Challenges when collecting, classifying and analyzing photos of the respondents' books at home

WEB DATA OPP workshop / March 18-19

Patricia A. Iglesias | Research and Expertise Centre for Survey Methodology, Pompeu Fabra University

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 849165; PI: Melanie Revilla), the Agencia Nacional de Investigación y Desarrollo under the "Becas Chile" Doctoral Fellowship programme (grant agreement No. 72220301; Beneficiary: Patricia A. Iglesias), and GESIS – Leibniz-Institut für Sozialwissenschaften.

Our study

- Study developed within the **WEB DATA OPP** project.

- Collaboration with **Dr. Clemens Lechner** (GESIS – Leibniz-Institut für Sozialwissenschaften), **Dr. Birgit Heppt** (Humboldt-Universitat zu Berlin), and **Dr. Anna Volodina** (Institute for Educational Quality Improvement at the Humboldt-Universitat zu Berlin).

- Focus on the **books at home** → survey protocol by Iglesias *et* al. (2023) available at Open Research Europe (still under review).

- Two main objectives:
  - **Methodological**: Compare how different methods to collect information about the books at home perform in terms of participation and quality.
  - **Substantive**: Assess the relation between the number of books at home and the academic performance of children in primary school.

Why books at home?

- The **number of books at home** is a variable broadly present in social science surveys. However, it presents some limitations:
  - It is very likely that respondents **do not know the exact number of books** at their household → at best, such number can be estimated.

Why books at home?

- The **number of books at home** is a variable broadly present in social science surveys. However, it presents some limitations:
  - It is very likely that respondents **do not know the exact number of books** at their household → at best, such number can be estimated.
  - Social desirability bias might play a role in **over reporting the estimated number of books**.

Why books at home?

- The **number of books at home** is a variable broadly present in social science surveys. However, it presents some limitations:
  - It is very likely that respondents **do not know the exact number of books** at their household → at best, such number can be estimated.
  - Social desirability bias might play a role in **over reporting the estimated number of books**.
  - The response categories for this question are usually composed of broad/different intervals.

Response category
"10 books or less"
"11 to 25 books"
"26 to 100 books"
"101 to 200 books"
"201 to 500 books"
"More than 500 books"

(Sieben & Lechner, 2019)

Número de libros en el hogar
0-10 libros
11-25 libros
26-100 libros
101-200 libros
Más de 200 libros

(Gil Flores, 2011)

0-10 book(s)
11-25 books
26-100 books
101-200 books
201-500 books
More than 500 books

(Güre et al., 2023)

None
Less than 10
10 to 20
21 to 30
More than 30 books

(McNally et al., 2023)

Why books at home?

- The **number of books at home** is a variable broadly present in social science surveys. However, it presents some limitations:
  - It is very likely that respondents **do not know the exact number of books** at their household → at best, such number can be estimated.
  - Social desirability bias might play a role in **over reporting the estimated number of books**.
  - The response categories for this question are usually composed of broad/different intervals.
  - As it is a variable used as proxy of cultural and socioeconomic capital, **the number of books alone might not be sufficient** → type of books (from the title), which would be too burdensome for respondents.

- These limitations could be overcome by photos of the books at home, **potentially increasing the quality of the data and providing new insights.**

Methods

- Online survey that **could only be answered from a smartphone or tablet**.
- Collected through the Netquest online panel in Spain, in June 2023.
- Target population: parents of children in first, third, or fifth year of primary school.
- Quotas for gender, age, and level of education.
- 1,202 respondents completed the survey.
- We did not ask about e-Books.

Methods

Participants were asked a) to answer questions about the books they have at home, and/or b) to send photos of such books.

Methods

Participants were asked a) to answer questions about the books they have at home, and/or b) to send photos of such books.

