MANAGEMENT AND ANALYSIS OF NATIONAL MULTISITE PROGRAM EVALUATION DATA: CENTER FOR SUBSTANCE ABUSE PREVENTION'S DATA ANALYSIS COORDINATION AND CONSOLIDATION CENTER

SESSION CHAIR AND DISCUSSANT:

Beverlie Fallik, Ph.D.

Center for Substance Abuse Prevention
Division of Systems Development

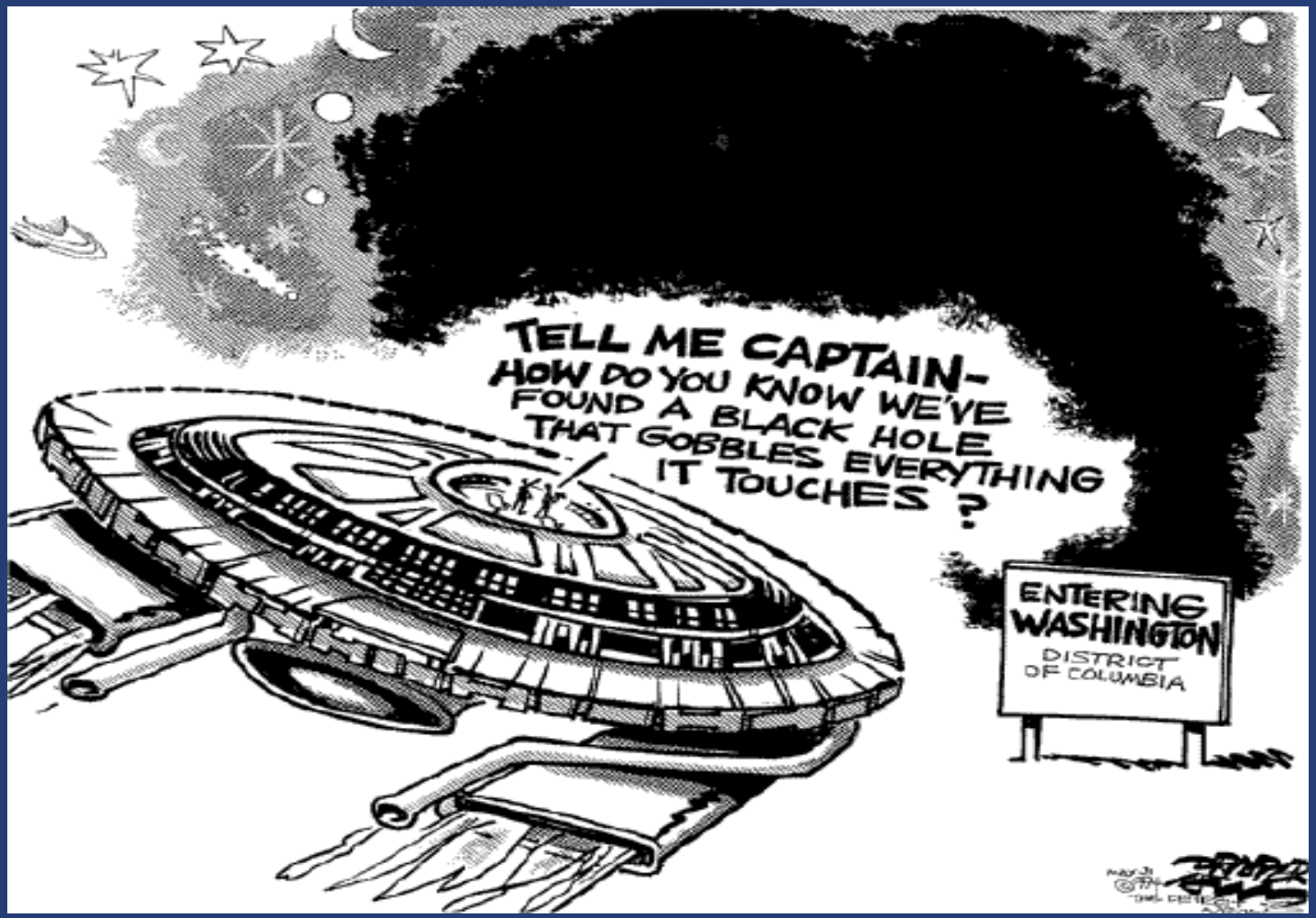
PRESENTERS:

Allison Minugh, Ph.D.
Nilufer Isvan, Ph.D.

