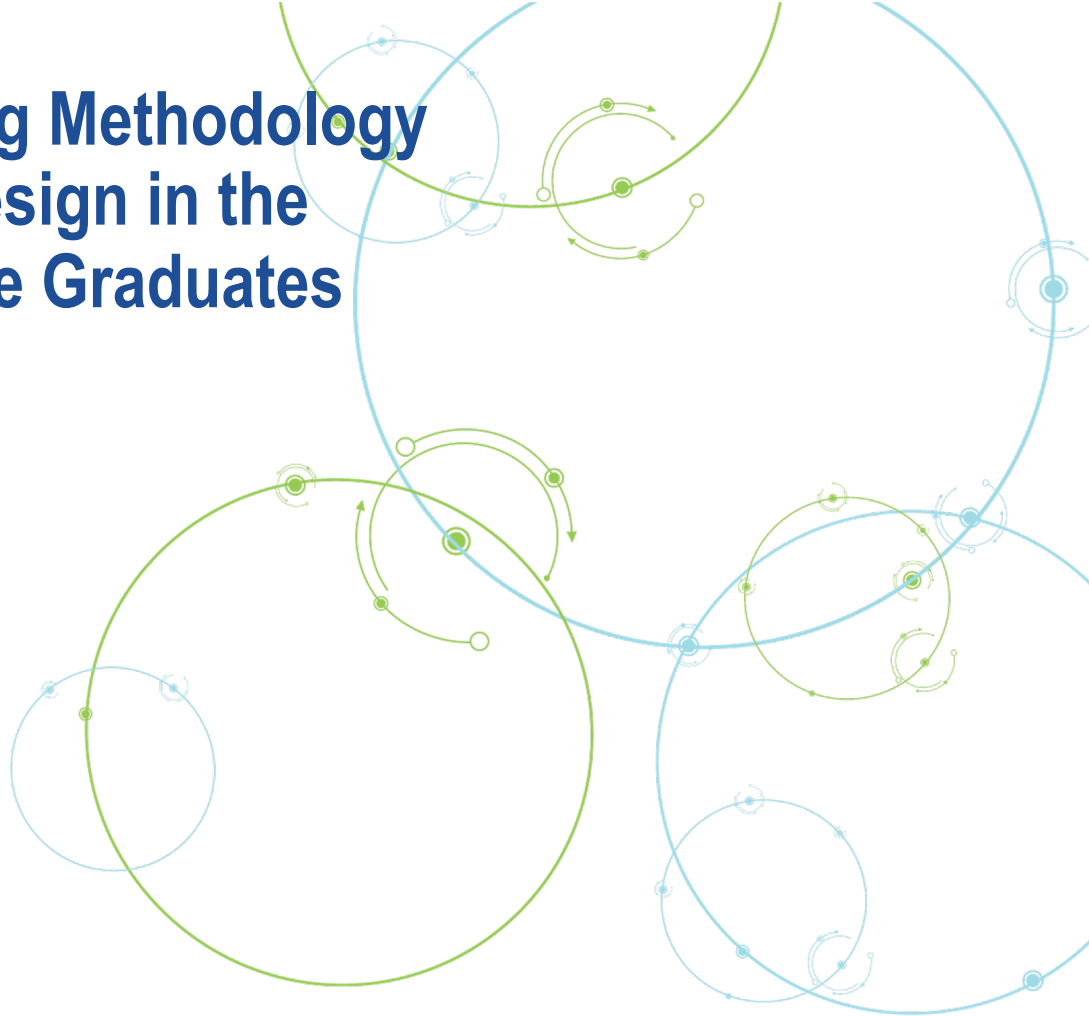
Developing and Evaluating Methodology for Split Questionnaire Design in the National Survey of College Graduates

Andy Peytchev

FCSM Conference
November, 2021



Acknowledgement

- Collaboration with Emilia Peytcheva, Darryl Cooney, Darryl Creel, Dave Wilson, and Jeremy Porter

