Abstract

For large surveys, creating a comprehensive codebook can present many challenges. Without automation, this process can be labor intensive and error-prone; data in the codebook can be outdated or inconsistent with the actual data in the datasets. Another major challenge is that information/data for codebooks can come from multiple sources, such as questionnaire specifications, the questionnaire design system, and database or SAS datasets containing collected data. Our poster presents an integrated approach to codebook generation using modern tools and technologies, including SAS dictionary tables and SAS Integrated Object Model (IOM) for data management, HTML/CSS for codebook presentation, and the .NET framework for integrating the pieces together.

1. Introduction

Assumptions:

The SAS dataset from which the codebook will be created from has gone through all data processing required for the project and is ready for delivery. As such, it should contain formats and labels for all variables in the dataset. Additional survey metadata, such as question text, special notes, imputation formulation, and other information, may also be available for inclusion with the codebook.

Technologies:

The following technologies were used to leverage the strengths in each technology:

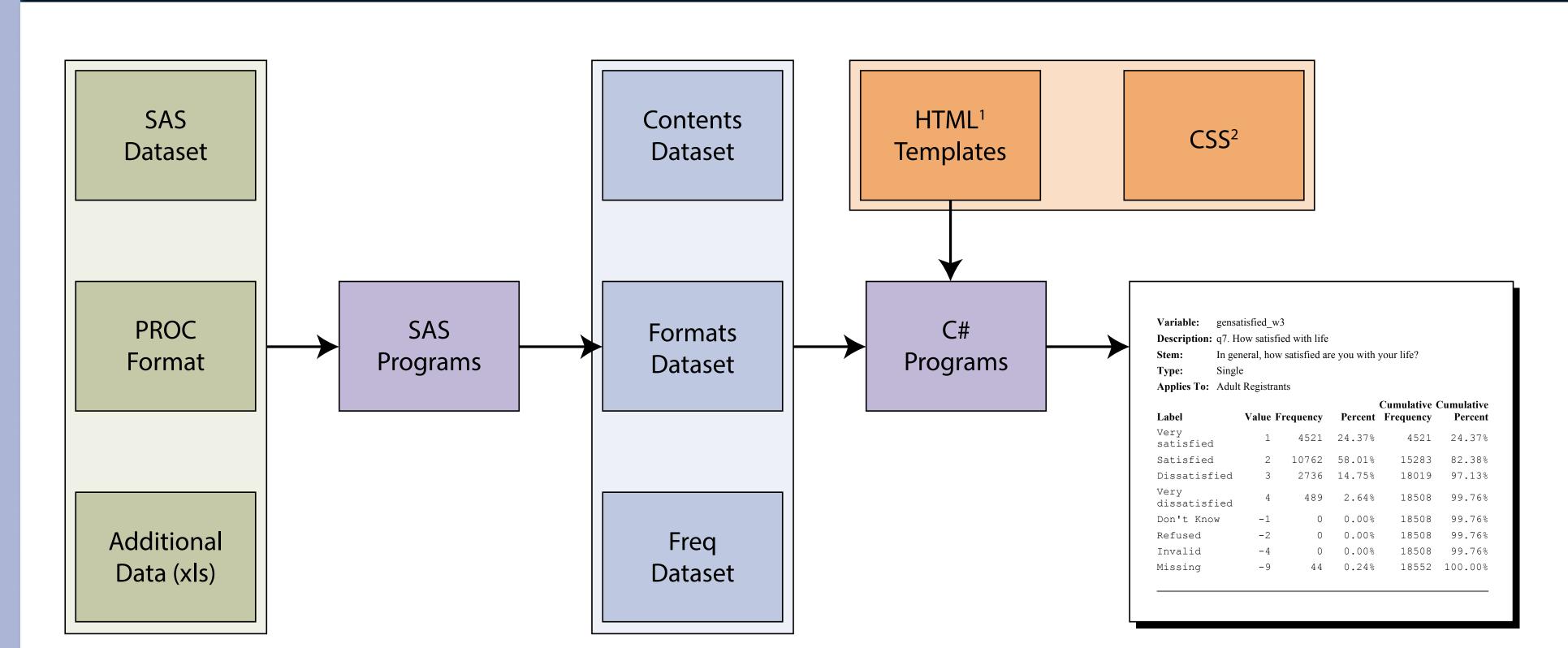
- SAS® Software Statistical analysis, handling large amounts of data or computation-intensive applications
- .NET Framework Microsoft's platform for building windows and web applications
- SAS Integrated Object Model (IOM) SAS IOM provides a bridge between SAS and .NET platforms
- HTML/CSS Standardized styling and structure of data

