limit Distortion

Although the expected distortion for any given cell under EZS noise infusion is zero, the probability that a given cell has distortion of exactly zero approaches zero.[53] Sometimes the luck of the draw means that a given cell, even if highly aggregated, can be more distorted than anticipated. There are potential modifications to the basic EZS methodology that can lower or eliminate distortions in important cells or important tables.

"Balanced EZS" is one of the more common of such modifications. Under this modification, within an important cell, respondents are sorted by the size of their contribution, and the signs of their

51. For example, in BEA's outward activities of multinational enterprises (AMNE) statistics a single U.S. parent company may report host-country employment separately for affiliates in both Thailand and Cambodia.

52. For example, imagine that an NFDI cell covering the intersection of industry X and country Y is composed of contributions from reporters A (50 percent), B (30 percent), and C (20 percent). It is uncommon for A, B, or C to also contribute to a cell covering industry W or country Z, and it is uncommon for the same cell the next year to include reporters A, B, or C.

53. Rounding of the published statistics implies that the probability is positive.

noise factors are alternated between positive and negative so the cell does not randomly end up dominated by factors of one sign or the other.[54] Strictly speaking, this is a more complex instance of using dependent rather than independent draws; the dependence in this case does not stay within the realm of a given reporter but allows the sign component of one reporter's draw to affect that of another reporter.

A different modification to limit distortion is to modify the tables after noise factors have already been drawn and applied. This "post-processing" modification consists of identifying key cells within a perturbed table, resetting the values of these cells back to the unperturbed values, and raking the remaining perturbed cells to restore table additivity. This modification allows for selected key aggregates to be entirely undistorted, but requires that any vulnerable cell be excluded from the set of key aggregates and may further distort those cells that are not restored to their unperturbed values.[55]

For the NFDI data, we do not plan to apply either of these modifications. Both come with costs, including increased complexity for implementing noise infusion. Based partly on our Monte Carlo simulation, we view these costs as exceeding the benefits they would bring for the NFDI data. For other statistics, these costs may well be worth the benefit of preserving specific aggregates or otherwise minimizing distortions.

3.4 Impact of Noise Infusion in the NFDI Tables

3.4.1 Simulations

This section presents results quantifying the impact of our selected approach to noise infusion on the NFDI tables based on a simulation exercise. For this exercise we performed 500 replications; that is, we applied our selected approach to noise infusion to the NFDI microdata 500 times, generating 500 sets of noise-infused microdata. We then converted each of these separate microdata instances into the equivalent noise-infused NFDI tables, generating 500 noise-infused versions of each NFDI table published between 2017–2024.

3.4.2 Results

To avoid providing too much information about the specifics of the noise factor distributions, we provide only limited information about the impact noise infusion has on the NFDI tables in this

54. See Paul B. Massell and Jeremy M. Funk (2007), Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection in proceedings of the *2007 Federal Committee on Statistical Methodology Research Conference* for a more detailed and precise description of balancing.

55. In addition, if noise factors are drawn from a split distribution, this modification eliminates the guarantee that the published (perturbed) value of a single-contributor cell is at least a certain distance away from the unperturbed value of that cell.

exercise. Specifically, we divide the NFDI tables based on the level of aggregation represented by each cell. Our expectation is that noise infusion will generally have smaller impacts on more aggregated cells due to offsetting perturbations. First, we look at how estimates of the single highest-level aggregate for first-year investment expenditures are distorted by noise infusion. This value is generally used to summarize changes in the NFDI data (i.e., it is the first number mentioned in BEA's news releases on NFDI). Then we look at distortion and other impacts of noise at three levels of aggregation that, together, cover all cells in the published tables.

