

# **Balancing Confidentiality Requirements with Data Users' Information Needs**

Background Document Prepared by the BLS Disclosure Review Board

for the BLS Data Users' Advisory Committee

November 8, 2012

## **0. Overview**

For this session of the BLS Data Users' Advisory Committee (DUAC), the BLS Disclosure Review Board (DRB) is requesting guidance on practical ways in which to explore the impact that disclosure limitation policies and methods may have on BLS data users. Sections 1 through 4 of this document provide some general background on disclosure limitation methods that are relevant to BLS establishment survey programs. In particular, Section 1 reviews the primary aspects of the legal, regulatory and reputational environment that lead to standard disclosure limitation policies and methods; Appendix A presents supplementary material on BLS confidentiality pledges to respondents. Section 2 summarizes some general ideas that inform current BLS publication practices related to disclosure limitation and data quality. Section 3 describes some relatively simple methods for disclosure limitation with tabular data, including cell suppression, cell aggregation and cell perturbation. Section 4 discusses current practices in some BLS programs; additional programs are discussed in Appendix B. Finally, Section 5 presents three sets of questions for DUAC members regarding (1) the general impact that disclosure limitation methods have on practical use of tabular data; (2) the reasons for the impact identified in (1); and (3) additional steps that BLS should take to explore the impact of disclosure limitation methods on the broader community of data users.

## **1. Elements of the Legal, Regulatory and Reputational Environment that Influence Disclosure Limitation Policies and Methods**

### ***1.1. What is statistical confidentiality?***

For the past four decades, the statistical community generally has used a definition of *confidentiality* that was given by the President's Commission on Federal Statistics (1971):

“[Confidential should mean that the dissemination] of data in a manner that would allow public identification of the respondent or would in any way be harmful to him [her] is prohibited and that the data are immune from legal process.”

### ***1.2. What is a statistical disclosure?***

The statistical literature considers several definitions of “statistical disclosure.” For the current discussion, two definitions are of primary interest.

A commonly cited definition from 1993 is: *Disclosure* relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (*identity*

*disclosure*), sensitive information about a data subject is revealed through the released file (*attribute disclosure*), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (*inferential disclosure*)<sup>1</sup>.

Section 512 of the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) defines statistical disclosure as:

“Disclosure of Statistical Data or Information.—

1. Data or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in identifiable form, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent.
2. A disclosure pursuant to paragraph (1) is authorized only when the head of the agency approves such disclosure and the disclosure is not prohibited by any other law.
3. This section does not restrict or diminish any confidentiality protections in law that otherwise apply to data or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes.”

Disclosure and confidentiality are potentially of concern in both household and establishment surveys. Most BLS household survey data are collected by the Census Bureau under Title 13. Confidentiality issues for those data are handled primarily by the Census Bureau. This paper focuses on BLS surveys that collect data from establishments or companies whether directly by BLS staff, or through agents (i.e., contractors or States).

### ***1.3. Confidentiality Issues Encountered by BLS Statistical Programs***

BLS programs take very seriously their responsibilities to protect respondent confidentiality. Three important factors are the following.

- 1) Compliance with applicable laws and regulations. These include the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA); the Privacy Act (5 U.S.C. 552a); the Information Quality Act (P.L. 105-554 § 515); applicable state Laws (for Federal/State programs); the Federal Statistical Confidentiality Order from 1997; and various BLS Commissioner’s Orders. Although all of these are important, CIPSEA has received special attention in the past decade. CIPSEA was a major step forward in providing legal protections for BLS data, and for providing confidentiality assurances to BLS respondents.
- 2) Respondent cooperation. Almost all BLS statistical programs rely on voluntary respondent cooperation. Consequently, it is important to ensure that respondents are comfortable that sensitive information they give to a statistical program won’t be revealed in a way that jeopardizes their interests. If not, respondents are less likely to cooperate with that program and perhaps other statistical programs.

---

<sup>1</sup> Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics,” edited by George T. Duncan, Thomas B. Jabine, Virginia A. de Wolf; published by the Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, D.C., 1993

- 3) Broader public perceptions. The general public perceptions of a statistical agency are enhanced by a clear understanding that the agency handles statistical data in a form that is consistent with both the letter and the spirit of commitments on respondent confidentiality.

Note that for the latter two concerns, the underlying reality of BLS practice is important, but issues of communication and perception are also important. Concerns would arise if respondents perceive they are at risk, or if the public perceives that respondents are at risk, even if technical features of confidentiality protection were entirely adequate.

#### ***1.4. Processes to Address BLS Confidentiality and Disclosure Issues, and to Meet Data Users' Needs***

BLS has a proud history of being very protective of sensitive respondent information, and of outreach to understand and address data users' needs. In the confidentiality area, all sample units receive explicit confidentiality pledges, as detailed in Appendix A. In addition, the BLS Disclosure Review Board (DRB) was chartered in 1999 to provide guidance on the policies and practices used by BLS programs to safeguard sensitive information provided by respondents. The DRB operates under the guidance of the BLS Disclosure Review Executive Committee (DREC), composed of BLS Senior Executives who rule on policy issues raised by the DRB.

In outreach to data users, BLS elicits input through broadly based groups like the BLS Data Users' Advisory Committee (DUAC). In addition, BLS obtains valuable input on program-specific and multi-program data users' needs through a wide range of outreach and consultation activities. For example, the Quarterly Census of Employment and Wages (QCEW) program receives a near continuous stream of feedback from users regarding their interests in additional data products from QCEW and improved ways for them to use our current products. This feedback comes from routine data user contacts by phone and e-mail to the BLS offices, as well as contacts through our cooperating state partners. In addition, QCEW staff frequently attend meetings and conferences of organizations whose members are users of QCEW data. At these meetings, such as of the Nation Association of Counties ([naco.org](http://naco.org)) and the Council for Community and Economic Research ([c2er.org](http://c2er.org)) staff gather information on both current use, and future desires for QCEW data. For the Consumer Price Index (CPI) program, CPI staff routinely interact with the public to answer customer inquiries and make presentations and seminars. The CPI public website provides contact information (email addresses and telephone numbers) for a variety of CPI programs. CPI staff review and respond to each customer inquiry within 3 working days. Beyond responding to public inquires, CPI staff provide presentations and seminars to outside organizations to gain feedback and better understand how customers use CPI data from the BLS public website. For the Occupational Safety and Health Statistics (OSHS) programs, OSHS staff learn about data users' needs primarily through direct interaction with the public. Staff economists in all OSHS programs are required to answer help line calls and E-mails from data users, and are routinely assigned special queries to process and check for public users. Furthermore, national office staff are sent to several conferences run by safety and public health organizations to present findings and staff an information booth on an exhibition floor. Cooperation with state and regional offices at these events and regular interaction during survey operations are another channel for indirect feedback at the local level from the public. Other BLS programs carry out similar outreach efforts for data users.

Efforts to meet data users' needs while remaining in compliance with confidentiality restrictions will invariably require a balanced consideration of multiple factors. Any decision to publish additional data is conditional on having assurance that the additional publication will not jeopardize quality, and will not compromise BLS confidentiality pledges. The remainder of this paper outlines some of the important factors that contribute to balanced decisions in this area, and requests input from DUAC on these factors. In the disclosure avoidance literature, the effort to balance disclosure risk and data usability is referred to as the Risk/Utility or R/U curve<sup>2</sup>. The current paper will not cover technical features of risk-utility curves in additional detail, but it is worthwhile to bear in mind the general need to balance disclosure risk and data utility.