- This research is supported by the National Center for Science and Engineering Statistics (NCSES) at NSF under a broad agency agreement
  - Jennifer Sinibaldi and Matthew Williams

- The views expressed in this document are those of the authors and do not necessarily reflect the views of the National Center for Science and Engineering Statistics within the National Science Foundation

Objective

- Evaluate SQD to reduce burden
  - How well is the data reproduced
  - What methods perform best, specifically for NCSES data

- National Survey of College Graduates
  - Fairly long (approximately half an hour to complete)
  - Data not in restricted access, allowing flexibility in statistical tools

Survey Length

- Reduce burden on each respondent
  - Sharp and Frankel (1983)

- Reduce nonresponse and potential nonresponse bias
  - Long questionnaires can have higher nonresponse rates - e.g., Heberlein and Baumgartner (1978); Adams and Darwin (1982); Dillman, Sinclair and Clark (1993)
    - Finding less consistent for interviewer-administered modes
    - Lack of evidence for nonresponse bias

- Reduce measurement error
  - Peytchev and Peytcheva (2017)

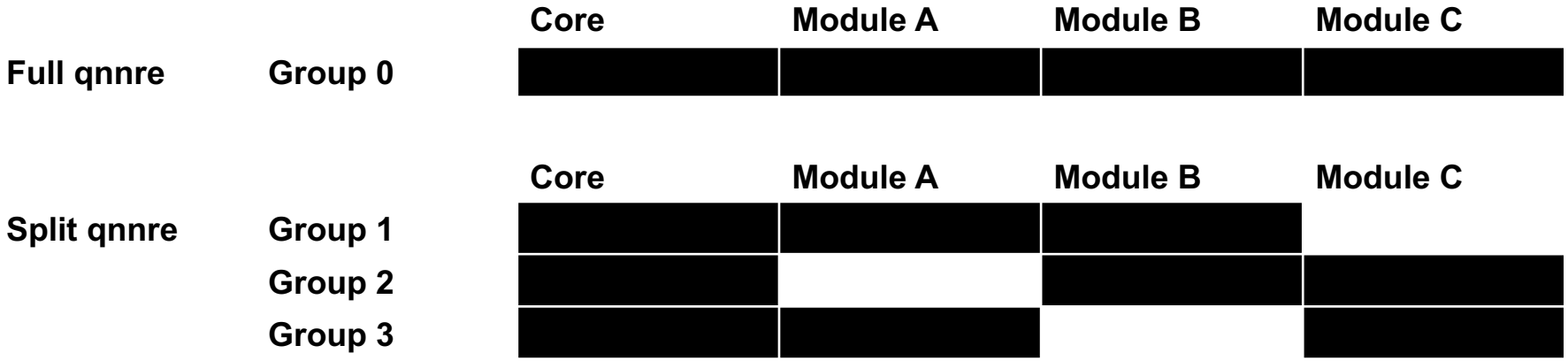
Split Questionnaire Design (Raghunathan and Grizzle, 1995)

- Main objective: shorten the survey instrument to reduce respondent burden while maintaining a rectangular dataset with all survey variables

- Extension of the multiple matrix sampling design (Shoemaker, 1973 and Munger and Lloyd, 1988)

Split Questionnaire Design

- Divide questionnaire into modules
- Administer a subset to each sampled individual, while observing all possible combinations of variables (i.e., bivariate associations)
- Multiply impute data for omitted module(s)



Key Factors to Evaluate Prior to Implementation

- How to create the splits

- How to impute the missing data

- What is the impact on:
  - Nonresponse rates
  - Nonresponse bias
  - Measurement error bias and variance

Can be simulated on existing data

Calls for an experimental design

Creating the Splits

- The cognitive perspective
  - Organize by topic

- The statistical perspective
  - Maximize associations across modules
  - Matrix sampling idea

Trying Marijuana in the National Survey for Drug Use and Health

2002 and Earlier

How do you feel about **adults smoking one or more** packs of cigarettes per day?

How do you feel about **adults trying marijuana or** hashish once or twice?