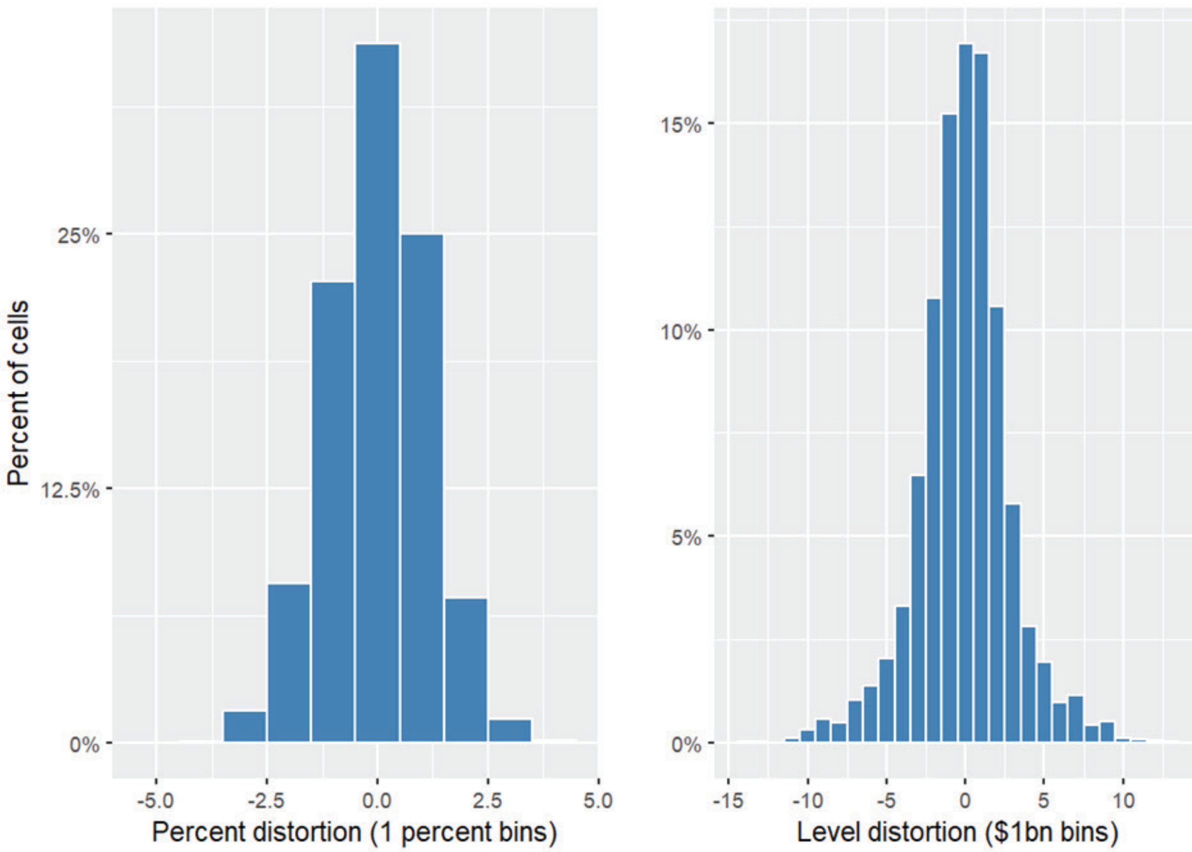
For aggregate investment expenditures, Table 7 provides summary statistics of the impact of noise infusion on aggregate first-year investment expenditures by year, while Figure 11 shows histograms of the percent and level perturbations from noise infusion in these cells. Both the table and the figure confirm that, for this aggregate, the perturbed estimates are unbiased—that is, the means of the perturbed estimates are (approximately) the unperturbed estimates, despite any given perturbed estimate differing from the corresponding unperturbed estimate. Unbiasedness of the perturbed estimates is a feature built into the design of the EZS noise infusion adopted for the NFDI statistics. Unbiasedness does not only hold for the highest-level aggregate; it also holds for any other cell.[56]

Table 7: Summary Statistics for Aggregate First-Year Investment Expenditures by Year
(Millions of USD)

Year	Published value	Summary of 500 replicates					
		Mean	Min	Percentiles			Max
				25th	50th	75th	
2017	272.8	272.8	265.1	270.9	272.8	274.7	281.1
2018	312.5	312.2	298.2	308.0	312.3	316.2	325.2
2019	221.2	221.0	215.2	219.7	221.1	222.4	226.5
2020	141.4	141.4	137.1	140.5	141.4	142.4	145.2
2021	362.6	362.7	352.2	359.9	362.6	365.7	374.8
2022	206.2	206.2	202.1	205.4	206.4	207.2	210.6
2023	176.0	175.8	170.8	174.4	175.7	177.3	180.9
2024	151.0	151.2	146.3	150.0	151.3	152.5	155.9

56. The rest of the results discussed in this section take unbiasedness as a given. For example, instead of looking at mean perturbations, attention is focused on mean absolute perturbations and related measures.

Figure 11: Percent and Level Distortions of Aggregate First-Year Investment Expenditures



At an aggregate level, the impact of noise infusion is very small, with the great majority of distortions less than 2 percent and $10 billion in absolute value. The largest percentage distortion in absolute value among the 500 replicates of the 8 aggregates is 4.49 percent, with 5 of the 8 aggregates distorted less than 3 percent in all 500 replications. A supplemental simulation with 10,000 replicates of each aggregate (not shown) confirms that expanding the number of replicates only slightly increases the largest absolute distortion for each of the 8 estimates.[57]

57. For example, the largest percentage distortion of the 8 aggregates is 5.06 percent in the 10,000-replicate simulation compared with 4.49 percent in the 500-replicate simulation.