## 2. General Approaches to Disclosure Limitation

BLS establishment survey programs publish large amounts of data, generally in tabular form. The structure and volume of these data publications arise from a combination of stakeholder requests; budgetary decisions; and (when feasible) historical patterns of previous publications. Published tables often include many dimensions; specific dimensions vary across programs, but may include classification factors defined by geography, industry, occupation, product and consumption items, and illness/injury types. In principle, the intersection of several of these classification factors can lead to tables with large numbers of cells. In practice, the BLS publishes data for some, but not all of these cells. A decision not to publish data for a given cell generally involves concerns related to (a) insufficient precision (based on, e.g., large relative standard errors for the cell); (b) resource limitations (relative to priorities among the primary data users); and (c) disclosure limitation concerns.

Issues (a) and (c) often involve closely related constraints. For example, there is often some degree of commonality between policies, practices, and procedures that affect the confidentiality of reported data, on the one hand, and the policies, practices, and procedures that produce reliable estimates, on the other hand. Many programs have standards for the minimum number of contributors to an estimate, and/or the minimum size of the population the estimate is for. Thus, reliability rules provide confidentiality protection in a number of programs.

More specifically for tabular data, many cells or potential cells have only one respondent, or population member, so if the cell aggregates were to be released, that would correspond exactly to the individual establishment's reported or imputed values. Also a large number of cells have only two or three respondents, or are dominated by the contributions of just one of the establishments. Cells dominated by two establishments are vulnerable because either respondent might deduct their own contribution from the cell aggregate to approximate the other dominant contributor's values. In very small establishments, cell information might actually be attributable to a single individual. In some surveys, a large number of observations are collected, but all or most of the observations for a particular item or characteristic might come from the same company. In those cases disclosing the estimate might be perceived as disclosing the company's

---

<sup>2</sup> See Duncan, Keller-McNulty, Stokes: Disclosure Risk vs. Data Utility: The R-U Confidentiality Map (<http://www.heinz.cmu.edu/research/122full.pdf>), and Karr, Kohnen, Oganian, Reiter, and Sanil: A Framework for Evaluation the Utility of Data Altered to Protect Confidentiality (<http://www.jstor.org/stable/pdfplus/27643781.pdf?acceptTC=true>)

information. At the other extreme, there are a number of programs whose finest cells reflect the contributions of a large number of contributors and are fundamentally safe to disclose.

Consequently, in exploration of issue (c) for tabular data, the statistical literature defines a cell to be “sensitive” if release of data from that cell allows identification of one or more respondents; or discloses sensitive information for one or more respondents. The current discussion will focus on publication issues related to disclosure limitation, but it should be noted that non-publication issues (a) through (c) are often closely related. For example, a cell that contains a small number of sample units, or contains one or two dominant sample units may be non-publishable both due to its large relative standard error, and due to disclosure limitation concerns.

### **3. Options for Disclosure Limitation with Sensitive Cells**

#### ***3.1. Three Options: Cell Suppression, Cell Coarsening and Cell Perturbation***

For the reasons cited in Sections 1 and 2 above, it is not feasible to publish standard direct estimates for cells that are identified as sensitive. Instead, a statistical program may consider three realistic options.

***Option 1:*** Do not publish data for the cell at all. Cells handled through Option 1 are called “suppressed cells.” For such cases, data are replaced by a “flag” indicating that an estimate could not be published for the specified cell. There is a large technical statistical literature on methods for identification and suppression of sensitive cells. The current paper will not explore that literature in detail, but it is worthwhile to note that due to additivity constraints within many tables (e.g., entries in a given row or column will sum to a separately published “row total” or “column total”), identification of one cell as “sensitive” may require suppression of data from one or more additional cells. The resulting changes (referred to as “secondary suppression” or “complementary suppression”) can have an important impact on the overall pattern of non-publication of cells in a given table.

***Option 2:*** Combine cells in a way that the resulting published combined-cell estimates are not sensitive. Under Option 2, users interested in data from a specific suppressed cell are referred to data from the published higher level/coarser cells. For example, a statistical program may seek to publish many of its estimates at a four-digit NAICS level (over 300 industry categories), but some four-digit NAICS cells may have estimates that are sensitive, and thus cannot be published. For these cells, the program may refer data users to related estimates published for coarser level (e.g., three-digit or higher) NAICS cells.

***Option 3:*** Cell perturbation. In recent years, some statistical agencies (e.g., the Census Bureau and the National Agricultural Statistics Service) have developed and/or implemented a set of “cell perturbation” methods to address disclosure limitation concerns; and have used these methods for some of their statistical publications<sup>3</sup>. The main idea of cell perturbation is that for a given sensitive cell, instead of publishing the

---

<sup>3</sup> For general background on cell perturbation methods, see, Evans, Zayatz, Slanta of Census, Using Noise for Disclosure Limitation of Establishment Tabular Data, (1996), <http://www.census.gov/srd/papers/pdf/bte9601.pdf>.

standard direct estimate based on observed data, one would publish a number equal to the sum of the direct estimate plus a “noise” term.

### ***3.2. Illustrative Example from the Quarterly Census of Employment and Wages (QCEW)***

To illustrate some features of Options 1 through 3, we consider some simple examples based on the Quarterly Census of Employment and Wages (QCEW). The QCEW publishes data on establishment counts, employment, and wages for detailed tables defined by industry and geographical area. For a more general description of QCEW, see Section 4.1 below.

Taken together, QCEW publications involve over four thousand areas and over two thousand industries. The cells defined by the relevant combinations of areas and industries include 3.5 million active combinations. However, due to disclosure limitation issues, approximately 60 percent of these 3.5 million cells currently are suppressed. For example, Table 1 presents data from the broadcasting industry in Clarke County, Alabama for the years 2010 and 2011. Note especially that for 2010, establishment counts were published for some three, four, five and six-digit NAICS levels. However, employment counts and average annual pay for 2010 was published only at the three-digit level. In addition, for 2011 only establishment counts were published, and the employment counts and average annual pay was suppressed for all of the cells in this example. Thus, Table 1 involves a combination of Options 1 and 2 as outlined above.

To illustrate Option 3, Tables 2 and 3 present hypothetical data for an industry-area combination that includes a total of fourteen establishments. The leftmost column presents progressively finer levels of industrial classification involving a specific industry A and sub-industries A1 and A2. The next three columns present hypothetical raw data for establishment counts, employment counts and average annual pay. The final three columns present the table that would be published after application of standard cell-suppression techniques. Table 3 has the same general organization as Table 2. The only difference is that its final three columns present data that could be published after application of the cell-perturbation methods described in Section 3.1. Note that the establishment counts remain as given previously, but that values of the employment counts and average annual pay are different from those given in the raw data.