Implementation:

The system has 2 components:

- 1. SAS programs are responsible for reading the input SAS datasets and creating a set of SAS datasets:
 - Contents dataset Contains a list of variables in the input dataset and their attributes
 - Formats dataset Contains all format definitions that will be used in the codebook
 - Frequencies dataset Contains the frequency table for each variable
 - Survey Metadata dataset Contains additional metadata specific to a survey
- 2. .NET program utilizes data from codebook datasets and a set of predefined HTML templates to produce the codebook. The key features of this program are:
 - Reads SAS datasets using SAS IOM and .NET ADO.
 NET technologies. Data access is done in SQL.
 - Uses HTML to provide document structure.
 - Uses CSS to format and style codebook.
 - Changes the presentation of the codebook without having to change programming codes by using HTML template and CSS.

2. Data Flow Diagram



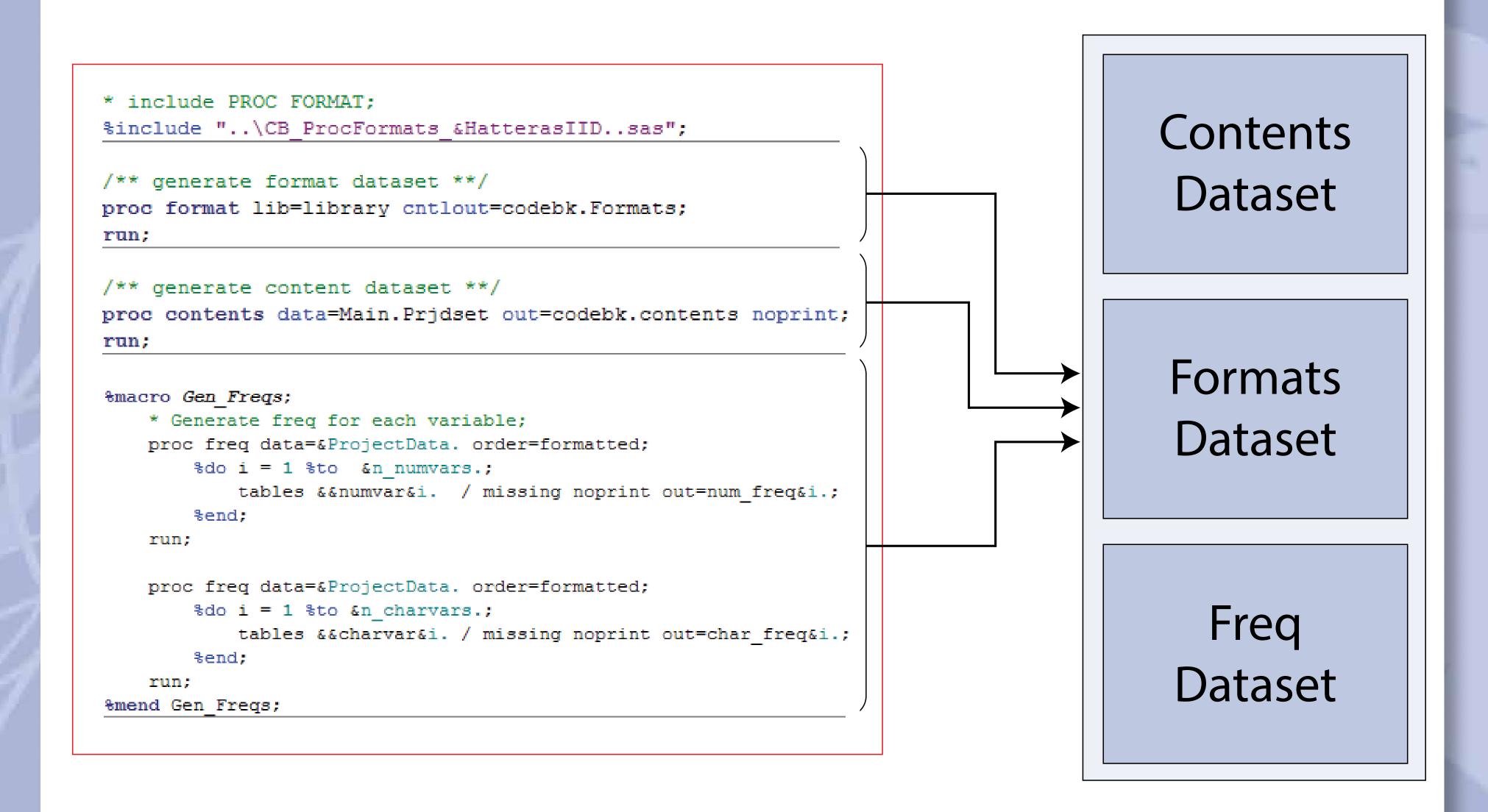
SAS® software launches the codebook process and then the resulting template is stylized and saved using C#