Moving to consideration of other cells in the NFDI tables, we define the following three levels of aggregation:

1. *Top-level aggregates*, an aggregate whose published dimensions include at least one of "all states," "all industries," "all countries," or "all investment/expenditure years." In some tables, these may still contain some level of disaggregation in another dimension: for example, many tables show several levels of investment type detail for each top-level aggregate. Note that one of the aggregates at this level of aggregation is the highest-level aggregate, which was the focus of Table 7 and Figure 11.
2. *Second-level aggregates*, any aggregate that is directly below one of these top-level aggregates in at least one dimension (for example, "Europe" is directly below "all countries").
3. *Interior cells*, representing all other cells.

For each cell we then calculate the absolute percent distortion (i.e., the absolute percent difference between the unperturbed and perturbed values) for each of the 500 replications. Table 8 shows the mean absolute percent difference, median absolute percent difference, and the standard deviation of absolute percent differences by series types and level. For all three calculations, we weight each cell by its (unperturbed) absolute value (dollars for investment expenditures or employees for employment).

Table 8: Distortions from Noise Infusion in NFDI Tables by Cell Level

Series type	Level	Weighted absolute percent distortion		
		Mean	Median	Standard deviation
Investment expenditures	Top-level aggregates	1.3	0.9	1.3
	Second-level aggregates	2.1	1.6	1.8
	Interior cell	2.8	2.4	2.2
Employment	Top-level aggregates	1.2	1.0	0.9
	Second-level aggregates	2.1	1.6	1.7
	Interior cell	3.4	3.1	2.3

The results are consistent with our expectations: weighted by the absolute size of the cell, larger cells generally have smaller impacts from noise infusion. As shown in the "interior cell" rows, the impact of noise infusion is largely offsetting even for more granular cells.

Table 9 shows similar results classified by whether replicated cells are primary suppressions (i.e., whether BEA currently considers them to be vulnerable).

Table 9: Distortions from Noise Infusion in NFDI Tables by Vulnerability

Series type	Primary suppression	Weighted absolute percent distortion		
		Mean	Median	Standard deviation
Investment expenditures	Yes	4.7	4.4	2.5
	No	1.9	1.3	1.7
Employment	Yes	5.4	5.0	2.3
	No	2.2	1.6	1.9

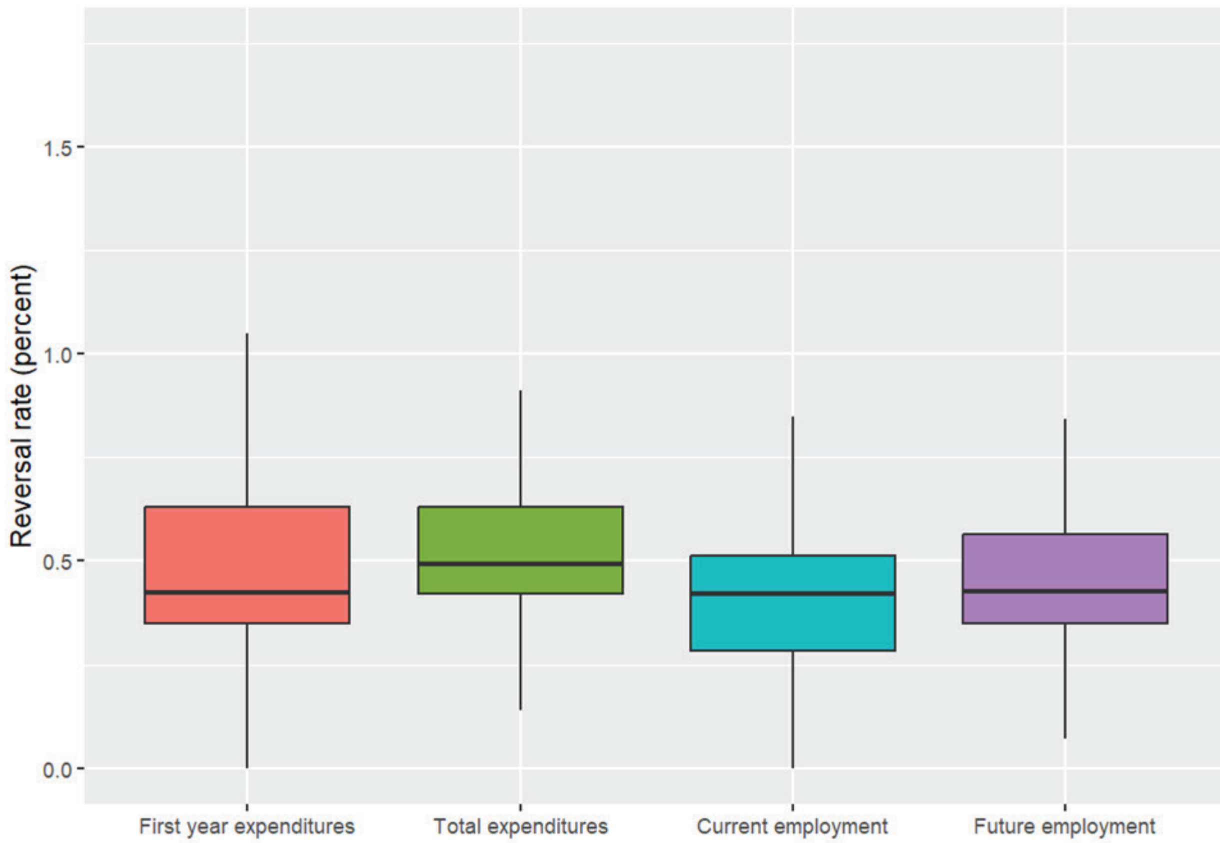
The results in this table are also consistent with our expectations: weighted by the absolute size of the cell, more aggregate cells are generally less distorted from noise infusion, as measured by the mean absolute percent distortion, than lower-level cells. Nonetheless, record-level perturbations are largely offsetting at the cell level even for the most granular (i.e., interior) cells. In this table, median absolute distortions are smaller than mean absolute distortions, suggesting a degree of right skewness in the distribution of distortions. Standard deviations are mostly smaller than corresponding means.