**Table 1:  
Example from the Quarter Census of Employment and Wages:  
Clarke County Alabama (ND = Not Disclosed)**

		2010	2010	2010	2011	2011	2011
NAICS	Industry	Number of Establishments	Employment	Average Annual Pay	Number of Establishments	Employment	Average Annual Pay
515	Broadcasting, except Internet	6	18	26755	6	ND	ND
5151	Radio and television broadcasting	2	ND	ND	2	ND	ND
51511	Radio broadcasting	2	ND	ND	2	ND	ND
515112	Radio stations	2	ND	ND	2	ND	ND
5152	Cable and other subscription programming	4	ND	ND	4	ND	ND

**Table 2:**  
**Hypothetical Example of Cell Suppression from QCEW Data:**  
**Comparison of Raw Data (Left Columns) with Cell Suppression (Right Columns)**

	Raw Data			Cell Suppression		
Industry	Number of Establishments	Employment	Average Annual Pay	Number of Establishments	Employment	Average Annual Pay
Total	14	52	43,051	14	52	43,051
Industry A	7	30	42,212	7	ND	ND
Sub-Industry A1	3	6	18,317	3	ND	ND
Sub-Industry A2	4	24	48,186	4	ND	ND

**Table 3:**  
**Extension of Hypothetical Example of Cell Suppression from QCEW Data:**  
**Comparison of Raw Data (Left Columns) with Perturbed Data (Right Columns)**

	Raw Data			Cell Suppression		
Industry	Number of Establishments	Employment	Average Annual Pay	Number of Establishments	Employment	Average Annual Pay
Total	14	52	43,051	14	53	43,093
Industry A	7	30	42,212	7	29	42,661
Sub-Industry A1	3	6	18,317	3	7	16,485
Sub-Industry A2	4	24	48,186	4	22	50,989

#### **4. Current Disclosure Limitation Practices in Selected BLS Programs**

This section summarizes current disclosure limitation approaches for four BLS programs: the Quarterly Census of Employment and Wages (QCEW) program (with emphasis on data by industry); the National Compensation Survey (NCS) program (with emphasis on data by occupation); the Survey of Occupational Injuries and Illnesses (SOII), and the Consumer Price Index (CPI). These summaries place primary emphasis on:

- 1) the type of data in the program;
- 2) currently released data and the associated disclosure risk;
- 3) potential new or expanded data that could be released from the program;
- 4) the extent to which that new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react; and
- 5) various aspects of how that program might trade reduced precision in its data releases for more detail.

The appendix includes similar reports for a number of other BLS programs.

##### **4.1. *Quarterly Census of Employment and Wages (QCEW)***

###### **4.1.1. The type of data in the program**

QCEW publishes employment levels, total wages and wages per employee, by industry and area. These are establishment/place of work measures, derived from the quarterly Unemployment Insurance (UI) tax filings made to the employers' state workforce agencies. The published data are stratified to the most detailed industry classification codes (North American Industry Classification System – NAICS 6-digit), and to the county level. Unlike primary economic indicators, these data are released approximately six months after the reference period – e.g. September release of first quarter data, and are subject to typically minor revisions over a period that ranges from three months to one year.

A real world example is 2010 Broadcasting industry data for Clarke County, Alabama. There were six establishments in this industry (NAICS 515), two in Radio (NAICS 515112) and four in Cable (NAICS 515210). BLS released data showing that there were 18 employees with an average weekly wage of \$515 in Broadcasting overall, but didn't release a more detailed industry breakout of employment or wages. It is easy to imagine the calculations that would allow a tight estimation of some of the individual establishment's employment or wages if more detailed data were released.

###### **4.1.2. Currently released data and the associated disclosure risk**

Cell suppressions, both primary and complementary, are done to limit the exposure of individual responses. Many cells have only one respondent, or population member, so if the cell aggregates were to be released, that would correspond exactly to the individual establishment's values. Also a large number of cells have only two or three respondents, or are dominated by the contributions of just one of the establishments. Cells dominated by two establishments are vulnerable because either of them might deduct their contribution from the cell aggregate to approximate the other dominant contributor's values. In very small establishments, cell information might actually be attributable to a single employee.

There seems to be general agreement that employers (establishments or companies) have an interest in keeping their detailed wage measures confidential. The somewhat obvious rationales are that the disclosed information might: 1) affect compensation bargaining between the company and its employees; 2) provide an advantage to current or potential competing companies; 3) increase local labor demand resulting in higher market wages; 4) affect bargaining position with clients; and 5) affect the relationship between insiders (employees or owners) and their families or communities. Note that it seems that for most rationales, and at least for those cited above, there is a countering interest by someone, for the information that the employer would be concerned about. Some of those interests are societal, such as higher wages or more jobs, and others, such as business competition, are individual.

On the other hand, it has been difficult to come up with generally applicable rationales as to why businesses would want to keep their overall employment levels confidential. Anecdotal situations can be imagined, but, it isn't clear how that might be expected to be translated into real world concerns. The biggest concerns would likely be only among the smallest of companies. For example, an insider questioning why there are four rather than three people on the payroll.

#### **4.1.3. Potential new or expanded data that could be released from the program.**

There is intense interest, conveyed directly from the public as well as through the cooperating state agencies for data tabulated at alternative geographic and industry levels – BLS itself only currently produces data for official State, county, and MSA areas and for official levels in the NAICS hierarchical structure. If these alternate aggregates are assembled from disclosure processed data, there is very little detail that is disclosed. On the other hand, special disclosure processing that takes into consideration the various levels and aggregates that have already been released is complex, error prone, and, because of potential failure in the method, might compromise the confidential expectations. BLS has generally avoided that type of processing for a number of years.

#### **4.1.4. The extent to which that new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react.**

Alternative cell suppression techniques are available. More robust complementary processing could be introduced that would both increase suppression and reduce utility. An alternative might be to consider employment as not sensitive and thus not suppressed. This alternative would pose greater risk – a least an appearance of not protecting the data – but offer greater utility to data users. A third dimension in the trade off might be to offer perturbed or less-exact data, but with reduced or eliminated suppression. In this case the utility would be greater for some users (perturbed data being better than suppressed data) but reduced utility for others (perturbed data being worse than exact data).

#### **4.1.5. Various aspects of how that program might trade reduced precision in its data releases for more detail.**

One possibility would be to eliminate the finest industry detail. How might that work? Sometimes that fine detail poses complementary risks for the higher level aggregates. Less fine detail would result in substantially higher overall disclosure rates. Another possibility would be treating employment as not-sensitive, assuming we met our

legal obligations to warn data providers about that treatment. That is, would it be expected to affect data collection?

A third possibility would be one of various types of micro data perturbation, possibly combined with limited data synthesis – the perturbations or synthesis parameters be derived from reported data. For general background information on a perturbation technique getting greater usage, see: Evans, Zayatz, Slanta of Census, Using Noise for Disclosure Limitation of Establishment Tabular Data, (1996), <http://www.census.gov/srd/papers/pdf/bte9601.pdf>.

A fourth possibility is to assess or reassess the sensitivity of wage data reported to BLS. It may be that BLS is providing too little or too much protection to that data. If too much, that inherently reduces the utility and detail available to data users.

## **4.2. National Compensation Survey (NCS)**

### **4.2.1. The type of data in the program**

The NCS is a sample of jobs chosen from a sample of establishments. Data is collected on pay and benefits, including costs, access and participation rates, and various benefit plan details.

The NCS publishes a variety of estimates. There are 5 main products:

- **Employment Cost Index (ECI):** Tracks changes in compensation costs. Estimates include the Indexes and their 3-month and 12-month percent changes, for total compensation, wages and salaries, and total benefits. Twelve-month changes are published for health insurance.
- **Employer Costs for Employee Compensation (ECEC):** Estimates include mean hourly cost per worker, for: total compensation, wages and salaries, total benefits, 6 benefit groups, and 18 individual benefits.
- **Incidence and Provisions of Benefits:** Estimates include benefit access and participation rates, the percent of participants with a given plan-provision, employer and employee shares of a cost or premium, and various means (premiums, deductibles, leave, etc.). Most estimates are the percent of workers in a group that lie inside one of its sub-groups. The “group” and “sub-group” definitions vary: for example, it could be all workers, all with access to a specific benefit, all who participate, or all with a given plan type or provision.
- **Pay Agent Deliverables:** Special wage estimates for the Federal Salary Council, which are based on a model that uses data from both the NCS and the Occupational Employment Survey (OES). The OES, with its larger sample size, provides the backbone of the model. Yet the NCS collects data on “generic levels” (designed to mimic federal GS levels) whereas the OES does not, so the NCS data is required. The OES and NCS inputs are private industry and state and local government data, so estimates are benchmarked to the federal employment distribution.
- **Special Tabulations:** These are prioritized based on requestor. Not all are computed due to resource constraints. Reliability and confidentiality concerns influence what is released.

