Data Discovery and re-use (citations)

Data citation refers to the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to other scholarly resources. They are references that can be included at the study level which point to published works that have used the data from a particular study such as a journal article, working paper, or news article. A citation gives credit to the data source and distributor and identifies data sources for validation. They are also a good way of showing the funders of surveys that the data are being used for policy and research purposes. Citations support researchers to manage and share data and enabling data citation and linking data with publications increase visibility and accessibility of data and the research itself.

Bibliographic references are important when you are using the data or ideas of others in your written work: references credit your sources and permit your readers to find those sources. Citing statistics and data has been a neglected grey area in academic publishing and often citation styles preferred by scientific journals largely ignore datasets and tables.

One of the features of the Pacific Data Hub – Microdata Library is a bibliography of publications that have cited the use of a dataset listed in the catalog. Selecting a publication from the list will show which study dataset was used and provide a link to the study in the catalog. This helps to improve data discovery and re-use. Data citation is an important practice in publishing of research. As data is shared with more frequency, data citation provides numerous advantages including reproducibility through direct reference to the data used in a research study; providing credit to data producers and authors, and ability for researchers to track the use of their datasets in other studies.

One key benefit from making microdata more accessible and usable is increasing the quantity and diversity of research. Experience shows that, when quality, well-documented data are made easily accessible, more analytical output is generated. It is important to document these uses of microdata. To promote access to microdata, the PDH Microdata Library citation catalog develops and maintains a database of citations and quotes from surveys listed in our central survey catalog. Citations include articles or chapters from academic journals, books, reports, working papers, dissertations, etc. To build this database of citations, SPC relies primarily on Google Scholar (other tools and databases may be used in the future).

Why should I Cite Data?

The need for a standard citation mechanism to give credit to data collectors and to locate and examine the data used in investigations is a long-standing requirement of data management.

Proper citation ensures that research data can be:

- Discovered (help find the data)
- Re-used
- Replicated for verification of results
- Credited for recognition
- Leave a trail - tracked to measure usage and impact, measure how data are used.
- Give credit to those who created the data
- Identify research already undertaken in a region, country, or on a particular topic and reduce the risk of duplicating research work

A citation gives benefit to the original data collectors:
They get credit for their hard work. Collecting and compiling data takes a lot of effort and money. They can demonstrate how their work is impacting their field (and others) to supervisors and donor funders, which may lead to more funding.

A citation gives benefit to the research community:
Data citation make life easier. They act like a map, helping you discover research data available for reuse. It's easier for publications to be found and linked to the dataset. Readers can see exactly which data you used for example a citation tells the reader: who collected the data and where the data came from.

**How data citation works**
The data citation is included in the standard references list of an article, and treated on equal footing with article citations. That also means readers will enjoy the same benefits as for article citations, including clickable links to the referenced material and the ability to quickly jump to the point in the article where this work was first cited.

**Data Citation Principles**
For data to be discovered and acknowledged it must be widely accessible and cited in a consistent and clear manner in literature. The following Data Citation Principles from Research Data Alliance\(^1\) reflect the various efforts to achieve this:

---
\(^1\) [https://rd-alliance.org/groups/data-citation-wg.html](https://rd-alliance.org/groups/data-citation-wg.html)
1. **Importance**: Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

2. **Credit and Attribution**: Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

3. **Evidence**: In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

4. **Unique Identification**: A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

5. **Access**: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6. **Persistence**: Unique identifiers, and metadata describing the data, and its disposition, should persist – even beyond the lifespan of the data they describe.

7. **Specificity and Verifiability**: Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time-slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

8. **Interoperability and flexibility**: Data citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data citation practices across communities.

**Data citation styles and formats and DOIs**

A Digital Object Identifier or DOI are a unique, persistent identifier or handle used to identify objects uniquely that can be used to track data citation metrics and to link related outputs such as journal articles, research data and software. DOIs are not essential but are considered best practice for data citation.

- For the dataset ID, using a Digital Object Identifier (DOI) is recommended.

The citations should provide information on the source and version of the dataset. The "citation requirement" is one of the elements of the Data Documentation Initiative metadata standard. The citation should include, at a minimum, information on the primary investigator, title and abbreviation of the dataset, reference year(s), version number, and dataset ID number. For data obtained on-line, it can also include the URL of the website from which the data were downloaded and the date of the download.

Data should be cited and referenced in publications and other media using a bibliographic reference similar to that used for journal articles. Individual researchers or data set owners may define data citation requirements for their data and research; however, the expectation of the Pacific Data Hub is that certain specific elements are included. These elements are:

- **Author/Principal Investigator/Data Creator**
  Who is the creator of the data set? This can be an individual, a group of individuals, or an organisation.

- **Date of Publication – year of release, for a completed dataset**
  What year was the data set published or disseminated? When was the data set posted online?

- **Title of Data Source – formal title of the dataset including version number, if applicable**
  What name is the data set called, or what is the name of the study?

- **Version/Edition Number**
  Version of the dataset used in the study

- **Format of the Data – physical format of the data**
What type of file is the data set? Or what software format is it stored in.