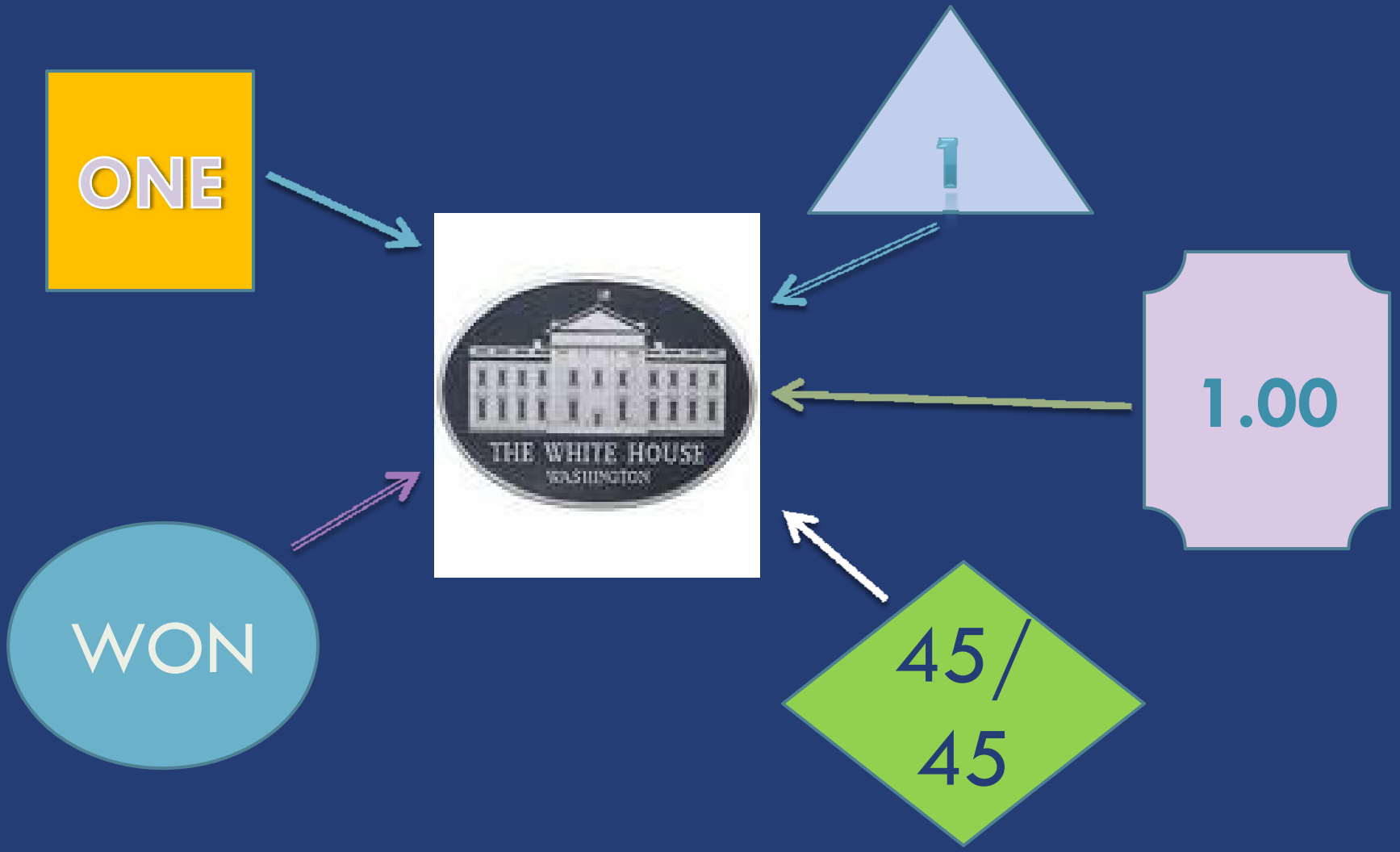
Center for Substance Abuse Prevention
Data Analysis Coordination and Consolidation Center