To give a sense of the relative scale of these impacts, we note that the median weighted absolute percent revision in the NFDI tables is roughly 14 percent. In other words, BEA's annual update typically changes estimates more than the envisioned noise infusion approach.

Users of the NFDI statistics have expressed particular concern about how noise infusion might impact state-level results, which are of great interest to NFDI data users. Specific interest has been expressed in states' *rankings*. To examine the impact of noise infusion on these rankings, we calculated all pairwise combinations of state-level totals over all types of investment for each year and series within each replicate, thereby generating, for instance, every combination of "Alaska" and all other states for each series and year, etc. For each of these combinations we then identified how frequently the application of noise infusion reversed the *sign* of the inequality of values between those two states. In other words, we recorded a reversal of, for example, Alaska > Maine in the unperturbed data but Maine ≥ Alaska in the perturbed data.

For a given state-level series and replication, we calculate the reversal rate of each combination of year and pairwise state comparison. Figure 12 shows boxplots of these annual reversal rates by series; the boxplots represent the distributions over the 500 replicates.[58]

Figure 12: Ranking Reversal Rates by Series for State-Level Tables

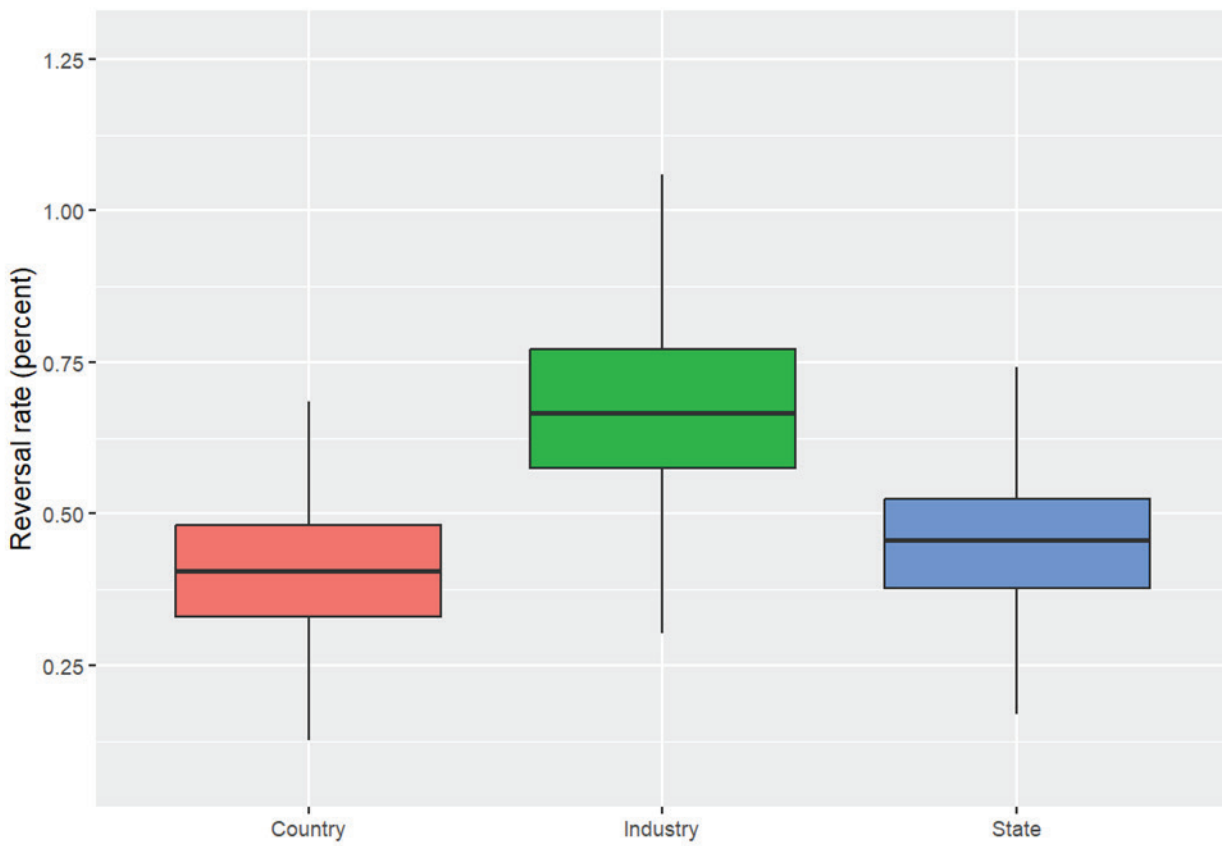


It is evidence that applying noise infusion to the NFDI data has minimal impacts on the various rankings between states, with median and mean reversal rates well below 1 percent.

58. The whiskers in these boxplots and all other boxplots in this paper represent 1.5 times the interquartile range (IQR), the difference between the 75th and 25th percentiles.

While data users have primarily expressed interest in the value rankings for by-state tables, for completeness, Figure 13 shows boxplots of the distribution of reversal rates of investment expenditures and employment series types for the three key table types. Each plot covers both series types.

Figure 13: Reversal Rates by Table Type



Ranking reversals are most common for the industry tables (which have more granular levels than other tables), but are still almost always less than 1 percent. They are least common for country tables.

EZS noise infusion does not only affect values explicitly published in the annual NFDI tables but also affects certain measures that can be inferred from those values. Most prominently, noise infusion affects growth rates and revision rates. For example, using cell suppression for SDL, an unperturbed cell value might be 200 in the preliminary NFDI statistics and 210 in the revised statistics published a year later. If noise infusion were instead to be used for SDL, the corresponding perturbed values might be 206 in the preliminary statistics and 202 in the revised statistics.[59] Using the unperturbed data, the cell value is revised up 5.0 percent. By contrast, the cell value is revised *down*

59. These represent distortions of 3.0 percent and −3.8 percent, respectively, of the estimates of the levels.