¹ HTML — Hypertext Markup Language, http://www.w3.org/community/webed/wiki/HTML

² CSS — Cascading Style Sheet, http://www.w3.org/Style/CSS/Overview.en.html

3. Codebook Data Gathering with SAS

Code snippets to generate codebook datasets from the SAS programs are shown below. The SAS programs are responsible for generating codebook data: labels, formats, frequencies, and other notes. These tasks were done using standard SAS procedures: PROC CONTENTS, PROC FORMAT, PROC FREQ, and PROC SQL. Codebook data are stored in SAS dataset formats, which are accessible to the .NET program via SAS IOM.



SAS code can be used to format data and to perform the statistical calculations such as PROC FREQ

4. Templates with HTML/CSS

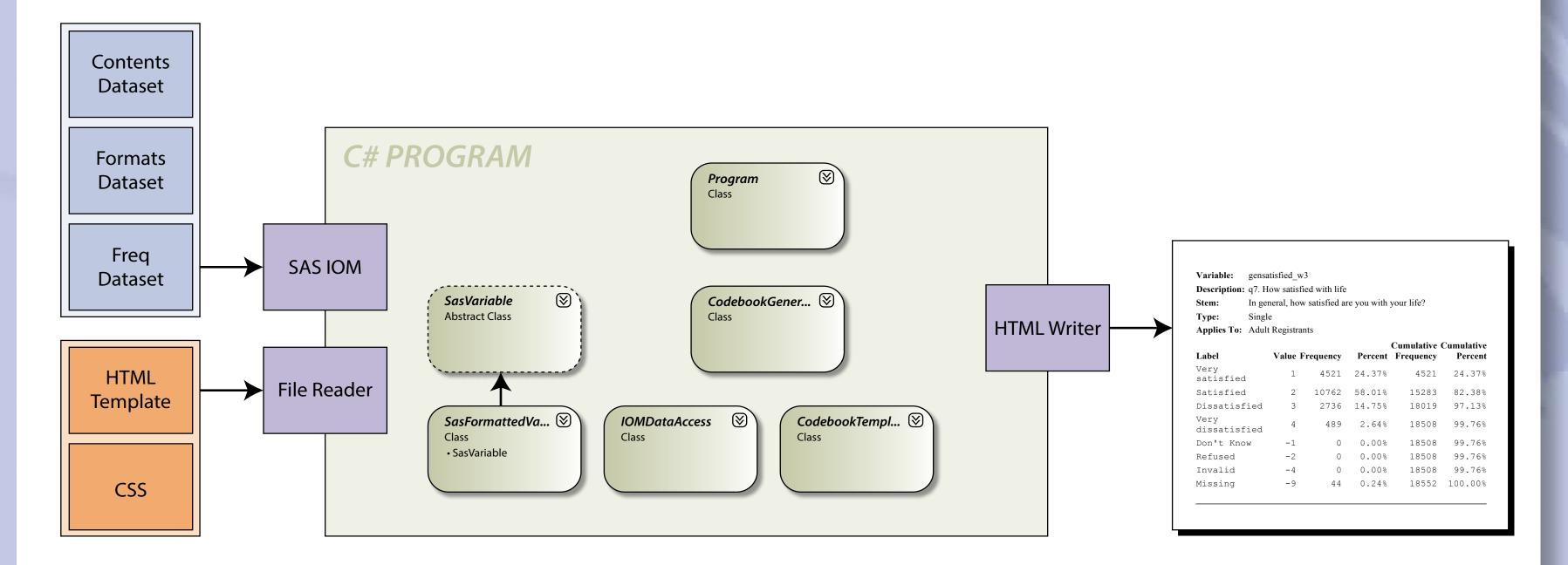
Several examples of the HTML templates and the CSS styles are shown below. These templates and CSS provide several benefits:

- Codebook structure can be easily manipulated without code changes
- Codebook layout, font selection, color, and styles can be easily controlled and adapted to the requirements
- CSS is easy to change and requires nothing more than a text editor like Notepad
- Provides separation of style from content
- Easy to apply 508-compliant styles
- Allows for lightweight content and presentation code

Templates are easily modified with a simple text editor and can be customized for content

5. Generating the Codebook

A .NET program, implemented in C#, is responsible for reading the codebook data, merging the data into a set of HTML templates, and producing the codebook file in HTML format. The program uses CSS styles to control the presentation of the codebook. The use of HTML templates and CSS stylesheets is an important design element of this program since they provide ability for codebook customization and many other benefits, as shown below. In addition, many modern programming constructs were used in the implementation, such as object-oriented programming, OLE DB, SQL, and SAS IOM.



SAS creates variable information, then C# applies the final style and template

6. Conclusion

A portion of a sample codebook is shown below. Some of the highlights are:

- Data Display
 - Double-column layout
 - Table layout for data
 - Heading labels in bold
 - Use of fixed-sized font to align frequency data
 - Horizontal break line between variables
- Usability
 - Easy to read
 - Conveys information clearly and distinctly
 - Reflects explicit research questions
 - Can be used as a summary to perform initial research exploration
 - Easily incorporated into word processing applications for reporting