American Evaluation Association Conference, Orlando, Florida
November 14, 2009

Federal Data Requirements: Grantee Perspective

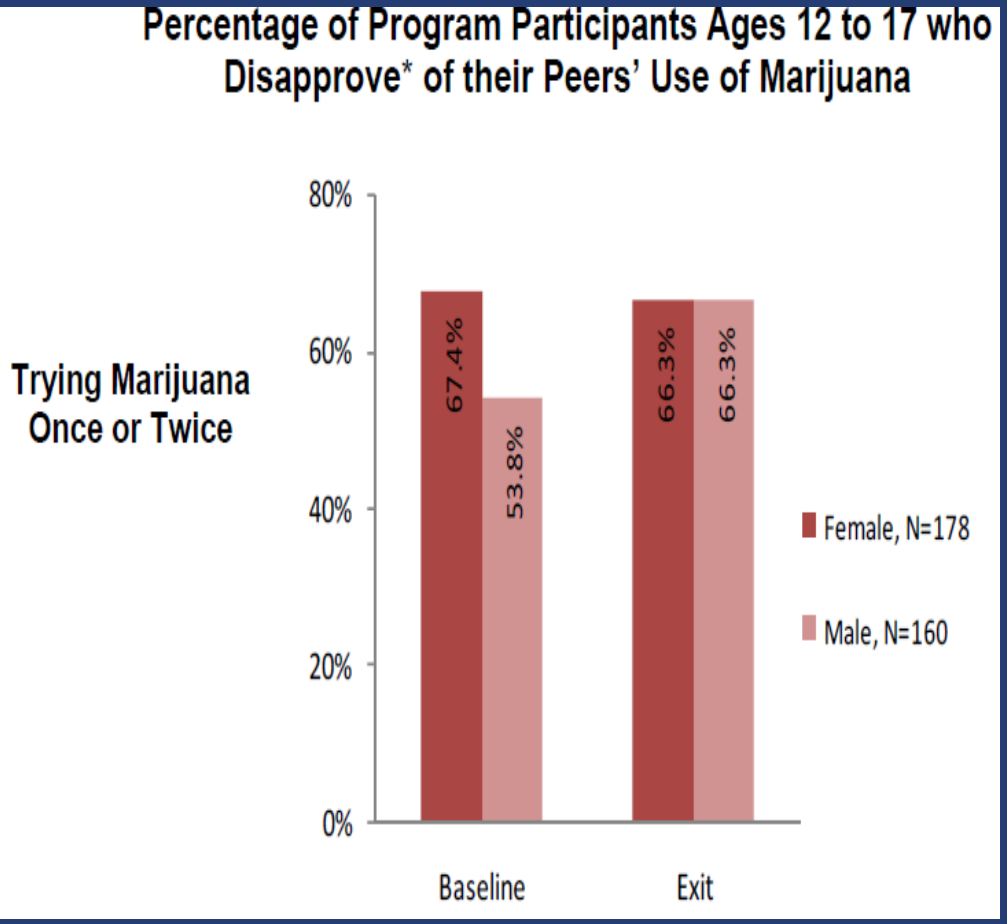


Data Requirements: Incoming data



Is it magic, sleight of hand or skill and hard work?

ALA PEANUT BUTTER SANDWICHES!



CSAP's DACCC

- Process, clean, and consolidate all data submitted by grantees and contractors
- Analyze data for performance assessments and cross site evaluations
- Prepare scheduled, ad hoc and special reports
- Support measure development and review activities
- Provide training and technical assistance to grantees, contractors and SAMSHA/CSAP staff on data related topics
- Work closely with CSAP's Data Information Technology Infrastructure Contract (DITIC)

CSAP's Data Pathway

Grantee/Contractor Data Submissions

Data Information Technology Infrastructure Contract (DITIC)

Coverage Report

Monthly Inventories

Cleaning Sheets to Grantees/ Contractors/POs

Responses to Cleaning Sheets

Application of Cleaning Rules

Cleaning
Matching
Harmonizing

Data Analysis Coordination and Consolidation Center (DACCC)
Data Management Team

Analysis

Data Analysis Coordination and Consolidation Center
Data Analysis Team

Report Production

CSAP's Reports
- Accountability
- NOMs, GPRA, PART
- Congressional Reports
- Program & Policy Decision Support

- The focus of this session is two-fold:
  - Our DMT lead, Allison Minugh Ph.D., will describe the steps, obstacles and solutions undertaken by the DACCC to deal with the myriad types of data issues that have been identified
  - Our DAT lead, Nilufer Isvan, Ph.D., will then discuss how the types of data issues and resolution choices can affect the results of the analyses used to meet accountability requirements.
- Share experiences and solutions: Similar? Different?

DATA QUALITY ASSESSMENT AND DATA MANAGEMENT PRACTICES: AN EXAMPLE FROM THE CENTER FOR SUBSTANCE ABUSE PREVENTION'S PROGRAM EVALUATION DATA

P. Allison Minugh, Ph.D.
Nicoletta A. Lomuto, M.A.
Susan L. Janke, M.S.

Center for Substance Abuse Prevention

Data Analysis Coordination and Consolidation Center

American Evaluation Association Conference, Orlando, Florida
November 14, 2009

National Minority AIDS Initiative

- Established by Congress in 1998
- Designed to address health disparities
- Intended to improve HIV/AIDS health outcomes
- CSAP's program funds 80 grantees

MAI Program Goals

- Deliver sustainable, effective services
- Prevent/reduce substance abuse onset
- Prevent/reduce HIV and Hepatitis transmission
- Target minority and minority re-entry populations
- Target disproportionately affected populations

History of the DACCC Cleaning Rules

What we needed:

- Standardized rules
- Applied CSAP-wide

What we did:

- Reviewed existing survey rules
- Examined scenarios in CSAP's data that appear in national surveys.