1.9 percent when the perturbed data are used, representing a 6.9 percentage point negative distortion to the revision rate and reversal of the sign of the revision.

We conclude this section with information on how noise infusion affects implicit measures of growth and revisions in the NFDI tables. To give some sense of how the proposed noise infusion process impacts growth rates in the NFDI tables, we calculated the growth *ratio* (i.e., current year value divided by previous year value) for each NFDI cell for both the perturbed and unperturbed values in each replication of the NFDI tables. We then calculated the natural log of these ratios and, finally, the absolute difference between the two logs (the "distortion").[60] Table 10 shows the mean, median, and standard deviation of these logs by cell level. Although the bounds for distortions of log growth, when multiplied by 100, are wider than the bounds for the percent distortion of levels (for non-negative values), the mean absolute distortion of growth rates is quite moderate.[61]

Table 10: Absolute Distortion of Log Growth from Noise Infusion by Cell Level

Level	Mean	Median	SD
Top-level aggregates	0.040	0.029	0.035
Second-level aggregates	0.043	0.033	0.037
Interior cell	0.050	0.040	0.039

Log growth measured as $ln(X^c/X^p)$ where *c* denotes the current year's value and *p* denotes the previous year's value.

60. We use log growth instead of percentage growth (and do the same for revisions) because growth (and revision) rates can be very high (positive or negative) in the NFDI data. With high rates, using mean percentages can be misleading, in part because positive and negative rates are not symmetric. Log rates do not suffer from the same deficiency. For relatively low growth (or revision) rates, using logs gives nearly the same result, when multiplied by 100, as using percentages.

61. Distortion of growth could be expected to be even lower if 1) there was more year-to-year continuity in the composition of contributors to cells than in the NFDI data, and 2) the noise factor for a given contributor was similar or the same from year to year.

Table 11 shows similar results for the impact of noise infusion on revisions by cell level. For this table we calculated how noise infusion would have impacted the level of revision for any cell revised during the 2025 release of the NFDI tables (revisions in the 2025 release primarily pertain to statistics for 2023). To a certain extent, results are driven by many cells that receive no revisions (and whose distortion to log revisions are thus precisely zero); to emphasize the relative frequency of these unrevised cells, we also show the percentage of cells at each level that are unrevised.

Table 11: Absolute Distortion of Log Revision by Cell Level

Level	Mean	Median	SD	Addendum: Share of cells with exactly zero distortion because of no revision (percent)
Top-level aggregates	0.005	0	0.015	61
Second-level aggregates	0.010	0.000	0.021	43
Interior cell	0.010	0.000	0.023	47

Log revision measured as $ln(X^r/X^p)$ where *r* denotes the revised estimate and *p* denotes the preliminary estimate.