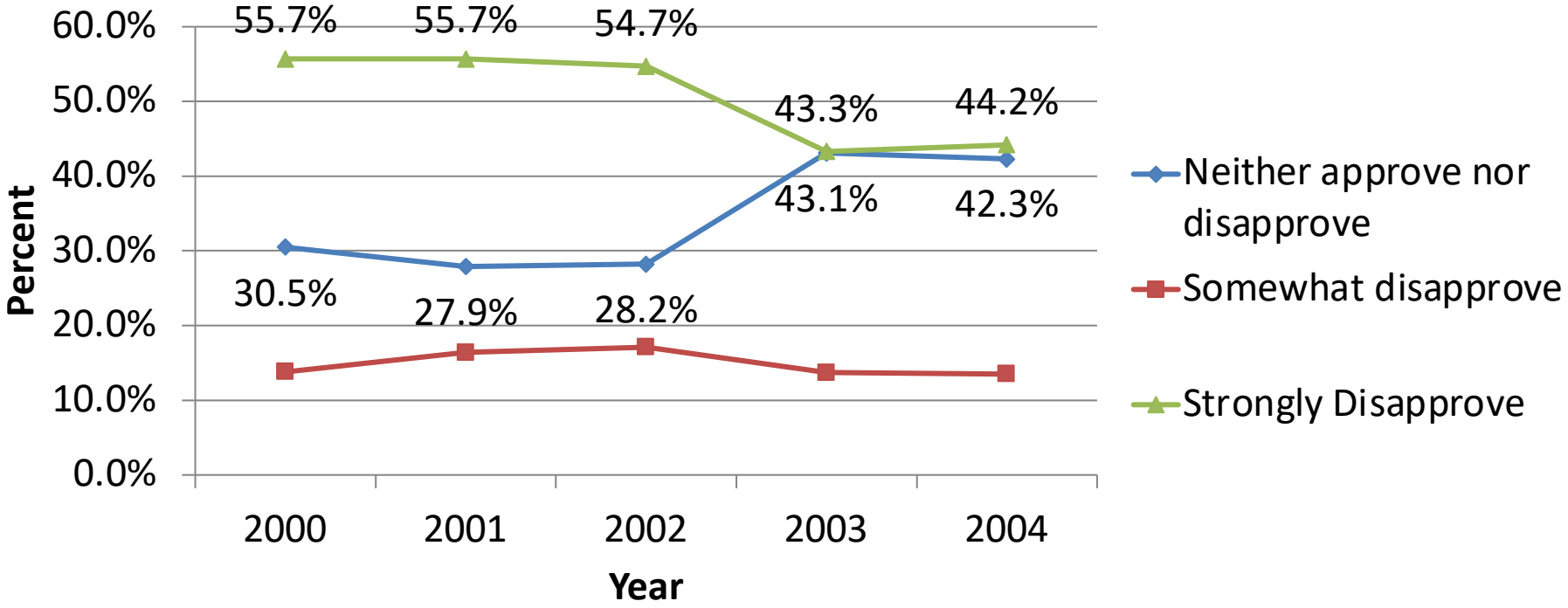
2003 and After

Question on smoking was dropped:

During the past 12 months, how many times have you attacked someone with the intent to seriously injure them?

How do you feel about **adults trying marijuana or** hashish once or twice?

Attitudes Towards Trying Marijuana, 2000-2004 NHSDA/NSDUH



Data source: Wang, K., R. Baxter, and D. Painter. (2005). Modeling Context Effects in the National Survey of Drug Use and Health (NSDUH). Proceedings of the Joint Statistical Meetings.