### **4.2.2. Currently released data and the associated disclosure risk**

In general, the sensitivity of NCS publication cells is quite low because the cells are quite large. Disclosure avoidance techniques are applied that ensure respondent data are protected.

#### **4.2.3. Potential new or expanded data that could be released from the program.**

Far more estimates are computed than published. There are several factors that determine what is published, such as relevance, size of the cell, size of the publication table, and whether the data pass all reliability and confidentiality tests. In some cases the NCS may be able to publish smaller cells or breakouts than normal, yet we still prefer they meet our reliability standards, and insist they meet our confidentiality standards.

#### **4.2.4. The extent to which that new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react.**

Most large cells are already published, so any expansion would most likely occur for smaller cells, which intuitively would tend to have higher sensitivity. But our confidentiality tests are sufficient to protect the data.

#### **4.2.5. Various aspects of how that program might trade reduced precision in its data releases for more detail.**

It is doubtful that we would use less reliable estimators, just so we can lower cell sensitivity enough to be able to publish more data.

### **4.3. Survey of Occupational Illnesses and Injuries (SOII)**

#### **4.3.1. The type of data in the program**

The SOII is a mandatory participation survey that collects OSHA-required occupational injury and illness data from a sample of establishments. Establishments selected for the survey are required to transcribe the information recorded on their OSHA log (Form 300), OSHA summary (Form 300A) and individual case forms containing detailed information on work-related injury and illness incidents where employees missed work (Form 301).

SOII estimates are divided between two major news releases with supplemental tables:

- **Annual Summary Estimates (AS):** Estimates of numbers of incidents and rates by NAICS industry. Case counts and rates are produced for cases involving days away from work (DAFW), cases involving job transfer or restriction (DJTR), and other recordable cases (ORC). A commonly used combination is the DART rate, which is an incidence rate that includes both DAFW and DJTR data. Various tables showing incident counts and rates by broad size classes of establishments by employment and tables for incident counts and rates for injuries only or illnesses only are also produced.
- **Case and Demographic Estimates (C&D):** Estimates of DAFW incidents and rates by various worker characteristics and incident circumstances. Worker demographics include race, age, occupation, gender, and job tenure. Case circumstances include time of shift, time of incident, NAICS industry, nature of injury or illness, part of body affected, source contributing to incident, and event.

Various two-way cross-tabulations and supplemental tables such as musculoskeletal disorder tabulations are also produced.

- **Quartile, Resource, Supplemental, and Special Tabulations:** In addition to the tables included in the news release, larger and more detailed resource and supplemental tables are posted on the BLS website.<sup>4</sup> Quartile tables showing coarse distributions of highest and lowest rates and case counts in published industries are generated for the AS estimates.<sup>5</sup> Cross-tabulations that are not part of regular SOII outputs are provided to the public on a first-come first-served basis as staff has time available to produce the tables. Special tabulation data must pass the same reliability and confidentiality standards used for regular SOII products.

#### 4.3.2. Currently released data and the associated disclosure risk

Disclosure risk is low due to relatively large survey samples and magnitudes of estimates.

- **Annual Summary:** Estimates rounded to the nearest hundred and rates per 100 full time workers are reported by industry at the state and national geographic levels of aggregation. Most state-industry combinations have sufficiently many eligible establishments such that it is not feasible for non-BLS entities to determine which establishments were surveyed and which establishments responded. Whenever state-industry cell employment is concentrated enough in a few dominant firms that there could be identification risk, State partner agencies may obtain informed written consent from such firms to allow publication of the estimates.
- **Case and Demographics:** Estimates rounded to the nearest ten, rates per 10,000 workers, and median numbers of days away from work are published for all tabulations. For example, in 2010 there were 563,850 estimated private industry DAFW cases nationwide involving males, the incidence rate was 127.6 cases per 10,000 workers, and the median number of days missed was 9. Due to the more sensitive nature of the data because it involves individual workers' health information, BLS suppresses estimates that round to less than 20 or have too few unweighted cases contributing to the estimate. This protects injured and ill workers whose combination of characteristics or circumstances would otherwise make their case easy to isolate in our data.
- **Quartile, Resource, Supplemental, and Special Tabulations:** The disclosure risks involved in these data products are identical to the risks for the standard products produced by BLS and thus carry the same protections and procedures.

#### 4.3.3. Potential new or expanded data that could be released from the program.

Participating States work with BLS National Office to decide which industries are targeted relatively more heavily in sample selection. This allows the States to ensure that industries they desire estimates for will receive enough sample allocation to be published. Most end data users are interested in prominent industries or ones where there are many injuries or illnesses. Most data requests from public health or safety researchers not

---

<sup>4</sup> For example, the Case and Demographic resource and supplemental tables are available at: <http://www.bls.gov/iif/oshednew.htm>

<sup>5</sup> This is on the same page as the Annual Summary resource and supplemental tables: <http://www.bls.gov/iif/oshsum.htm>

already served by existing tables can be satisfied through our special tabulations process (few people are interested in categories where “nobody is getting hurt”). Large case and establishment counts provide good disclosure/identity protection for such tabulations.

#### **4.3.4. The extent to which that new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react.**

Published estimates based on too few unweighted cases or establishment reports could put firms who have experienced unusual and/or publicized events at risk of being identified. For example, suppose there were a single mass food poisoning in an industry where such poisonings are rare. A published poisoning estimate in this industry would inform a public data user that the establishment with that mass poisoning was surveyed and in the dataset. Moreover a two way special tabulation between food poisonings and some other characteristic like race or gender would reveal information not just about the industry as a whole, but obviously about that specific incident and the employees involved. Normally, a rare event with no other observations at any other establishment in the cell to introduce uncertainty as to the identity of the firm where this happened would prevent publication of the estimate to prevent ex post identification.

High response rates in SOII and a special waiver process that lets states obtain informed consent from respondents offer states ways to keep cells publishable. Estimates that remain suppressed are generally cells in which there are simply too few companies or too few cases to publish estimates which still protect our respondents. For example, one large service industry at the three digit NAICS level is so dominated by two competitors that despite more than 1000 worksite level responses in the cell, those two companies contributed more than three quarters of the usable responses nationwide received by BLS. The largest firm contributed roughly half the responses and another quarter of the responses were the second largest firm. In particular, the largest worksites in the industry were overwhelmingly from those two large companies. It is difficult to conceive of any way in which an estimate with enough precision to be useful could be published (especially for size classes) without making it possible for the second largest company to know more about the safety and health experience of its larger competitor. Furthermore, it would not be hard for regulators to use information about “large establishments” to conclude something about those two specific companies.

#### **4.3.5. Various aspects of how that program might trade reduced precision in its data releases for more detail.**

It is unlikely that data users will prefer reduced precision or confidentiality relaxing to obtain estimates not already available. Where we are unable to publish in very small industries or for rare case types, reduced precision would render estimates to be of little use because the magnitudes for “small” cell estimates would be swamped by relatively large errors.

### **4.4. *Consumer Price Index***

#### **4.4.1. The type of data in the program**

The Consumer Price Index (CPI) is a Principal Federal Economic indicator that measures price change over time for urban consumers for a market basket of consumer goods and

services. The US government uses the CPI to adjust income payments for wages, Social Security, Federal Civil Service retirees and survivors, food stamps, and school lunches.