Finally, Table 12 shows reversal rates (i.e., the percent of cells for which noise infusion reverses the direction of a change) for cell growth and cell revisions by cell level. Noise infusion has minimal effects on the direction of revisions. For growth, noise infusion reverses the direction more frequently, but this is largely due to reversals of cells with low growth rates (the median absolute growth rate for cells with a reversal is 1.7 percent, compared to 84.4 percent for cells without a reversal).

Table 12: Reversal Rates for Growth and Revision by Cell Level (Percent)

Level	Direction of growth	Direction of revisions
Top-level aggregates	1.30	0.19
Second-level aggregates	1.24	0.21
Interior cell	1.47	0.42

4. Future Plans for Noise Infusion

As noted previously, BEA plans to adopt EZS noise infusion as its primary SDL method for the NFDI data in 2026. This is expected to be the beginning of a process of transitioning from cell suppression to noise infusion for SDL for statistics based on surveys administered by BEA itself.[62] In broad terms, BEA administers two survey programs:[63]

- International trade in services (quarterly, annual, and benchmark surveys)
- Direct investment

The direct investment survey program is composed of three subprograms:

- New foreign direct investment (NFDI) (annual survey)
- Activities of multinational enterprises (annual and benchmark surveys)[64]
- Direct investment transactions and positions (quarterly, annual, and benchmark surveys)[65]

BEA expects to adopt noise infusion for its other survey programs, with timing and sequencing to be determined. This section discusses challenges that are expected to accompany this transition based on the unique features of each program.

A major difference between the NFDI program and the trade in services and direct investment transactions and positions programs is that the latter programs provide source data for multiple statistical products. For example, the direct investment transactions and positions program contributes to the International Transactions Accounts, the International Investment Position Accounts, the direct investment by country and industry statistics, and the statistics on services trade by enterprise characteristics. When survey data feed into multiple statistical products, aside from potential challenges of coordinating the timing of the transition, the specific EZS implementation used must be satisfactory vis-à-vis each statistical product to avoid inconsistencies. For instance, the perturbed direct investment financial transactions in the International Transactions Accounts must be the same as the perturbed direct investment financial transactions that contribute to the change in the perturbed direct investment position in the International Investment Position Accounts. If some post-processing occurs in one (or both) of the statistical products to reduce the distortion of a given aggregate, it becomes difficult to ensure such consistency.

62. The bulk of BEA statistics are based on surveys administered by other government agencies or from other public and private source data. Most of these other source data come to BEA having already had any necessary SDL methods applied. However, for BEA's International Economic Accounts, much of the source data (roughly a fifth for the International Transactions Accounts and a quarter for the International Investment Position Accounts; much more for other international economics accounts) comes from surveys BEA administers itself, for which BEA is required to provide confidentiality protection to respondents.

63. See figure A.2 and the "Data Sources" subsection in U.S. Bureau of Economic Analysis, *U.S. International Economic Accounts: Concepts and Methods* (Washington, DC: BEA, September 2025, 277 and 21–23).

64. Surveys are conducted separately for outward direct investors and inward direct investment enterprises.

65. Surveys are conducted separately for outward direct investors and inward direct investment enterprises.

Another complication that is more prominent in some of the other survey programs, though present to some extent in the NFDI program, is data items that are additively or otherwise mathematically related. For instance, the multinational enterprises (MNE) surveys collect balance sheet information in which the sum of assets of various types must equal total assets. Similarly, owners' equity must equal total assets minus total liabilities. If, however, different items from a given contributor are assigned different noise factors, the perturbed versions of these relationships usually will not hold.[66] While, in principle, BEA could publish statistics that do not reflect such accounting relationships, doing so could sow confusion among data users.

A third characteristic of the other survey programs that is notably different than the NFDI program is that they predominantly collect data items that show continuity from period to period because they reflect contributions from a similar set of firms conducting similar activities from one period to the next.[67] As a consequence, growth rates are typically relatively moderate, so data users can use the published statistics to calculate growth rates that are more meaningful and interpretable than those obtained from the NFDI statistics. In a situation where EZS noise factors are applied independently from period to period, growth rate distortions can be larger than distortions of levels, particularly for cells with few contributors.[68] This can generate an incentive to use a scheme in which this period's noise factor depends on, and is not too dissimilar from, last period's noise factor. Such a scheme, however, comes with the tradeoff of less protection for survey respondents in terms of growth rate outcomes compared to the independent noise factor baseline. In the extreme of a single-contributor cell and no change from last period in the noise factor, the growth rate is entirely unprotected despite the level estimate being sufficiently protected.