Conventional questions:

- **the number of books**: four open-ended questions about 1) the total number of books at home, and the number of books 2) for toddlers and children who do not know how to read, 3) for literate children and teenagers, and 4) aimed at a general audience.
- **language**: three open-ended questions asking for the percentage of books 1) in Spanish, 2) in one of the three co-official languages in Spain (Catalan, Galician, and Euskera), and 3) in other languages.
- **storage**: four radio-button questions asking whether books are stored 1) in shelves, 2) inside closets or drawers, 3) in center, coffee, or night tables or over a desk, and 4) in other places.

Methods

Participants were asked a) to answer questions about the books they have at home, and/or b) to send photos of such books.

Conventional questions:

- **the number of books**: four open-ended questions about 1) the total number of books at home, and the number of books 2) for toddlers and children who do not know how to read, 3) for literate children and teenagers, and 4) aimed at a general audience.
- **language**: three open-ended questions asking for the percentage of books 1) in Spanish, 2) in one of the three co-official languages in Spain (Catalan, Galician, and Euskera), and 3) in other languages.
- **storage**: four radio-button questions asking whether books are stored 1) in shelves, 2) inside closets or drawers, 3) in center, coffee, or night tables or over a desk, and 4) in other places.

Methods

Participants were asked a) to answer questions about the books they have at home, and/or b) to send photos of such books.

Conventional questions:

- **the number of books**: four open-ended questions about 1) the total number of books at home, and the number of books 2) for toddlers and children who do not know how to read, 3) for literate children and teenagers, and 4) aimed at a general audience.
- **language**: three open-ended questions asking for the percentage of books 1) in Spanish, 2) in one of the three co-official languages in Spain (Catalan, Galician, and Euskera), and 3) in other languages.
- **storage**: four radio-button questions asking whether books are stored 1) in shelves, 2) inside closets or drawers, 3) in center, coffee, or night tables or over a desk, and 4) in other places.

Methods

Participants were asked a) to answer questions about the books they have at home, and/or b) to send photos of such books.

Conventional questions:

- **the number of books**: four open-ended questions about 1) the total number of books at home, and the number of books 2) for toddlers and children who do not know how to read, 3) for literate children and teenagers, and 4) aimed at a general audience.
- **language**: three open-ended questions asking for the percentage of books 1) in Spanish, 2) in one of the three co-official languages in Spain (Catalan, Galician, and Euskera), and 3) in other languages.
- **storage**: four radio-button questions asking whether books are stored 1) in shelves, 2) inside closets or drawers, 3) in center, coffee, or night tables or over a desk, and 4) in other places.

Images-based question:

Sending photos of all the books in the household.

The information of interest could be classified from the photos.

Methods

The conventional questions were collected through two methods:

Text

Asked for the number, language, and storage of books by using conventional questions.

TextPlus

Similar to **Text**, but with a visual example for the number-of-books questions.

Methods

Message for group TextPlus: *To help you estimate the total number of books that you have in your main residence, please, look at the examples below: you can see that a 74 centimeters long shelf can contain from around 30 to almost 80 books, depending on the thickness of the books.*

Methods

Group **Choice** *(n=301)*	Text or Images		
Group **Text-TextPlus** *(n=301)*	Text	→	TextPlus
Group **TextPlus-Images** *(n=300)*	TextPlus	→	Images
Group **Images-Text** *(n=300)*	Images	→	Text