The CPI is based on 4 surveys: the Commodity and Services (C&S) price survey, the Housing price survey, the Telephone Point of Purchase Survey (TPOPS) for a sample of retail establishments, and the Consumer Expenditure Survey (CE) for weighting. Each of these surveys has unique disclosure risks given the source of the data, and how it is used in the construction of the CPI. This evaluation is for surveys collected and published as final results; the TPOPS and CE surveys are conducted by Census, and therefore excluded from this evaluation

BLS economic assistants (EAs) conduct the C&S and Housing price surveys, which track prices over time. These surveys' authorization and confidentiality is based on the SO-1291 Confidential Burden Statement, which states that participation is voluntary, that the information will be confidential, and that the information will only be used for statistical purposes. The EAs request permission from the manager of the outlet to price a set of unique C&S surveyed Items, which are specific goods or services that can be purchased by a general consumer. In the case of Housing surveyed Items, the EA requests permission from the head of the household. Although advertised price quotes are not a disclosure risk, the CPI is obligated to protect the confidentiality of the retail establishment, including brand name, and housing unit information.

The headline products released by the CPI each month are for the urban population (CPI-U), the wage earner population (CPI-W), and the chained CPI for the urban population (C-CPI-U).

#### **4.4.2. Currently released data and associated disclosure risk**

Estimation of the CPI is divided into two stages. Lower-level estimation aggregates about 100,000 C&S and 7,500 Housing price quotes each collection month into 8018 elementary Item and Area cells, or basic level indexes. Area cells are made up of 87 geographic Primary Sampling Units (PSUs). Upper level estimation then aggregates elementary level cells into aggregate indexes such as the U.S. City Average All items index.

Publication status of aggregate CPI indexes is determined in a two step process. First, aggregate indexes are classified as eligible or ineligible for publication based on the existence of price and weight data, historical standards, and the variance estimate for the series.<sup>6</sup> Then, the CPI evaluates the "adequacy ratio" to determine which specific aggregate indexes will be published for a given month. The adequacy ratio is a measure of sufficiency of the collected sample within the aggregate cell, and is equal to the number of individual cells within that aggregate level cell that contain collected prices. If approximately 50% or more of the (weighted) Items and or geographic areas within that aggregate group contain collected prices, then the CPI will publish the price index for the aggregate level. As an example, a national banana index is made up of 38 unpublished area banana indexes. If the areas were weighted equally, the national banana index would be deemed adequate for publication if at least 19 of the individual cells contain good prices. Note, geographic areas are not in fact weighted equally; therefore prices are

---

<sup>6</sup> For additional information about CPI publication standards see Grandits: Publication strategy for the 1998 revised Consumer Price Index. Monthly Labor Review, 1996. (<http://www.bls.gov/mlr/1996/12/art4full.pdf>)

needed from Areas that represent at least 50% or more of the national Area weight in order to publish a national banana index. It is rare that CPI-U aggregate indexes are not published due to insufficient sample of collected price quotes.

BLS produces a host of products that are listed below, but acknowledges the limitations of experimental indexes such as the CPI for the Elderly population (CPI-E). Limitations about measurement can be classified as either sampling errors, or non sampling errors. Sampling errors occur due to the sampling of outlets and items. If the CPI had infinite resources, then retail prices could be evaluated for the entire index population. The CPI measures sampling error to demonstrate how the CPI allocation of the current sample maximizes the accuracy of the index given the funds available. A secondary type of limitation in measurement is non sampling error. Non sampling errors are problematic because they can come from a number of sources and potentially cause bias, which is hazardous to the accuracy of the index. Sources of non sampling errors range from the respondent, to the data collector, to even the survey instrument or survey questionnaire.

Official CPI products include:

1. CPI-U, CPI-W, C-CPI-U, and Seasonally Adjusted CPI-U and CPI-W
2. CPI-U and CPI-W Relative Importances
3. Average Prices for individual goods or services such as Gasoline
4. Additional Data Available from News Release includes: Monthly Relative Importance, Standard Error, Effect on All Items, and Largest / Smallest Change.
5. Variance Estimates for Changes in the CPI
6. Department Store Inventory Prices employ the retail inventory method (LIFO).

Experimental CPI products include:

1. CPI-U Research Series (CPI-U-RS) attempts to estimate what the measured rate of inflation in the CPI for all urban consumers (CPI-U) would have been had the methods now used been in effect since 1978.
2. Experimental Indexes produced within the CPI production system like the CPI-E which is designed to represent the elderly population
3. Experimental Indexes produced outside of the CPI production system like the US-CPI-Harmonized Index of Consumer Prices (HICP)

#### **4.4.3. Potential new or expanded data that could be release from the program**

When a customer requests non-published information the CPI evaluates the feasibility of producing a special tabulation or the feasibility of the researcher accessing micro data. If the request is from a student or government researcher for the sole purpose of analytical research, then the request is evaluated by the BLS Micro Data Access Review Board. If approved, then the researcher becomes an Agent of BLS and agrees to terms and conditions of the BLS Agent Agreement.

#### **4.4.4. The extent to which new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react.**

For the C&S surveyed Items, the potential damage to respondent confidentiality depends on the environment of prices. If multiple outlets offer similar prices for the same

elementary Item and Area combination, then it is unlikely that the confidentiality will be breached. If a unique Item is priced within an Area where there are few outlets that offer the comparable good or service, then confidentiality would need to be preserved. Lack of outlet sample and excessive weight for any given price quote are disclosure risks that would need to be evaluated and resolved at the lower levels. BLS preserves the confidentiality of the outlet, price, and even lower level specification of products, regardless of the data shared by the company such as with advertising programs.

#### **4.4.5. Various aspects of how that program might trade reduced precision in its data releases for more detail.**

The CPI evaluates each customer request on an individual basis, as highlighted above. The CPI makes every effort to fulfill the customer request. If the request is not feasible, the CPI will suggest alternatives, such as an aggregate level summary addressing the request. To date, the CPI does not perturb data or create noise at lower levels in order to publish more cells with less precision.

### **5. Questions for the Data Users' Advisory Committee**

Ideally, a statistical agency would be able to meet the information needs of all of its potential data users. However, as noted in Sections 1 and 2, publication decisions for specific tables and cells will generally require a statistical agency to balance a complex set of factors involving stakeholder needs, resource limitations, precision requirements and disclosure limitation concerns. To explore the disclosure-limitation dimension of this balance in additional detail, the BLS Disclosure Review Board is seeking guidance from the BLS Data Users' Advisory Committee on the following topics.

A. What is the current or potential impact on data users from the three disclosure limitation methods described in Section 3?

Option 1: Cell suppression - No data for that cell

Option 2: Cell coarsening - Value depends on relevance of data from the "coarser" cell

Option 3: Cell perturbation - Value depends on perceptions of "real data" vs. "perturbed data"

B. What are the primary reasons for the effects identified in (A)?

B.i. If possible, please provide concrete examples to illustrate the reasons for these effects.

B.ii. What subject-matter factors will dominate the trade-offs between Option 2 (coarser cells) and Option 3 (perturbed cell data)? For instance, cell perturbation may be of strongest interest for cases in which (a) there are important practical differences (e.g., substantially different market dynamics) between finer-level cells and the corresponding coarser-level cells; and (b) users would be willing to accept data subject to a moderate amount of perturbation, as described in Section 3.

C. What are some recommended additional steps for BLS to explore the impact of Options 1-3 on the broader community of data users?