NLSY

- Avoid via skip instructions

YRBS

- Mark missing

MTF

- Mark missing

CTC

- Leave as-is

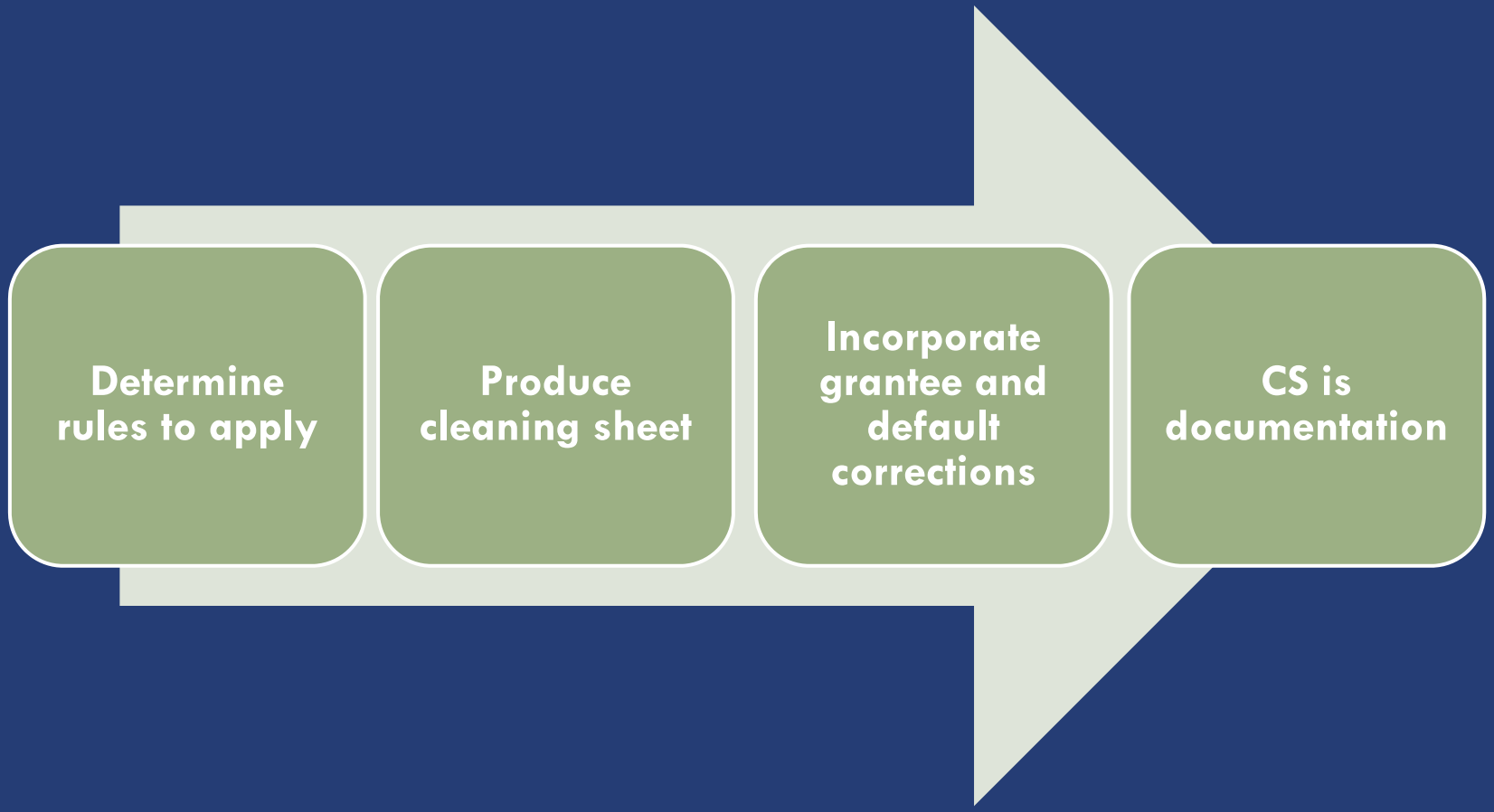
NSDUH

- Multiple approaches

DACCC Approach

- Record level cleaning rules
  - Missing design group
  - Inconsistent design group
  - Duplicated IDs
- Variable level cleaning rules
  - Inconsistent reporting within and across time
  - Outliers
  - Incorrect values