Variable: Cont	trolNo					Variable: ma	ar_status_w	3				Variable: ge	nder_w3				
Description: Cont	trol number					Description: q4	. Current m	arital status				Description: q3	. Gender				
Stem: enro	llee ID					Stem: W	hat is your	current marit	tal status?			Stem: W	hat is your go	ender?			
Type: Sing	le					Type: Si	ngle					Type: Si	ngle				
Applies To: Adu	lt Registran	ts				Applies To: Ac	dult Registra	ants				Applies To: Ac	dult Registraı	nts			
			(Cumulative (Cumulative					Cumulative (Cumulative					Cumulative (Cumulativ
Label	Value F	requency		Frequency	Percent	Label	Value	Frequency		Frequency	Percent	Label	Value F	requency		Frequency	Perce
Data Present		3	0.02%	3	0.02%	Never marri	ed 1	2949	15.90%	2949	15.90%	Male	1	11305	60.94%	11305	60.94
Don't Know	-1	0	0.00%	3	0.02%	Married	2	12041	64.90%	14990	80.80%	Female	2	7237	39.01%	18542	99.95
Refused	-2	0	0.00%	3	0.02%	Not married	-					Don't Know	-1	0	0.00%	18542	99.95
Invalid	-4	0	0.00%	3	0.02%	living with partner	a 3	0	0.00%	14990	80.80%	Refused	-2	0	0.00%	18542	99.95
Missing	- 9	0	0.00%	3	0.02%	Widowed	4	362	1.95%	15352	82.75%	Invalid	-4	0	0.00%	18542	99.95
out of range		18549	99.98%	18552	100.00%	Divorced or						Missing	- 9	10	0.05%	18552	100.00
						separated	5	2293		17645	95.11%						
						Don't Know	-1	0	0.00%	17645	95.11%						
						Refused	-2	0	0.00%	17645	95.11%						
							_	•	0 000	17645	95.11%						
						Invalid	-4	0	0.00%	17645	90.116						
						Invalid Missing	-4 -9	28	0.00%	17673	95.26%						
	_	(mail, pho	ne, web)			Missing		28				Variable: ei	mploy_part_v	w3			
Description: Inter	_	(mail, pho	ne, web)			Missing	-9	28 w3	0.15%			Variable: en Description: q			part time		
Description: Inter Stem:	_	(mail, pho	ne, web)			Variable: en Description: q5	-9 nploy_full_v .1: Currentl	28 w3	0.15%			Description: q		y employed	•		
Variable: mod Description: Inter Stem: Type: Applies To: Adu	rview mode		ne, web)			Variable: en Description: q5 Stem: Er	-9 nploy_full_v .1: Currentl	w3 y employed	0.15%			Description: q. Stem: en	5.2: Currently	y employed	•		
Description: Inter Stem: Type: Applies To: Adu	rview mode	ts	(Cumulative (Variable: en Description: q5 Stem: Er	nploy_full_v .1: Currentl nployed for	w3 y employed full-time wa	0.15%	17673	95.26%	Description: q. Stem: en	5.2: Currently mployed for jungle	y employed part-time wa	ages		
Description: Inter Stem: Type: Applies To: Adu Label	rview mode It Registran Value F	ts	Percent	Frequency	Percent	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac	-9 .1: Currentl nployed for ngle dult Registra	w3 y employed full-time wa	0.15%	17673	95.26%	Description: q Stem: en Type: S Applies To: A	5.2: Currently mployed for plant ingle dult Registra	y employed part-time wa unts	ages	Cumulative Frequency	
Description: Intersteem: Type: Applies To: Adu Label Don't Know	tview mode It Registran Value France	ts requency	Percent	Frequency 0	Percent	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac Label	nploy_full_v .1: Currentl nployed for ngle dult Registra Value	w3 y employed full-time wa	full time ages	17673 Cumulative Control of the con	95.26% Cumulative Percent	Description: q Stem: en Type: S Applies To: A Label	5.2: Currently mployed for plantingle dult Registra Value	y employed part-time wa nnts Frequency	ages Percent	Frequency	Perc
Description: Interstem: Type: Applies To: Adu Label Don't Know Refused	treew mode It Registran Value France -1 -2	requency 0	Percent 0.00% 0.00%	Frequency 0 0	Percent 0.00% 0.00%	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac Label Unchecked	-9 .1: Currentl nployed for ngle dult Registra	w3 y employed full-time wa ants Frequency 7324	full time ages Percent 39.48%	Cumulative Crequency	Cumulative Percent 39.48%	Description: q Stem: en Type: S Applies To: A Label Unchecked	5.2: Currently mployed for plant ingle dult Registra	y employed part-time wa unts Frequency 17355	Percent 93.55%	Frequency 17355	Perc 93.5