- **Locator or Identifier** – includes Digital Object Identifiers (DOI),
  What web address is the data set available at? Is there a persistent identifier available? If a DOI or other persistent identifier is associated with the data set it should be used in place of the URL.

- **Editor or Contributor** – reference to a person who compiled data
  Is there a person or team responsible for compiling or editing the data set?

- **Publication Place** – city and country of the distributor of the data
  Physical location associated with producing and/or distributing the data set.

- **Data within a Larger Work** – refers to the use of data in a compilation or a data supplement (such as published in a peer-reviewed paper)

These are the minimum elements required for dataset identification and retrieval. Fewer or additional elements may be requested by author guidelines or style manuals. Be sure to include as many elements as needed to precisely identify the dataset you have used.

---

**Steps involved in Data Discovery and citation on the Pacific Data Hub**

To provide data producers with proper credit for their input in data analysis outputs, and ensure the necessary transparency required to guarantee the "replicability" of analytical work, it is essential that data used by researchers be properly cited. This also allows data repositories like the Pacific Data Hub to maintain catalogs of citations (such as the one maintained by SPC), which are useful to measure and assess the uses of existing data, which in turn can be used by data producers to justify their data collection investments.

If you are a researcher analysing statistical data you will need to cite primary data sources in your published papers, just as you would cite other sources. The term "dataset" therefore in this case refers to the raw microdata files used for research, as well as the documents that give provenance and usage information about the data.

The process of citing a dataset from the *Pacific Data Hub* consists of the following steps:

1. Cite data in your references, for example reference the data used in your data tables. Identify which one(s) of the survey datasets is (are) quoted in the paper. Cite the exact version of the data used in your research, to support data discovery.
2. Verify that the citation is not a “false positive”, i.e. that it indeed quotes one or several of the survey datasets.

3. Find out if any other dataset is mentioned in the document

4. Identify the type of paper (journal, working paper, e.t.c)

5. Identify if there is a URL that exists. If the URL of cited web site in bibliography is no longer online then also check whether the content has a Digital Object Identifier DOI (see step 9)

6. Check that the citation is not already entered in the citation catalog

7. Add the citation to the catalog by entering in the bibtex

8. When all citations related to keywords you searched are entered, activate the “Email alert”

9. Include a unique identifier in your citation, such as Direct Object Identifiers (DOIs). These will enable the data to be accessed even if URLs change and thus provide a permanent link to the data.

**How to use Google Scholar to manage Citations**

Automatically creating a bibliography with Google Scholar

**Google Scholar** can facilitate citing a reference on-the-fly. Below each citation after running a search, there are a few links offering actions like exporting, saving, and more. If you don’t see a link called *Cite*, then click on the *More* option. This will display the *Cite* action where Google Scholar generates the reference in a few of the major citation styles.

Firstly, set up your Google Scholar account by signing into Gmail account
Open Google Scholar: click on the settings (on top left) = settings should shift to show links to bibTex (must be ticked)

Next step search on Google Scholar for a title or study of interest. Once the articles have been added, click on “Next step,” and find your newly created Google Scholar account with all of your publications’ reference data.

Final step download BibTex file In order to export your Google Scholar references into the Pacific Data Hub-Microdata Library Citations catalog, you will need to download them as a BibTex file.

Select all references by clicking in the box to the right of “Title.” Then click “Export” and select the BibTex file type.
Tip: Check Google Scholar regularly to find new references and import them into your Data Catalog.

Do the search with the key words, then make sure that:
- The correct word is searched (not another homonym)
- Then import into BibTex
- Then import to the “Citations” into the Pacific Data Hub – Microdata Library (select which study it will be attached to)
- Then add information
- Create “Alert query” for each research in Google Scholar so that further citations can be collated

Instructions for the maintenance of the PDH citations database (including bibliography of data-related citations)

Open the “Citations” catalog at https://microdata.pacificdata.org/index.php/citations

Login as administrator

Click on “Site administration” to open the Admin interface

Click on the Citations tab
The list of citations already entered will be displayed.

In another window, open Google Scholar

In Google scholar, search for the keywords related to the survey(s) for which you are looking for citations.
The process will then consist of the following steps, which are described below:

1. Verify that the citation is not a “false positive”, i.e. that it indeed quotes one or several of the dataset(s).
2. Identify which one(s) of the datasets is (are) quoted in the paper.
3. Find out if any other dataset is mentioned in the document.
4. Identify the type of paper (journal, working paper, etc).
5. Check that the citation is not already entered in the citation catalog.

If relevant, add the citation to the catalog.

When all citations related to keywords you searched are entered, activate the “Email alert” in Google Scholar.

1. Verify that the citation is not a “false positive”, i.e. that it indeed quotes one or several of the Fiji Population Census.

This can often be done simply by reading the short text below the title.

In some cases, this will require opening the PDF document (when available). The search tool in Acrobat will allow you to quickly find the survey dataset name (or whatever the survey acronym is for the specific country dataset you are searching).

2. Identify which one(s) of the datasets is (are) quoted in the paper, and
3. Find out if any other dataset is mentioned in the document.

The short text below the survey title will give you some information, but not necessarily all. Whenever available, open the PDF document and search the PDF file to find out if more than one dataset is quoted.