Data Cleaning Steps



Major Data Quality Issues

- Incorrectly formatted ID numbers
- Duplicate ID numbers
- Too much missing data
- Age too young

Common Threats to Data Quality

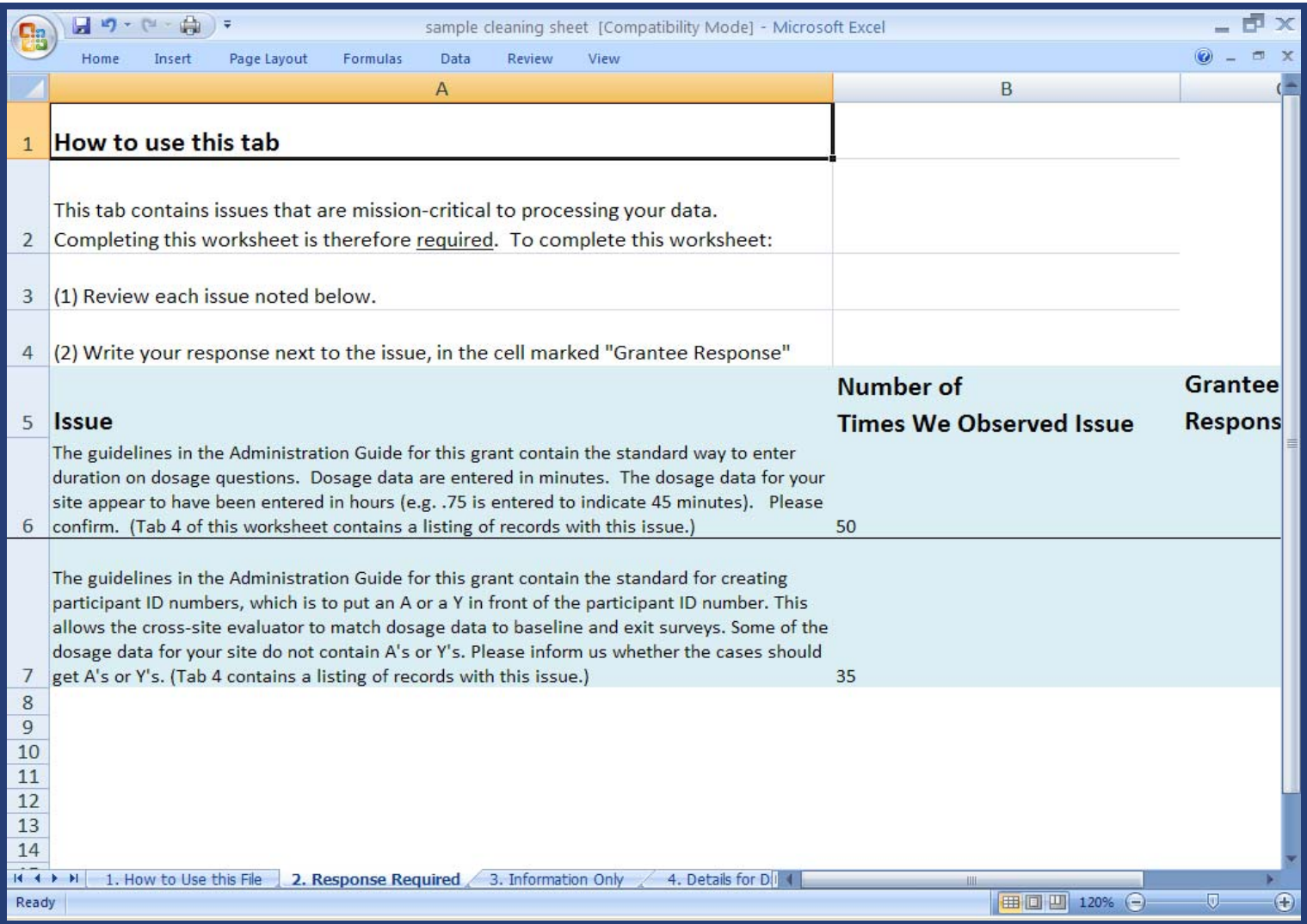
Inconsistent Reporting within a Time Point

- Age of first use older than current age
- Never use on lifetime, use on past 30 days
- No use on general question, use on specific question