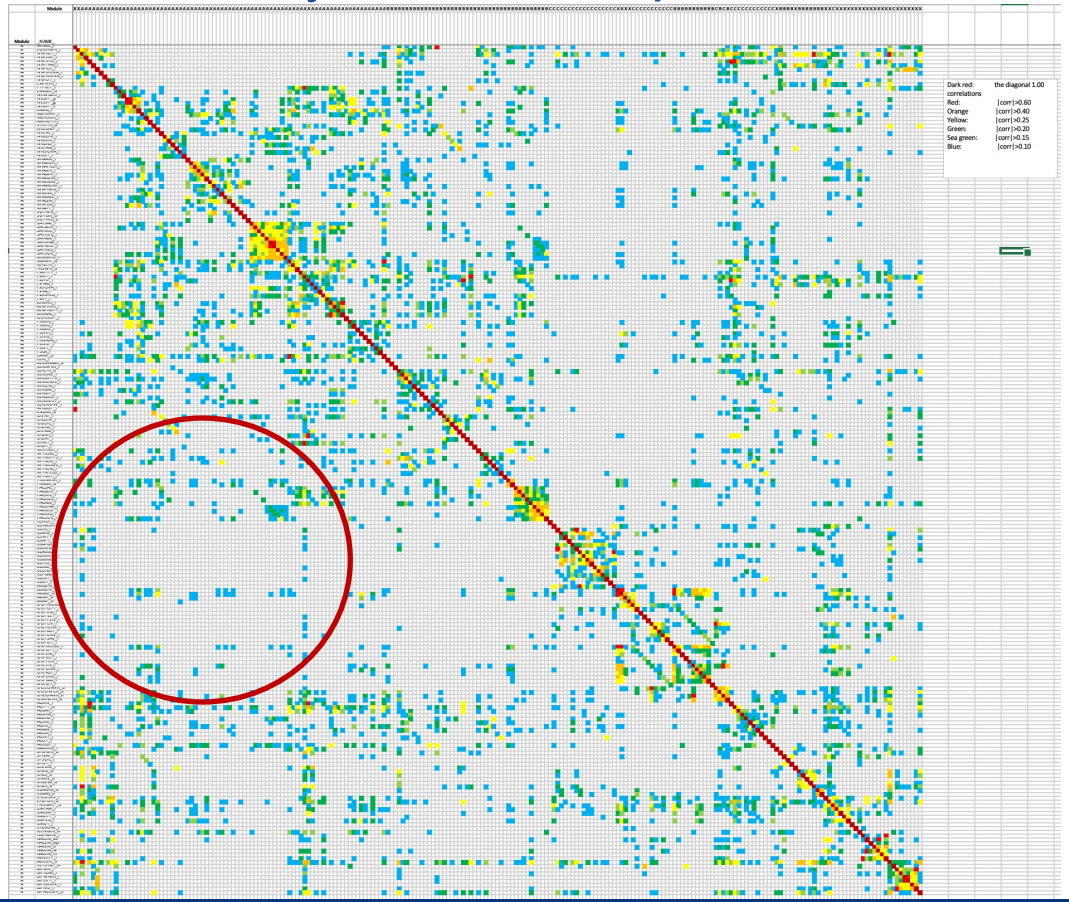
Creating the Splits Revisited

- The cognitive perspective
  - Organize by topic

- ~~The statistical perspective~~
  - ~~Maximize associations across modules~~
  - ~~Matrix sampling idea~~

- Statistically informed splits
  - Organize by topic and modify based on missing associations

Statistically Informed Splits: National Survey of College Graduates

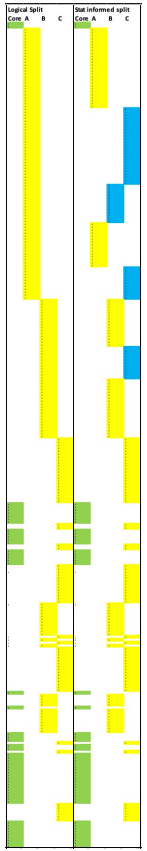


- Correlations between all variables
- Ordered by sequence in the questionnaire
- Heatmap to identify groups of questions that lack associations with questions in other modules