4. Identify the type of paper (journal, working paper, etc).
5. Check that the citation is not already entered in the citation catalog.

To do that, use either the author’s name, or any “rare” keyword from the title and search the PDH citations catalog. For example, if you want to find out if the following citation is already in the catalog:
Search for “Binge” or for “eating” in the PDH citations catalog. You will see that the citation was already added.

In that case, verify that the survey you are documenting is listed. Remember that one citation might have quoted surveys from more than one country. It is possible that someone else (or you) have already entered this citation when documenting dataset and find it again when looking for dataset. If at the time you documented the dataset you did not notice that the document also quotes the study, this citation will not be attached to the survey. In such case, simply edit the list of “Related studies” to add the relevant dataset reference.

**Edit Citation**

Select citation type
- Journal Article

Article title
- Binge eating and binge eating disorder in a small-scale, indigenous society: The view from Fiji

Journal
- International Journal of Eating Disorders

Author(s)

<table>
<thead>
<tr>
<th>First name</th>
<th>Last name</th>
<th>Middle initial</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Becker</td>
<td>E</td>
</tr>
<tr>
<td>R</td>
<td>Burwell</td>
<td>A</td>
</tr>
<tr>
<td>K</td>
<td>Navara</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Gilman</td>
<td>E</td>
</tr>
</tbody>
</table>

Click here to add more...

Volume  | Issue | Publication (Day/Month/Year) |
--------|-------|-----------------------------|
34      | 4     | -                           |

Page from | Page to | ISSN/ISBN/Other number |
---------|---------|------------------------|
423      | 451     |                        |

URL
6. **Add the citation to the catalog**

If the citation found in Google Scholar is not yet in the Pacific Data Hub and is not a “false positive”, add it to the catalog.

New citations can be “Added” (created from scratch) or “Imported” by copy/pasting BibTex information. In both cases, some visual checks and manual edits will be needed.

![Add New Citation](image)

Select the type (mandatory)

![Add New Citation Form](image)

Enter information on author(s); at least one author must be entered.

![Author(s) Form](image)

URL is only for links to the document (typically in PDF) ; SPC does not create links to sites which only provide the bibliographic citation and an option to purchase the PDF.
Abstract is optional, but we want to include it whenever possible. When copy/pasted from a PDF document, make sure to edit it to ensure that the flow of the text is continuous:

instead of:

Abstract

Objective: Although the cross-cultural prevalence of anorexia and bulimia nervosa has been investigated in multiple studies, published studies to date have explored BDD in small-scale, indigenous, or developing societies. The current study investigated the prevalence and correlates of binge eating in a community sample of Fijian women living in rural Fiji. Methods: Fifty ethnic Fijian women completed a self-report measure developed for this study on dieting and attitudes toward body shape and change, a Malaise-distress questionnaire on body image, and the Questionnaire on Eating and Weight Patterns-Revised (QBE-R). Their height and weight were also measured. Patterns of dieting, high body mass index (BMI), and attitudes toward eating and body image were compared between women with and without a history of binge eating. Results: Ten percent of respondents reported at least weekly episodes of binge eating during the past 6 months and 4% endorsed symptoms consistent with BDD. Binge eating in this sample was associated significantly with a BMI value above 30, a history of dieting, and a high concern with body shape.

we want:

Abstract

Objective: Although the cross-cultural prevalence of anorexia and bulimia nervosa has been investigated in multiple studies, little is known about the prevalence and correlates of binge eating and binge eating disorder (BDD) cross-culturally. In published studies to date have explored BDD in small-scale, indigenous, or developing societies. The current study investigated the prevalence and correlates of binge eating in a community sample of Fijian women living in rural Fiji. Methods: Fifty ethnic Fijian women completed a self-report measure developed for this study on dieting and attitudes toward body shape and change, a Malaise-distress questionnaire on body image, and the Questionnaire on Eating and Weight Patterns-Revised (QBE-R). Their height and weight were also measured. Patterns of dieting, high body mass index (BMI), and attitudes toward eating and body image were compared between women with and without a history of binge eating. Results: Ten percent of respondents reported at least weekly episodes of binge eating during the past 6 months and 4% endorsed symptoms consistent with BDD. Binge eating in this sample was associated significantly with a BMI value above 30, a history of dieting, and a high concern with body shape.

Click on related studies —— attach studies

Select in the list all studies quoted in the citation. Every citation should have at least one survey listed here.

Citations can also be entered by importing (some of) the metadata from BibTeX. Click on “Import citations”.

In Google Scholar, click on “Import into BibTeX.”
Binge eating and binge eating disorder in a small-scale, indigenous society: The view from Fiji

AC Becker, RA Burwell, K Navara... - ... Journal of Eating ..., 2003 - Wiley Online Library

Objective Although the cross-cultural prevalence of anorexia and bulimia nervosa has been investigated in multiple studies, little is known about the prevalence and correlates of binge eating and binge eating disorder (BED) cross-culturally. No published studies to date have ...

In the Pacific Data Hub – Microdata Library Citation manager, click on “Import citations” and paste the text you copied in the clipboard.