Methods

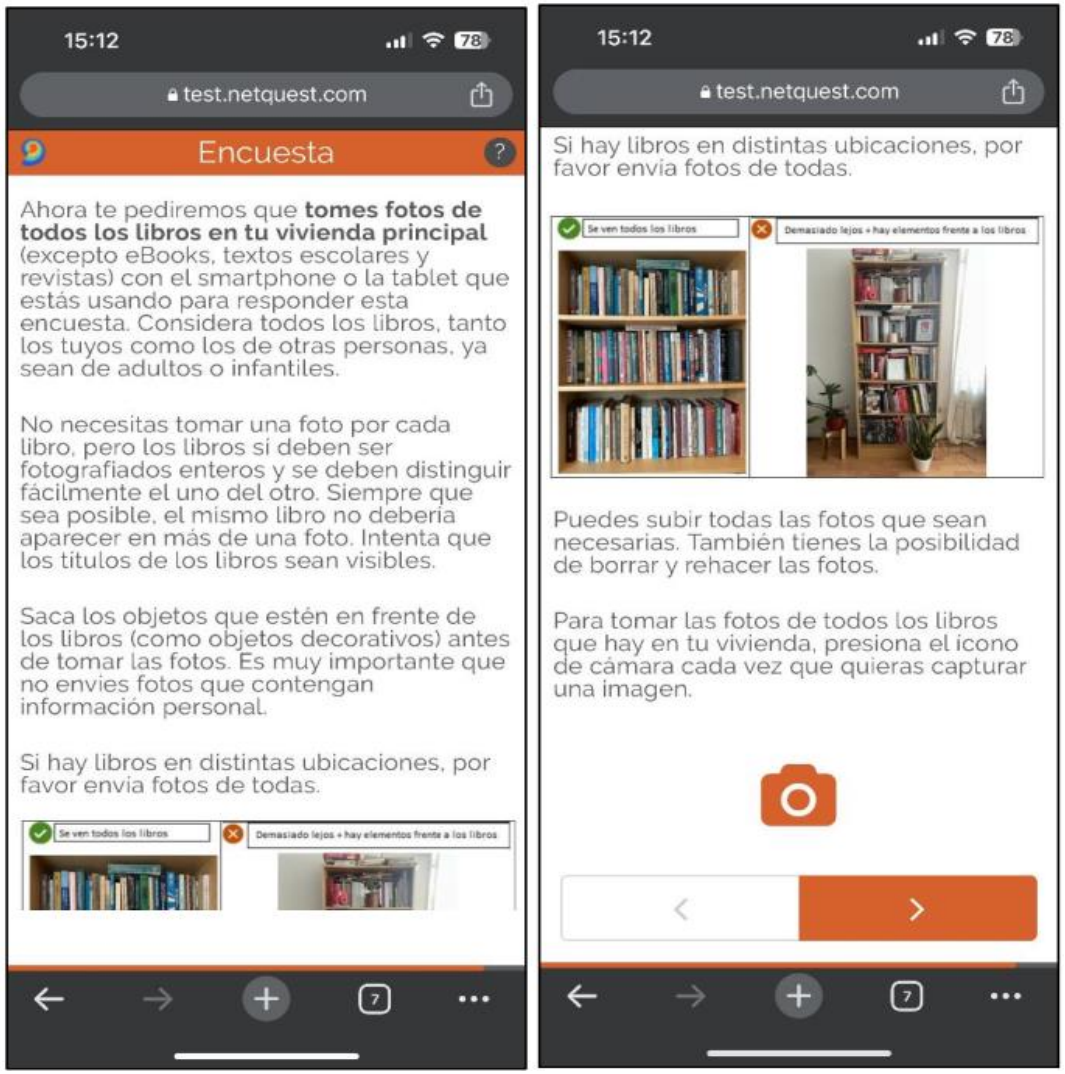
- Tool for collection:

WebdataVisual (Revilla *et al.*, 2022)

- The tool works within the survey → respondents do not need to leave the survey to do the photos.

- Respondents could send as many photos as they wanted.

- They could preview, delete, and re-capture the photos.



Methods

- For the classification of the photos:
  - Human classification with 2 classifiers sharing the work → overlap of 100 images (out of 723).
  - Classifiers trained by the research team.
  - Detailed guidelines were developed.
- Algorithms for classification were not available at the moment… but could be in the future.

Main challenges for the images-based question: collection

- Participants **skipping** the images-based question or **breaking-off** the survey when seeing it.
  - 66% out of 703 participants asked for photos did not send a photo or left the survey.

Main challenges for the images-based question: collection

- Participants **skipping** the images-based question or **breaking-off** the survey when seeing it.
  - 66% out of 703 participants asked for photos did not send a photo or left the survey.

179 respondents (**26%**) did not share photos due to privacy concerns…

Main challenges for the images-based question: collection

- Participants **skipping** the images-based question or **breaking-off** the survey when seeing it.
  – 66% out of 703 participants asked for photos did not send a photo or left the survey.