Implied ratios in surveys—for example, a "wage" or "salary" that is not collected or published but can be calculated implicitly as the ratio of employee compensation and total employment—pose similar issues as implied growth rates. Namely, 1) if noise factors are drawn independently for each data item, the maximum percent distortion of the ratio can be higher than the maximum percent distortion of the data items and 2) if noise factors are not drawn independently, the protection of the implied ratio can suffer. As with implied growth rates, implied ratios are more of a consideration for some of the other survey programs than for NFDI. The rest of this section notes some of the specific features of each survey program, including somewhat representative information on the current prevalence of suppression in published statistics, that may influence BEA's decision-making in implementing EZS noise infusion.

---

66. For instance, suppose owners' equity = 100, assets = 500, and liabilities = 400 in a contributor's unperturbed data. If the items are assigned noise factors of 7 percent, −12 percent, and 4 percent, respectively, the necessary equality fails to hold in the perturbed data: $100 \neq 440 - 416$.

67. By contrast, to a large extent, the new investment reflected in the NFDI statistics tends to be lumpy and transitory.

68. Specifically, for a cell with only positively valued contributions, the upper bound of the (absolute) percentage point distortion of the growth rate exceeds the upper bound of the percent distortion of the estimate of the level, even for small (and zero) growth rates.

International trade in services:[69]

- For exports and imports in International Services table 2.2 for 2024, which presents annual trade in services by service type and by country or affiliation, 23 percent of nonzero cells based primarily on BEA survey data were suppressed.[70] In International Transactions Accounts tables 1.3 and 1.4 for 2020–2024, which present *quarterly* service type by country data for more aggregated service types and a smaller number of countries, only 7 percent of nonzero cells based primarily on BEA survey data were suppressed. By contrast, in International Services table 3.3, which presents annual trade in information and communication technology services and in digitally deliverable services by country or affiliation, 45 percent of exports and imports cells were suppressed.
- The trade in services data feed into several International Economic Accounts statistical products and into BEA's National Income and Product Accounts (NIPAs), including into GDP. Given the prominence and impact of these statistical accounts, especially GDP, BEA is carefully evaluating whether limiting—through modifying of the basic EZS methodology— distortions at the highly aggregated levels at which these data enter the NIPAs would be warranted in applying noise. BEA is unlikely to use an aggressive modification but is still investigating the suitability of options that could be feasibly implemented and with less impact on the more disaggregated statistics.
- The number of relationships between data items is relatively small, both in terms of additivity constraints and implied ratios of interest. This reduces the complexity of designing the noise infusion approach compared to a program with more of these relationships.
- Noise factors are applied at the contributor level, for which data are not seasonally adjusted. They are then aggregated, and the aggregated (monthly and quarterly) data are often adjusted for seasonality, which yields the seasonally adjusted statistics that are of primary interest to many data users. The use of perturbed nonseasonally adjusted data to develop seasonal adjustment factors may result in muting or distorting actual seasonal patterns.

69. In previous years, BEA examined several implications and potential outcomes of adopting EZS noise infusion for SDL of the trade in services statistics; see John Bockrath and Dan Yorgason, "Exploring Noise Infusion for Disclosure Avoidance at BEA" (background report for June 9, 2023 presentation at Federal Economic Statistics Advisory Committee (FESAC) meeting) for a discussion. As of 2023, BEA planned to adopt noise infusion for the trade in services statistics before other products. BEA's sequencing plans have since evolved to begin with the NFDI statistics.

70. Some International Services statistics in table 2.2 are based primarily on data sources other than BEA surveys. Including such cells, and cells with values of zeros, in the denominator reduces this share to 17 percent from 23 percent.

MNEs:

- In a selection of the 2022 outward data tables, 26 percent of nonzero cells were suppressed.[71] In a selection of the 2022 inward data tables, 37 percent of nonzero cells were suppressed.[72]
- The MNE surveys cover the finances and operations of MNEs (both affiliates and, to a lesser extent, parents). This results in more data items and tables than for the other survey programs. Moreover, certain data items are included in several different types of tables. For any potential approach to reduce the distortion of key aggregates (e.g., balanced EZS or post-processing), the challenge is heightened when it is difficult to single out a particular table as "key."
- Several of the data items are related to one another (balance sheet and income statement items, among others), resulting in a very consequential decision as to whether noise factors across different data items should be dependent (or the same) or independent.

Direct investment transactions and positions:

- In a selection of 2020–2024 annual data tables, 24 percent of nonzero cells were suppressed for outward investment and 31 percent of nonzero cells were suppressed for inward investment.[73] In a selection of 2020–2024 *quarterly* data tables (across both outward and inward investment), only 3 percent of nonzero cells were suppressed.[74]
- As with the trade in services data, these data feed into several International Economic Accounts and then on to the NIPAs (as part of corporate profits), so there may be interest in minimizing distortions to key aggregates.
- As with the MNE data, several data items are related to one another. In particular, the relationship between direct investment positions and financial transactions and other flows is key. It is important that the stocks and flows can be reconciled in both the perturbed statistics and the underlying unperturbed data.
- As with the international trade in services data, some of the transactions estimates are adjusted for seasonality.