- 1) If we were to provide more detail, please explain why that detail has added value.
- 2) Would more detailed data be desired, if it were at reduced quality? (That quality reduction might come from either accepting increased standard errors, or utilizing the data perturbation methods described in Section 3.)

**Appendix A: BLS Use of the CIPSEA Pledge 2011**

Thirteen BLS programs use the same common CIPSEA Pledge:

- (CES) Report on Employment, Payroll, and Hours
- (CFOI) Census of Fatal Occupational Injuries
- (CPI Housing) Consumer Price Index Housing Survey
- (CPI C&S) Consumer Price Index Commodities and Services Survey
- (GGS) Green Goods and Services
- (GTP) Green Technologies and Practices
- (IPP) International Price Program – U.S. Export and Import Price Indices
- (JOLTS) Job Openings and Labor Turnover Survey
- (NCS) National Compensation Survey: Private industry forms only
- (OES) Report on Occupational Employment
- OES Green Technologies and Practices Forms Pre-testing
- (PPI) Producer Price Index Survey
- (SOII) Survey of Occupational Injuries and Illnesses

Other BLS programs use program specific pledges:

- Cognitive and Psychological Research Program
- (MLS) Mass Layoff Statistics – LMI Cooperative Agreement
- (NCS) National Compensation Survey: Government forms
- (NLSY79) National Longitudinal Survey of Youth 1979
- (NLSY97) National Longitudinal Survey of Youth 1997
- (QCEW) Quarterly Census of Employment and Wages– Labor Market Information –  
(LMI) Cooperative Agreement

TITLE OF SURVEY	CIPSEA PLEDGE
Common BLS CIPSEA Pledge	The Bureau of Labor Statistics, its employees, agents, and partner statistical agencies, will use the information you provide for statistical purposes only and will hold the information in confidence to the full extent permitted by law. In accordance with the Confidential Information Protection and Statistical Efficiency Act of 2002 (Title 5 of Public Law 107-347) and other applicable Federal laws, your responses will not be disclosed in identifiable form without your informed consent.
Cognitive and Psychological Research	<p><u>Conducted at the cognitive laboratory at BLS:</u></p> <p>In accordance with the Privacy Act of 1974, as amended (5 U.S.C. 552a), you are hereby notified that this study is sponsored by the U.S. Department of Labor, Bureau of Labor Statistics (BLS), under authority of 29 U.S.C. 2. Your voluntary participation is important to the success of this study and will enable the BLS to better understand the behavioral and psychological processes of individuals, as they reflect on the accuracy of BLS information collections. The BLS, its employees, agents, and partner statistical agencies, will use the information you provide for statistical purposes only and will hold the information in confidence to the full extent permitted by law. In accordance with the</p>

TITLE OF SURVEY	CIPSEA PLEDGE
	<p>Confidential Information Protection and Statistical Efficiency Act of 2002 (Title 5 of Public Law 107-347) and other applicable Federal laws, your responses will not be disclosed in identifiable form without your informed consent.</p> <p><u>Cognitive and Psychological Research conducted outside of the BLS laboratory:</u></p> <p>Current approved pledge for that survey (if it has an approved OMB number) or the above Privacy Act Statement.</p>
<p>Mass Layoff Statistics (MLS) – LMI Cooperative Agreement</p>	<p>Upon receipt by the BLS of UI data for the MLS program, the BLS will use the UI data for exclusively statistical purposes and will hold this information in confidence to the full extent permitted by law.</p>
<p>National Compensation Survey Government forms</p>	<p>The BLS publishes statistical tabulations from this survey that may reveal the information reported by individual state and local governments. Upon your request, however, the BLS will hold the information provided on this survey form in confidence.</p>
<p>National Longitudinal Survey of Youth 1979 (NLSY79)</p>	<p>We want to reassure you that your confidentiality is protected by law. In accordance with the Confidential Information Protection and Statistical Efficiency Act of 2002, the Privacy Act, and other applicable Federal laws, the Bureau of Labor Statistics, its employees and agents, will, to the full extent permitted by law, use the information you provide for statistical purposes only, will hold your responses in confidence, and will not disclose them in identifiable form without your informed consent. All the employees who work on the survey at the Bureau of Labor Statistics and its contractors must sign a document agreeing to protect the confidentiality of your data. In fact, only a few people have access to information about your identity because they need that information to carry out their job duties. Some of your answers will be made available to researchers at the Bureau of Labor Statistics and other government agencies, universities, and private research organizations through publicly available data files. These publicly available files contain no personal identifiers, such as names, addresses, Social Security numbers, and places of work, and exclude any information about the States, counties, metropolitan areas, and other, more detailed geographic locations in which survey participants live, making it much more difficult to figure out the identities of participants. Some researchers are granted special access to data files that include geographic information, but only after those researchers go through a thorough application process at the Bureau of Labor Statistics. Those authorized researchers must sign a written agreement making them official agents of the Bureau of Labor Statistics and requiring them to protect the confidentiality of survey participants. Those researchers are never provided with the personal identities of participants. The National Archives and Records Administration and the General Services Administration may receive copies of survey data and materials because</p>

TITLE OF SURVEY	CIPSEA PLEDGE
	those agencies are responsible for storing the Nation’s historical documents.
National Longitudinal Survey of Youth 1997 (NLSY97)	<p>We want to reassure you that your confidentiality is protected by law. In accordance with the Confidential Information Protection and Statistical Efficiency Act of 2002, the Privacy Act, and other applicable Federal laws, the Bureau of Labor Statistics, its employees and agents, will, to the full extent permitted by law, use the information you provide for statistical purposes only, will hold your responses in confidence, and will not disclose them in identifiable form without your informed consent. All the employees who work on the survey at the Bureau of Labor Statistics and its contractors must sign a document agreeing to protect the confidentiality of your data. In fact, only a few people have access to information about your identity because they need that information to carry out their job duties. Some of your answers will be made available to researchers at the Bureau of Labor Statistics and other government agencies, universities, and private research organizations through publicly available data files. These publicly available files contain no personal identifiers, such as names, addresses, Social Security numbers, and places of work, and exclude any information about the States, counties, metropolitan areas, and other, more detailed geographic locations in which survey participants live, making it much more difficult to figure out the identities of participants. Some researchers are granted special access to data files that include geographic information, but only after those researchers go through a thorough application process at the Bureau of Labor Statistics. Those authorized researchers must sign a written agreement making them official agents of the Bureau of Labor Statistics and requiring them to protect the confidentiality of survey participants. Those researchers are never provided with the personal identities of participants. The National Archives and Records Administration and the General Services Administration may receive copies of survey data and materials because those agencies are responsible for storing the Nation’s historical documents.</p>
Quarterly Census of Employment and Wages (QCEW) – Labor Market Information (LMI) Cooperative Agreement	Upon receipt by the BLS of the QCEW files, the BLS will use the QCEW data for exclusively statistical purposes and will hold this information in confidence to the full extent permitted by law.

## **Appendix B: Description of Disclosure Limitation Approaches in Additional BLS Programs**

This appendix provides a general description of disclosure limitation approaches in additional BLS programs, following the same outline used for the programs described in Section 4.