179 respondents (**26%**) did not share photos due to privacy concerns…

… 48 (7%) reported technical issues when uploading the photos…

Main challenges for the images-based question: collection

- Participants **skipping** the images-based question or **breaking-off** the survey when seeing it.
  - 66% out of 703 participants asked for photos did not send a photo or left the survey.

179 respondents (**26%**) did not share photos due to privacy concerns…

… 48 (7%) reported technical issues when uploading the photos…

… 44 (6%) stated their camera was not working.

Main challenges for the images-based question: collection

- Participants **skipping** the images-based question or **breaking-off** the survey when seeing it.
  - 66% out of 703 participants asked for photos did not send a photo or left the survey.

- Some participants were **not at home**, thus they were unable of capturing photos of the books at home.
  - 53 (8%) stated they were not at home… but they could be even more!

[IF ANSWERING FROM PC] PC_PLACE. Thank you for your interest in participating in this survey. **This survey should be answered:**

- **From your main residence (meaning the place where you have most of your belongings)**. If you are not there right now, please, come back once you are in your main residence.

- From a **smartphone or tablet.** Please, access to the survey again from one of these devices.

Main challenges for the images-based question: collection

- Participants **skipping** the images-based question or **breaking-off** the survey when seeing it.
  - 66% out of 703 participants asked for photos did not send a photo or left the survey.

- Some participants were **not at home**, thus they were unable of capturing photos of the books at home.
  - 53 (8%) stated they were not at home… but they could be even more!

- At least 6 respondents wanted to send photos but moved forward in the survey and **could not return to the page with the camera feature** → not a specific survey question, but stated in the final open-narrative question. So, they could also be more!

Main challenges for the images-based question: collection

- **Only 4%** of participants in group Choice preferred to send photos → less than expected.

Main challenges for the images-based question: collection

- **Only 4%** of participants in group Choice preferred to send photos → less than expected.

- Of the 723 photos sent, 8% included **personal information** → revision by the fieldwork company and the ethics advisor.

Main challenges for the images-based question: classification

- 2% were of the photos **off-topic** or of **bad visual quality**.

Main challenges for the images-based question: classification

- 2% were of the photos **off-topic** or of **bad visual quality**.

- Difficulty to distinguish between the three books categories, particularly between **books for children who do not know how to read** and **books for literate children and teenagers.**

Main challenges for the images-based question: classification

- 2% were of the photos **off-topic** or of **bad visual quality**.

- Difficulty to distinguish between the three books categories, particularly between **books for children who do not know how to read** and **books for literate children and teenagers.**

- Further, some books **could not be categorized**.

Main challenges for the images-based question: classification

- It was difficult to tell **if certain items were books** (CDs, DVDs, notebooks, magazines and so on).

Main challenges for the images-based question: classification

- It was difficult to tell **if certain items were books** (CDs, DVDs, notebooks, magazines and so on).

- Special attention was aimed at identifying **overlap of books when respondents sent more than one photo.**

Main challenges for the images-based question: classification

- Example of overlap

Main challenges for the images-based question: classification

- It was difficult to tell **if certain items were books** (CDs, DVDs, notebooks, magazines and so on).

- Special attention was aimed at identifying **overlap of books when respondents sent more than one photo.**

- **Continuous update** of the classification guidelines → review of the photos.

Main challenges for the images-based question: analysis

- True value **unknown.**

Main challenges for the images-based question: analysis

- True value **unknown.**

- **Treatment of the dataset** → longer than expected:
  - **First, reviewing the dataset.**
  - Second, merging the classification datasets.
  - **Third, transforming/creating variables from the photo-level to the respondent-level.**
  - Fourth, merging the classification dataset at the respondent-level, to the main survey dataset.

Main challenges for the images-based question: analysis

- True value **unknown.**

- **Treatment of the dataset** → longer than expected:
  - **First, reviewing the dataset.**
  - Second, merging the classification datasets.
  - **Third, transforming/creating variables from the photo-level to the respondent-level.**
  - Fourth, merging the classification dataset at the respondent-level, to the main survey dataset.