Inconsistent Reporting across Time Points

- Demographics
- Age of first use

Sample Cleaning Sheet



sample cleaning sheet [Compatibility Mode] - Microsoft Excel

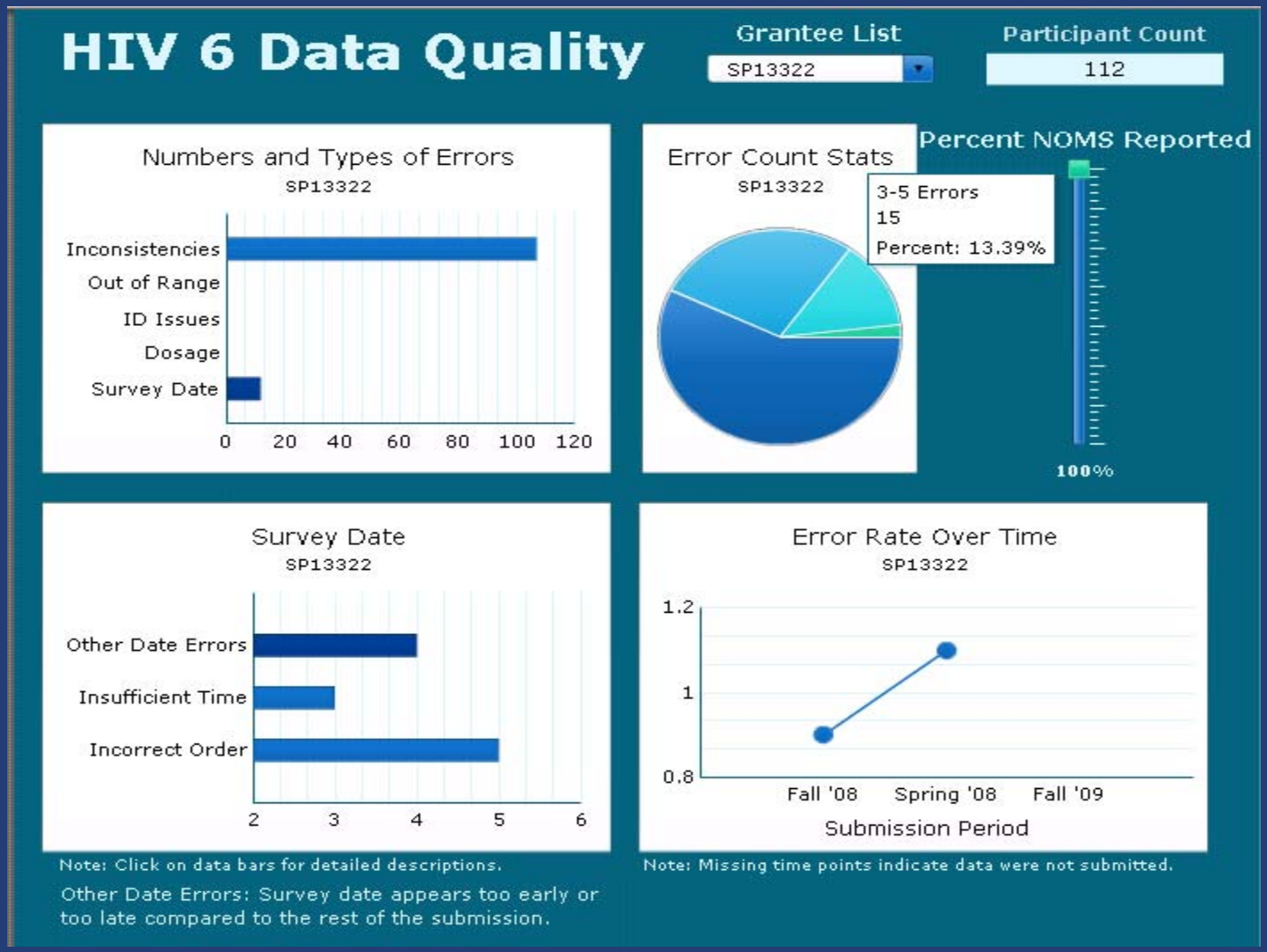
Home Insert Page Layout Formulas Data Review View