Statistically Informed Splits: National Survey of College Graduates

			t					Number of	Number of	Number of	Number of
			0.15					|corr| >t	|corr| >t	|corr| >t	|corr| >t
Type	Name	sec		mod	otal w/other			Core vars	A vars	B vars	C vars
CORR	FTPRET_r	A	8	A	11	11		2	11	8	1
CORR	EMSIZE_d	A	11	A	2	2		0	12	1	1
CORR	NEWBUS_r	A	12	A	1	1	topic incon	0	5	1	0
CORR	NEDTP_p	A	13.1	A	3	3		1	12	2	0
CORR	NEDTP_g	A	13.2	A	4	4		0	11	3	1
CORR	NEDTP_s	A	13.3	A	2	2		0	12	2	0
CORR	EMED_r	A	14	A	15	15		3	21	8	4
CORR	MGRNAT_r	A	19.1	A	10	10		4	17	2	4
CORR	MGRSOC_r	A	19.2	A	2	2		0	3	2	0
CORR	MGROTH_r	A	19.3	A	5	5		1	7	3	1
CORR	STRTYR_d	A	20	A	10	10		3	2	6	1
CORR	OCEDRLP_r	A	21	A	16	16	Gate, Move	5	11	7	4
CORR	NRPAY_r	A	22.1	A	4	4	Move to C	0	9	4	0
CORR	NRCON_r	A	22.2	A	1	1	Move to C	0	5	1	0
CORR	NRLOC_r	A	22.3	A	2	2	Move to C	0	4	2	0
CORR	NRCHG_r	A	22.4	A	1	1	Move to C	0	6	1	0
CORR	NRFAM_r	A	22.5	A	3	3	Move to C	2	2	1	0
CORR	NROCNA_r	A	22.6	A	1	1	Move to C	0	7	1	0
CORR	NROT_r	A	22.7	A	0	0	Move to C	0	4	0	0
CORR	WAACC_r	A	24.1	A	1	1	Move to C	0	9	0	1
CORR	WABRSH_r	A	24.2	A	4	4	Move to C	1	6	0	3
CORR	WAAPRSH_r	A	24.3	A	2	2	Move to C	1	7	0	1
CORR	WADEV_r	A	24.4	A	0	0	Move to C	0	8	0	0
CORR	WADSN_r	A	24.5	A	3	3	Move to C	1	11	1	1
CORR	WACOM_r	A	24.6	A	6	6	Move to C	3	7	2	1
CORR	WAEMRL_r	A	24.7	A	0	0	Move to C	0	7	0	0
CORR	WAMGMT_r	A	24.8	A	0	0	Move to C	0	9	0	0
CORR	WAPROD_r	A	24.9	A	0	0	Move to C	0	4	0	0
CORR	WASVC_r	A	24.10	A	9	9	Move to C	1	6	6	2
CORR	WASALE_r	A	24.11	A	0	0	Move to C	0	7	0	0
CORR	WAQM_r	A	24.12	A	1	1	Move to C	0	9	0	1
CORR	WATEA_r	A	24.13	A	4	4	Move to C	0	9	4	0
CORR	WAOT_r	A	24.14	A	0	0	Move to C	0	0	0	0
CORR	SUPWK_r	A	26	A	1	1	Move to B	0	6	1	0
CORR	SUPDIR_d	A	27.1	A	0	0	Move to B	0	1	0	0
CORR	SUPIND_d	A	27.2	A	0	0	Move to B	0	2	0	0
CORR	SATSAL_r	A	28.1	A	1	1	Move to B	0	17	0	1
CORR	SATBEN_r	A	28.2	A	5	5	Move to B	0	19	3	2

Logical Split



Statistically Informed Split

Multiple Imputation

Two very different types of approaches with different strengths and weaknesses

- Regression-based imputation
- Weighted sequential hot-deck imputation

Multiple Imputation: National Survey of College Graduates, 2019

- Almost exclusively categorical variables
- Some variables with large number of categories
- Many variables (over 200)
- Many cases (almost 100,000)

- Identifying software and hardware limitations
  - Breaking up processes
  - Choice of software
  - Both

Next Steps

- Complete imputation steps
  - Improve models
  - Finalize imputed datasets
- Evaluate and compare
  - Approach to creating splits
  - Imputation methods
- Offer recommendations
- Disseminate findings

Andy Peytchev
apeytchev@rti.org