### ***B.1. Current Employment Statistics State and Area (CES-SA)***

#### **B.1.1. The type of data in the program**

The Current Employment Statistics survey (aka, payroll survey) is a cooperative program conducted by the Federal Bureau of Labor Statistics, and State governmental agencies. The State and Area program of the payroll survey offers several pieces of economic information for State and Metropolitan areas. This information includes the following items: estimates of job growth or decline (all employees and production workers), of average weekly hours (AWH), of average hourly earnings (AHE), and of average weekly earnings (AWE). The payroll survey makes monthly estimates of establishment job level and growth for States and Metropolitan areas and industries within those areas. The State survey uses the same data used by the National payroll survey, but with slight differences in its estimation methodology. Statewide estimates use the same method used in National estimates when enough respondents are available for a sample-based estimates. When there is insufficient sample—such as in some metro area estimates—a model is employed that uses a combination of the following components:

1. Sample based estimate
2. Time series forecast of the employment
3. Other additional information

Given the additional components' inputs, it becomes more difficult for an intruder to disclose an establishment's reported data. While this summary's focus is on the State and Area program, its confidentiality concerns also apply to National CES estimates.

#### **B.1.2. Currently released data and the associated disclosure risk**

CES State and Area data releases consist of estimates for employment, average weekly hours, of average hourly earnings, and of average weekly earnings. The primary variable at risk is an establishment's employment. CES is a voluntary survey, and it is important to keep employment data confidential to maintain respondent confidence.

Secondary variables at risk are hours worked and earnings, and the CES program considers both items to be sensitive. Hours may give somebody an indication as to how many people a business employs. Large shifts in a business' hours could also indicate a shift in the number of employed persons. A similar statement can be made about payroll. The QCEW section has a longer discussion on sensitivity to payroll information, and it applies to CES as well.

Some industries in a metro area possess only a few firms. Since those areas possess desired economic data yet do not meet the requirements for a purely sample-based estimate, they are modeled. CES believes that the model provides enough protection to avoid disclosure.

For estimates that are not modeled, an estimate may fail disclosure tests because most of the employment or wages come from just a few of the involved businesses. In such a situation it may be possible for an individual to discern a business's payroll information.

### **B.1.3. Potential new or expanded data that could be released from the program.**

There is interest for additional industry and geographical information on payroll employment estimates. This interest comes primarily from the cooperating State agencies. CES carries out outreach efforts to gauge user interest in detailed industry and geographic estimates. Each year States review the estimates they publish for disclosure and reliability criteria, and they can suggest new or additional estimates that they would like to publish.

BLS currently publishes Statewide estimates for major NAICS sectors, Metropolitan estimates for major NAICS sectors, and Statewide estimates for various detailed industries (3 and 4-digit NAICS) when enough respondent establishments are available.

There may be additional interest in the AWH and AHE estimate series. These series are partially hampered by higher nonresponse rates for these data items. Additional AWH and AHE estimates are suppressed for confidentiality reasons. Relaxing confidentiality would increase the utility of these estimates, but the quality of these estimates would still be in question given the higher level of nonresponse.

### **B.1.4. The extent to which that new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react.**

Additional AWH and AHE estimates would require a change in the disclosure criteria. If the public knew that BLS intended to lower their disclosure-avoidance standards, then it's possible that fewer establishments would choose to respond. Fewer respondents could lead to suppressions due to quality considerations, so lowering disclosure standards could increase the number of suppressed cells and defeat the original intent.

### **B.1.5. Various aspects of how that program might trade reduced precision in its data releases for more detail.**

There are very few requests to CES to make special tabulations for unpublished estimates, so CES believes that it has met many of its users' data needs. At a very fine industry detail adequate sample—for either a modeled or a pure sample-based estimate—would not exist. In such a circumstance, the main component CES might use to model that estimate is publicly available—provided that the census data is also published.

State agencies are allowed to publish some estimates without official BLS approval, so they may be encouraged to publish additional estimates that do not meet BLS quality standards, so long as those estimates meet the proper disclosure avoidance criteria.

## ***B.2. Business Employment Dynamics Program (BED)***

### **B.2.1. Type of information.**

**Business Employment Dynamics** is a set of statistics generated from the Quarterly Census of Employment and Wages (QCEW) program. These quarterly data are measures of job gains and job losses from 1992 forward. These data help to provide a picture of the dynamic state of the labor market. Data types: gross job gains as total and by expanding and opening establishments; gross job losses as total and by contracting and closing establishments. Released as: employment; percent of employment; establishment count; percent of establishments; establishment births and deaths as counts and rates. These are private sector data that are released both seasonally adjusted and not seasonally adjusted, at the total level and by NAICS SuperSector. Additional tallies of the data show employment data for the private sector as a whole stratified by firm size, dynamic measures stratified by the size of the employment change, dynamic measures stratified by establishment age and NAICS SuperSector; and establishment survival by NAICS SuperSector and year of opening. Most of the data released at the National level is also released at the total private level without industry breakouts.

BED also publishes tables of employment by establishment size.

### **B.2.2. Disclosure Risk.**

The files are reviewed for instances where the data are substantially attributable to an individual establishment, and in those cases, the data are suppressed. However, at the levels currently disclosed, there is little chance for the dynamic information to be linked back to or re-identified to an individual establishment.

### **B.2.3. Variables at Risk.**

The only variables are employment levels and change.

### **B.2.4. BLS' current position on the Risk/Utility Curve – BED.**

Currently, BED has been fairly conservative in selecting the data levels they have been releasing. Most data meets the conservatively set disclosure standards. BED is interested in expanding the detail at which it publishes data, but, expects to explore disclosure options and data sensitivity further, before each expanded release.

### **B.2.5. Alternative positions of Risk/Utility – BED.**

BED perceives that it has been staying away from the margins of risk and utility with its data release. Data users have come up with new requests for BED data, and their interests need to be evaluated within the spectrum of possible further BED data products.

### **B.2.6. Questions for DUAC – BED.**

At this point in time, the best input from DUAC would likely be details about further ways BED data could be tabulated, and for what uses. BED wishes to avoid creating datatypes that would jeopardize future uses of BED data.

## **B.3. Occupational Employment Statistics (OES).**

### **B.3.1. The type of data in the program.**

OES is a stratified sample of business establishments that collects information on occupations and wages. The OES sample consists of 1.2 million establishments, allowing it to make detailed geographic estimates. While collecting an establishment's information on occupations, OES asks respondents to report employment by wage range. Wages are reported as the number of persons in a particular hourly wage bracket (\$7.51 - \$10.00, \$10.01 - \$15.00, etc.).

### **B.3.2. Currently released data and the associated disclosure risk.**

OES publishes National estimates of employment as well as mean and quantile wages for occupations, across the nonfarm economy and by industry. Industry estimates range from major industries (for example, Construction or Retail Trade) to detailed industries (such as specific industries within Construction or Retail Trade). OES also publishes occupational employment and wages by state, metropolitan area, metropolitan division and nonmetropolitan areas. OES data also allows States to publish Statewide and Metropolitan area estimates of similar information by industry.

Similar to other BLS programs, wages are a sensitive data item for respondents. Since the wages are reported by occupation, that information may become even more sensitive. Businesses may not want competing establishments to know the wages they pay for certain occupations. Respondents may also consider the number of employed persons per occupation to be sensitive.

### **B.3.3. Potential new or expanded data that could be released from the program.**

OES has required minimums with respect to the estimate quality and the number of responding establishments. OES could publish additional geographic, industry, and occupational information, but that publication would come at the expense of quality or disclosure risk. OES receives requests for state and area industry estimates but has left the discretion in releasing that information to the states, except in rare instances. The level of industry detail states find useful varies. OES does provide files of these estimates with primary suppressions and suggested secondary suppression. States are responsible for see that secondary suppressions are applied to any data that they release. (OES is currently reconsidering this position and considering issuing a standard data set of state, or state and area industry estimates.

### **B.3.4. The extent to which that new or expanded data release could threaten confidentiality or perceived confidentiality, including considerations of other publicly available material, and how data contributors might react.**

OES would screen any additional sets of estimates published using the same procedures in place for current publications. While this would lead to data sets with many suppressed entries, it should not present any significant increase in risk of disclosure.