- **Difficulties to assess data quality.**

Main challenges for the images-based question: analysis

- True value **unknown.**

- **Treatment of the dataset** → longer than expected:
  - **First, reviewing the dataset.**
  - Second, merging the classification datasets.
  - **Third, transforming/creating variables from the photo-level to the respondent-level.**
  - Fourth, merging the classification dataset at the respondent-level, to the main survey dataset.

- **Difficulties to assess data quality.**

- And possibly other issues → most analyses are yet to be conducted, both in the methodological and substantive spheres.

Conclusions

- Researchers should carefully consider whether photos **could provide better results than conventional survey questions**.

Conclusions

- Researchers should carefully consider whether photos **could provide better results than conventional survey questions**.

- The collection needs to be carefully planned to:
  - Avoid respondents **not being able to send photos** (e.g., not at home)
  - Avoid respondents **not being able to return to the camera feature**
  - Tackle the **privacy issues** mentioned by some respondents → how?

Conclusions

- Researchers should carefully consider whether photos **could provide better results than conventional survey questions**.

- The collection needs to be carefully planned to:
  - Avoid respondents **not being able to send photos** (e.g., not at home)
  - Avoid respondents **not being able to return to the camera feature**
  - Tackle the **privacy issues** mentioned by some respondents → how?

- As for the classification:
  - The items of interest should be **clearly defined**
  - Process is longer than what was initially expected, and more than one iteration is needed

Conclusions

- Researchers should carefully consider whether photos **could provide better results than conventional survey questions**.

- The collection needs to be carefully planned to:
  - Avoid respondents **not being able to send photos** (e.g., not at home)
  - Avoid respondents **not being able to return to the camera feature**
  - Tackle the **privacy issues** mentioned by some respondents → how?

- As for the classification:
  - The items of interest should be **clearly defined**
  - Process is longer than what was initially expected, and more than one iteration is needed

- Regarding the analyses:
  - Current main challenge**: assessing data quality**.

Thanks!

Questions?

Patricia A. Iglesias | Research and Expertise Centre for Survey Methodology, Pompeu Fabra University

patricia.iglesias@upf.edu

https://www.upf.edu/web/webdataopp

Universitat Pompeu Fabra Barcelona

RECSM Research and Expertise Centre for Survey Methodology

web data opp

References

- Gil Flores, J. (2011). Medición del nivel socioeconómico familiar en el alumnado de Educación Primaria. *Revista de Educación*, *362*. https://doi.org/10.4438/1988-592X-RE-2011-362-162

- Güre, Ö., Sevgin, H., & Kayri, M. (2023). Reviewing the Factors Affecting PISA Reading Skills by Using Random Forest and MARS Methods. *International Journal of Contemporary Educational Research*, *10*(1), 181–196. https://doi.org/10.33200/ijcer.1192590

- Iglesias, P.A., Revilla, M., Heppt, B., Volodina, A., & Lechner, C. (under review). Protocol for a web survey experiment studying the feasibility of asking respondents to capture and submit photos of the books they have at home and the resulting data quality [version 1; peer review: awaiting peer review]. *Open Res Europe* 2023, 3:202 (https://doi.org/10.12688/openreseurope.16507.1)

- McNally, S., Leech, K. A., Corriveau, K. H., & Daly, M. (2023). Indirect Effects of Early Shared Reading and Access to Books on Reading Vocabulary in Middle Childhood. *Scientific Studies of Reading*, 1–18. https://doi.org/10.1080/10888438.2023.2220846

- Revilla, M., Iglesias, P., Ochoa, C., & Antón, D. (2022). WebdataVisual: a tool to gather visual data within the frame ofweb surveys. OSF. http://doi.org/10.17605/OSF.IO/R7CAX

- Sieben, S., & Lechner, C. M. (2019). Measuring cultural capital through the number of books in the household. *Measurement Instruments for the Social Sciences*, *1*(1), 1–6. https://doi.org/10.1186/s42409-018-0006-0