	A	B	C
1	**How to use this tab**		
2	This tab contains issues that are mission-critical to processing your data. Completing this worksheet is therefore required. To complete this worksheet:		
3	(1) Review each issue noted below.		
4	(2) Write your response next to the issue, in the cell marked "Grantee Response"		
5	**Issue**	**Number of Times We Observed Issue**	**Grantee Respons**
6	The guidelines in the Administration Guide for this grant contain the standard way to enter duration on dosage questions. Dosage data are entered in minutes. The dosage data for your site appear to have been entered in hours (e.g. .75 is entered to indicate 45 minutes). Please confirm. (Tab 4 of this worksheet contains a listing of records with this issue.)	50	
7	The guidelines in the Administration Guide for this grant contain the standard for creating participant ID numbers, which is to put an A or a Y in front of the participant ID number. This allows the cross-site evaluator to match dosage data to baseline and exit surveys. Some of the dosage data for your site do not contain A's or Y's. Please inform us whether the cases should get A's or Y's. (Tab 4 contains a listing of records with this issue.)	35	
8			
9			
10			
11			
12			
13			
14			

1. How to Use this File | 2. Response Required | 3. Information Only | 4. Details for D

Ready 120%

Data Quality Dashboard



Conclusion

- Reporting to Congress versus Research Methods
- ONDCP Data Quality Audits
- Diversity among Grantees
- Resource Constraints
- Red Herrings

THE IMPACT OF PROGRAM DOSAGE AND INTERVENTION STRATEGY ON PROGRAM OUTCOMES: EXPLORING THE IMPACT OF CSAP'S DATA CLEANING PROCEDURES ON DATA ANALYSIS

Nilufer Isvan, Ph.D.
Lavonia Smith LeBeau, Ph.D.

Center for Substance Abuse Prevention

Data Analysis Coordination and Consolidation Center

American Evaluation Association Conference, Orlando, Florida
November 14, 2009

Analytic Question

To what extent do CSAP's data cleaning procedures affect analysis outcomes?

Analysis Strategy

1. Identify types of questions that are most commonly asked of CSAP's multisite program evaluation data
2. Conduct sample analyses to address each type of question using first raw and then cleaned data
3. Compare the results obtained from raw and cleaned data in terms of
   - Sample sizes
   - Frequency distributions
   - Mean levels of outcome variables
   - Model parameters and test statistics

Typical Questions Addressed by Program Evaluation Data

- What are the demographic characteristics of the individuals served by this program?
- What are the effects of the program on outcome measures?
- What are the predictors of program outcomes?
- Do participants with unmatched records have common characteristics that might result in attrition bias?

Sample Analysis I

Demographic characteristics of people served

Distribution of Race and Ethnicity

	Raw (Baseline)		Cleaned (Cross-time composite)	
	Number	Percent	Number	Percent
Ethnicity				
Hispanic	2,836	30.4	2,958	30.3
Non-Hispanic	6,508	69.6	6,808	69.7
Race				
African American/Black	4,920	54.0	5,108	65.6
American Indian or Alaska Native	272	3.0	284	3.6
Asian	103	1.1	110	1.4
Native Hawaiian or Other Pacific Islander	66	0.7	66	0.8
White	1,769	19.4	1,879	24.1
Other Race	1,634	17.9	N/A	N/A
Multiracial	352	3.9	339	4.4

Distribution of Age and Gender

	Raw (Baseline)		Cleaned (Cross-time composite)	
	Number	Percent	Number	Percent
Age				
17 or younger	1,529	16.7	1,615	16.6
18-25	1,884	20.6	1,962	20.2
26-35	1,686	18.4	1,784	18.4
36-45	2,169	23.7	2,303	23.7
46 or older	1,885	20.6	2,045	21.1
Gender				
Female	4,141	43.8	4,293	43.7
Male	5,220	55.2	5,389	54.9
Transgender	104	1.1	134	1.4

Sample Analysis II

Baseline-to-Exit changes in the frequency of past 30-day substance use

Average Number of Days of Use During the Past 30 Days (Matched-Pairs T-Tests)

	Raw				Cleaned			
	Valid N	Baseline	Exit	Diff. (E - B)	Valid N	Baseline	Exit	Diff. (E - B)
Alcohol	5,048	2.7	2.2	-0.51***	4,907	2.6	2.1	-0.46***
Cigarettes	4,733	10.5	10.4	-0.10	4,761	10.4	10.3	-0.09
Other Tobacco Products	4,819	2.7	2.6	-0.12	4,771	2.5	2.4	-0.06
Marijuana	5,093	2.2	1.6	-0.68***	5,109	2.2	1.6	-0.67***
Other Illicit Substances	5,126	1.8	1.3	-0.47***	5,153	2.2	1.7	-0.51***

*** $p \leq 0.001$, two-tailed matched-pairs t-test

Sample Analysis III

Multivariate analysis predicting program outcomes

OLS Regression Model Predicting Baseline-to-Exit Change in Number of Days of Alcohol Use