### **B.3.5. Various aspects of how that program might trade reduced precision in its data releases for more detail.**

Releasing sets of estimate with the added granularity of industry would not have to affect the data quality or level of disclosure risk of what OES currently publishes. Many of the estimates in this data set could be of higher quality than some of the estimates that we currently publish for small domains. Of course on average the estimates in these data sets would be based on fewer

responses and so would tend to have higher variances. In addition limited program resources would mean that these estimates would not receive the same level of quality review received by our current product line.

#### ***B.4. PPI Summary***

##### **B.4.1 Background on Producer Price Index**

The Producer Price Index (PPI) is a family of indexes that measures the average change over time in selling prices received by domestic producers of goods and services for their output. The legal authority authorizing the collection of data for the PPI is, *Title 29, Section 2 of the Code of Laws of the United States of America*. PPI indexes measure price change from the producer's perspective. Prices received by producers may differ from those paid by buyers due to subsidies, sales and excise taxes, and distribution costs.

The PPI classification structures, which draw from the same pool of price information provided to BLS by cooperating company reporters, are as follows: industry classification by NAICS code, commodity classification (using a coding system unique to PPI) based on product similarity or material composition, and commodity-based stage-of-processing (SOP) classification.

The PPI draws its samples by industry, using a probability proportionate to size methodology where employment is used as a proxy for revenue. Industry membership is obtained from the Unemployment Insurance (UI) file (also known as the Longitudinal Database) which provides employment data on every company classified within a specific industry. For the companies selected for inclusion in the PPI survey who agree to participate, a sample of their individual products are chosen for inclusion in the PPI survey using probability proportionate to size based on individual product revenue. Participation in the PPI survey is voluntary, and companies cannot be compelled to participate or provide pricing information. The PPI does not typically publish indexes on a geographical basis, so there is no geographical component to the sample.

##### **B.4.2. Uses for the Producer Price Index**

The PPI is a Principal Federal Economic Indicator of overall price movement at the producer level. Some feel PPI movement may foreshadow subsequent price changes for businesses and consumers at the retail level. Therefore the President, Congress and the Federal Reserve often employ these data in formulating fiscal and monetary policy.

Another use of PPI data is by businesses for contract escalation. These contracts typically specify a transaction to occur at some point in the future, so it is often desirable to include an escalation clause in the contract to protect both parties from significant changes in input prices between the signing of the contract and the agreed upon delivery date.

Other uses of the PPI are as a measure of price movement for particular industries and products, to allow companies to compare input and output costs, as a deflator of other economic series, for LIFO (last-in, first-out) inventory valuation, and for forecasting.

##### **B.4.3. The Balance of Disclosure Issues**

In publishing data, the PPI tries to balance disclosure issues and quality standards. They realize that since the PPI is a voluntary survey, disclosing respondents' proprietary pricing information would be detrimental to both the respondent and to PPI's ability to collect data. Overall, the PPI would like to publish the most accurate data as possible on a timely basis, while respecting the confidentiality of their survey respondents.

The lowest level which PPI commits to publish industry data is at the Census 7-digit NAICS code for their national indexes (they generally do not publish by geography), although there are many cases where the PPI attempts to publish to a greater level of detail depending on the sample size of a particular industry. The PPI also commits to publish commodity indexes at the 8-digit level. If further product disaggregation is possible and the sample size allows, the PPI will publish to a greater level of commodity detail.

For each cell, the tests for adequacy are 3-fold: 1) minimum number of price quotes; 2) minimum number of reporting companies; and 3) weight test

If a calculated cell does not meet these 3 standards it will be suppressed. If an index fails only the weight test however, the PPI may still be able to publish it if they receive a waiver from the highly weighted company in the cell. Additionally, if an industry analyst feels the volatility of prices received for a certain cell is too great in a given month, and the analyst cannot verify the accuracy of the reported prices, they will suppress the publication of that cell index. Generally, if a cell is suppressed it is noted in publication tables with a dash (-) and there is a note for suppressions at the bottom of the tables.

Suppression Statistics – of about 5,000 industry indexes eligible for publication: PPI publishes approximately 82.8% of these indexes each month, suppressing the remaining 17.2%. Of these suppressed cells, typically only 1 or 2 cells are due to quality standards or analyst judgment. The remaining cells are suppressed due to confidentiality issues or because they do not meet the PPI 3-fold test for adequacy.

Special tabulations are often calculated upon request if they meet adequacy standards.

In light of the utility of PPI indexes as contract escalators, the PPI would rather not publish potentially confidential cells with noise. PPI customers need precise data for escalation purposes, and any noise would unnecessarily increase or decrease contracts escalation values. Instead they would rather direct the customer to go up the aggregation tree to a more general index to get needed information when a cell is unavailable.

Overall, PPI customers understand the need for confidentiality and the suppression of cells because they are dealing with a voluntary survey. Most users feel, that if they were in a similar situation, they would rather have the confidentiality of their proprietary pricing information protected over data availability.

## ***B.5. IPP Summary***

### **B.5.1. Background for International Price Program**

The International Price Program (IPP) produces **Import/Export Price Indexes (MXP)** containing data on changes in the prices of nonmilitary goods and services traded between the

U.S. and the rest of the world. Prices are available for nearly all merchandise categories. Published series are based on the following classification structures: BEA End Use, NAICS and the Harmonized System. Selected categories of international services are available as well as monthly import indexes by locality of origin. The IPP sampling frame comes from the Consumption Entry Document for goods entering the US economy directly and the Customs 214 form for goods entering Foreign Trade Zones. These are mandatory import export forms that all businesses are required to fill out. For exports to Canada, we use the import data collected by the Canadians.

### **B.5.2. Coverage**

Product groups covered account for nearly 100 percent of U.S. commodity imports and exports, by value. Selected special miscellaneous exports such as works of art are excluded. Service industries covered are air freight and air passenger fares. Imports to the U.S. are also classified by locality of origin.

### **B.5.3. Uses for Import/Export Indexes**

The primary uses of import/export price indexes are to deflate trade statistics. For merchandise trade, the End Use classification system is the structure used by the department of Commerce in the construction of foreign trade sector of the National Income and Products Accounts. For trade in international services, Balance of Payments indexes are used for deflating National Accounts data. In addition to this use the IPP also 1) measures import and export price trends for detailed and aggregate product groups. 2) Analyze price elasticities which reflect changes in the volume of trade in response to changes in prices and income. And to 3) analyze terms of trade and exchange rates.

### **B.5.4. Disclosure Risk**

The most significant disclosure risk for the International Price Program is information on companies participating in their survey. Especially if any of the data provide to BLS under confidentiality agreements were disclosed. As with other pricing programs, this data is considered but not limited to company names, price data or price trends from a particular company.

The IPP surveys are voluntary. Any disclosure of information about a particular company would potentially jeopardize their participation in the survey.

### **B.5.5 Data Suppression vs. Added Noise**

The IPP realizes that disclosure risk would be ameliorated if the data released were less precise but they take great care to produce the highest level of accuracy possible and would not like to release data that was intentionally muddy or contained noise.

One thing that is different with the IPP when compared to publication practices of other prices programs is they review their data and revise when necessary. Within three months, analysts can review a continuous stream of data and make changes to previously published indexes. This is in effort to increase the quality and thus utility of their product.

There are very few suppressed cells in the MXP. Their suppression rate due to confidentiality issues is estimated to be less than 1% of published cells. IPP do not produce special tabulations under any circumstances. Therefore confidentiality of these special tables is not an issue.