Description: Interstem: Type: Applies To: Adu Label Don't Know Refused Invalid	Value Fi	requency 0 0	Percent 0.00% 0.00% 0.00%	Frequency 0 0 0	Percent 0.00% 0.00% 0.00%	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac Label Unchecked Checked	nploy_full_v .1: Currentl nployed for ngle dult Registra Value 0 1	w3 y employed full-time wa ants Frequency 7324 11217	full time ages Percent 39.48% 60.46%	Cumulative C Frequency 7324 18541	95.26% Cumulative Percent 39.48% 99.94%	Description: q Stem: en Type: S Applies To: A Label Unchecked Checked	5.2: Currently mployed for plantingle dult Registra Value 1 0 1	y employed part-time wa unts Frequency 17355 1186	Percent 93.55% 6.39%	Frequency 17355 18541	Perc 93.5
Description: Interstem: Stem: Type: Applies To: Adu Label Don't Know Refused Invalid Missing	Value Front -1 -2 -4 -9	ts requency 0 0 0	Percent 0.00% 0.00% 0.00% 0.00%	Frequency 0 0 0 0 0	Percent 0.00% 0.00% 0.00% 0.00%	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac Label Unchecked Checked Don't Know	nploy_full_v.1: Currentle nployed for ngle dult Registra Value 0 1 -1	w3 y employed full-time wa ants Frequency 7324 11217 0	0.15% full time ages Percent 39.48% 60.46% 0.00%	Cumulative (Frequency 7324 18541 18541	95.26% Cumulative Percent 39.48% 99.94% 99.94%	Description: q Stem: en Type: S Applies To: A Label Unchecked Checked Don't Know	5.2: Currently mployed for plant ingle dult Registra Value 1 0 1 -1	y employed part-time wa ints Frequency 17355 1186 0	Percent 93.55% 6.39% 0.00%	Frequency 17355 18541 18541	Perce 93.5 99.9
Description: Intersection: Stem: Type: Applies To: Adu Label Don't Know Refused Invalid Missing Mail	Value Front -2 -4 -9 M	requency 0 0 0 0	Percent 0.00% 0.00% 0.00% 0.00%	Frequency	Percent 0.00% 0.00% 0.00% 0.00% 0.00%	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac Label Unchecked Checked Don't Know Refused	-9 nploy_full1: Currentl nployed for ngle dult Registra Value 0 1 -1 -2	w3 y employed full-time wa ants Frequency 7324 11217 0 0	0.15% full time ages Percent 39.48% 60.46% 0.00% 0.00%	Cumulative (Frequency 7324 18541 18541 18541	95.26% Cumulative Percent 39.48% 99.94% 99.94%	Description: q Stem: en Type: S Applies To: A Label Unchecked Checked Don't Know Refused	5.2: Currently mployed for plant ingle dult Registra Value 1 0 1 -1 -2	y employed part-time wa ints Frequency 17355 1186 0 0	Percent 93.55% 6.39% 0.00% 0.00%	Frequency 17355 18541 18541 18541	Perce 93.5 99.9 99.9
Description: Interstem: Type: Applies To: Adu Label Don't Know Refused Invalid Missing	Value Front -1 -2 -4 -9	requency 0 0 0 0 0	Percent 0.00% 0.00% 0.00% 0.00%	Frequency	Percent 0.00% 0.00% 0.00% 0.00%	Variable: en Description: q5 Stem: Er Type: Si Applies To: Ac Label Unchecked Checked Don't Know	nploy_full_v.1: Currentle nployed for ngle dult Registra Value 0 1 -1	w3 y employed full-time wa ants Frequency 7324 11217 0	0.15% full time ages Percent 39.48% 60.46% 0.00%	Cumulative (Frequency 7324 18541 18541	95.26% Cumulative Percent 39.48% 99.94% 99.94%	Description: q Stem: en Type: S Applies To: A Label Unchecked Checked Don't Know	5.2: Currently mployed for plant ingle dult Registra Value 1 0 1 -1	y employed part-time wa ints Frequency 17355 1186 0	Percent 93.55% 6.39% 0.00%	Frequency 17355 18541 18541	Perce 93.5 99.9 99.9 99.9

The final codebook result is easy to read and allows for multiple formatting options

Contact Information

*Presenting author: Helen Smith

For further information: Mai Nguyen Email: mnguyen@rti.org Phone: 919.541.8757 • Fax: 919.541.6178

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.

® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