	Raw		Cleaned	
	Coefficient	p-value (t-statistic)	Coefficient	p-value (t-statistic)
Total dosage received: One-on-one services (hrs)	-0.399	.045**	-0.544	.016**
Total dosage received: Group-format services (hrs)	-0.007	.888	0.014	.819
Age (yrs)	-0.090	.017**	-0.129	.002***
Ever been in jail for more than 3 days	0.998	.308	1.193	.269
White	-1.617	.232	-2.402	.087*
Living with significant other	1.368	.167	2.055	.061*
Baseline frequency of marijuana (days)	-0.122	.001***	-0.106	.008***
Baseline alcohol-related emotional problems during past 30 days (days)	-0.452	.238	-0.474	.257
Perception of risk of harm from alcohol use	-0.613	.214	-1.033	.058*
Perception of risk of harm from cigarette use	0.894	.090*	1.237	.030**
Constant	3.351	.153	5.090	.045**
R^2	0.050		0.070	
Valid N	525		446	

* $p \leq 0.1$ ** $p \leq 0.05$ *** $p \leq 0.01$

Sample Analysis IV

Multivariate analysis predicting the likelihood of matching baseline and exit records

Logistic Regression Model Predicting the Likelihood of Matching Baseline to Exit Records

	Raw		Cleaned	
	Odds Ratio	p-value (Wald-statistic)	Odds Ratio	p-value (Wald-statistic)
Total dosage received: One-on-one services (hrs)	1.1	.000***	1.2	.000***
Total dosage received: Group-format services (hrs)	1.1	.000***	1.1	.000***
Baseline frequency of cigarettes (days)	1.0	.504	1.0	.157
Baseline frequency of other tobacco products (days)	1.0	.006***	1.0	.067*
Age of alcohol initiation (yrs)	1.2	.151	1.1	.198
Female	1.0	.826	1.0	.830
Age (yrs)	1.0	.000***	1.0	.000***
White	0.7	.009***	0.7	.004***
Hispanic	1.1	.645	1.1	.503
Baseline alcohol-related emotional problems during past 30 days (days)	1.2	.012**	1.2	.019**
Baseline alcohol-related stress during past 30 days (days)	0.9	.067*	0.9	.050**
Attended substance abuse education class prior to program	0.8	.021**	0.8	.056*
Attended HIV education class prior to program	1.2	.169	1.2	.118
Constant	0.5	.000***	0.5	.001***
-2 Log Likelihood	2,562.31		2,197.88	
Valid N	2,193		1,881	

* $p \leq 0.1$ ** $p \leq 0.05$ *** $p \leq 0.01$

Summary: Impact of Data Cleaning on Analysis Results

- Demographic distributions based on cleaned versus raw data are comparable except for race.
- Analysis of cleaned and raw data lead to roughly equivalent conclusions about baseline-to-exit changes in substance use.
- Predictive multivariate analysis using raw versus cleaned data may lead to different conclusions.
- Cleaning the data may improve our ability to match baseline and exit records, thus reducing attrition bias.

Conclusions

Trade-off between data accuracy and data currency.

- For relatively simple distributions and preliminary outcome analysis, using raw data may provide a quick overview of the sample without serious loss of accuracy.
- In some instances, matched comparisons using raw data may involve higher attrition bias.
- Using raw data for more complex analyses such as multivariate modeling may lead to unwarranted conclusions.

Discussion

Program and Policy Implications

- Increased emphasis on real-time data
  - Raw vs. cleaned
  - Direct service vs. environmental strategy
  - Greater than or less than 30 days and pre/post/follow-up data
- Increased emphasis on environmental strategies using epi-data (no control over types of data, samples or frequency of collection)
- Increased emphasis on cost efficiency of programs
- Obtaining overall program results if:
  - Grantees can choose data to report
  - Services, programs, strategies have different frequencies and intensities of dosage
- In short: tension between program-wide findings and relevance at grantee/contract level; between accuracy and speed

Balancing Conflicting Needs

- Provide online data analysis system offering both raw and cleaned data options.
- Submitted data extracted and made immediately available for quick, up-to-date analysis.
- Cleaned, less current data available for more detailed, finalized analysis.
- Users choose one or the other depending on the purpose of their analysis.

WHAT'S YOUR EXPERIENCE????

- How are your data quality issues similar/different?
- How are your cleaning rules developed? Are they similar/different?
- How do you deal with the tension between the demand for real time data vs. data accuracy?
- At what point is the difference between pre-post – follow-up meaningless?
- Other ideas? Suggestions? Observations?
- Questions?
- THANK YOU!