71. The selection covers the 26 country by industry of affiliate tables, which are relatively detailed tables and represent the most common breakdown in the outward data publication. (In all, there are 97 tables in the 2022 publication.) Twenty-three percent of all cells in these tables (zeros included) are suppressed.

72. The selection covers the 27 industry of affiliate by country of UBO tables, which are relatively detailed tables and represent the most common breakdown in the inward data publication. (In all, there are 93 tables in the 2022 publication.) Thirty-four percent of all cells in these tables (zeros included) are suppressed.

73. The selection covers country by industry tables for financial transactions, income, and position along with a table on the change-in-position decomposition for industry sectors and selected countries. The suppression share of all cells in these tables (zeros included) was 23 percent for outward investment and 29 percent for inward investment.

74. The selection includes International Transactions Accounts tables 4.5 and 6.2, which, respectively, present direct investment income and direct investment financial transactions by country and by industry.

5. Conclusion

In this paper, we have outlined BEA's plans to replace cell suppression with EZS noise infusion for SDL. The switch will first occur for NFDI statistics to be released in June 2026. BEA expects it will occur for other International Economic Accounts statistics over the next several years, though BEA has not yet determined the exact sequencing or timing of the switch for these other statistics. The application of SDL, and therefore the switch in SDL methods (and, indeed, SDL currently used for these statistics) pertains only to source data obtained in surveys conducted by BEA; with one exception, other source data are either already administered SDL by their providers or are not administered SDL at all.

Differences in the structure and content of BEA surveys result in different challenges and opportunities in moving to EZS noise infusion. Some surveys, such as the NFDI survey, comprise a relatively limited number of variables and reflect relatively few structural relationships between variables. This tends to make it simpler to assess and apply noise infusion. For some statistics, the impact of noise infusion on growth rates is a more important consideration than it is for NFDI. This tends to complicate decisions in applying noise infusion. Most BEA surveys underlie statistics with considerable cell suppression, though not necessarily to quite the same extent as with NFDI. For this reason, and because of the complexity of cell suppression, the switch to EZS noise infusion is expected to be of widespread benefit to data users and to BEA itself.

Our detailed study of the impact of using noise infusion for the NFDI statistics confirms that under reasonable assumptions the protection provided to survey respondents is (at least) maintained and that data utility is considerably strengthened. We show that the distortions of the NFDI statistics produced by noise infusion are generally small, but are considerably higher, on average, for cells requiring more protection. We also show that noise infusion has little effect on the rank order of state-, country-, or industry-level outcomes and only small impacts on revisions or on implicit growth outcomes.

The primary upside and primary downside of the switch to noise infusion are both clear. Using noise infusion instead of cell suppression increases the number of estimates that are published (that is, not suppressed) at the cost of some distortion—small in most cases—of the published estimates. Increasing the number of estimates published has different implications for different data users. For the most motivated and technically sophisticated users, it eliminates the need to undertake a complicated process of deriving and using range estimates—which, in some cases, might be very wide—for suppressed cells. For the more typical user, it provides estimates for cells that would otherwise be unknown.

A related benefit to increasing the number of estimates within a given set of tables that is published under an EZS noise infusion approach to SDL is that noise infusion can tip the balance of a statistical agency's evaluation of whether to expand the number of tables published. In particular, the prospect of publishing tables dominated by suppressed cells can strongly deter an agency from publishing additional tables, even when the requisite underlying source data are already available from the survey and when there is considerable interest from data users. The use of EZS noise infusion can improve the calculus in favor of increasing the number of tables provided, thereby making more complete use of data collected on surveys and increasing utility for data users. BEA views this benefit as a strong argument in favor of switching to noise infusion and, because it is continually looking for opportunities to expand its statistical offerings, expects that the switch to EZS noise infusion will eventually result in such an expansion.

One other potential downside to using noise infusion is that data users do not necessarily have a strong sense of how much noise is applied to records and of how much the noise distorts the estimates. This is particularly the case when the statistical agency provides only limited information in these areas. BEA has still not determined exactly how much information can be provided to data users without compromising the confidentiality protection provided to data users. However, it is likely that BEA will provide such information conservatively at first. Over time, it may determine that it can increase the type and/or extent of information provided.

BEA welcomes feedback from data users on its adoption of EZS noise infusion, both for the NFDI statistics as well as other BEA survey-based statistics. Of particular interest is feedback on how best to provide summary information on the effects of noise infusion and how much information is needed for data users to feel confident about using the statistics. BEA also welcomes suggestions as to which specific applications of noise infusion (for instance, specific noise factor distributions) would be most appropriate, either in general or for a given statistical product.