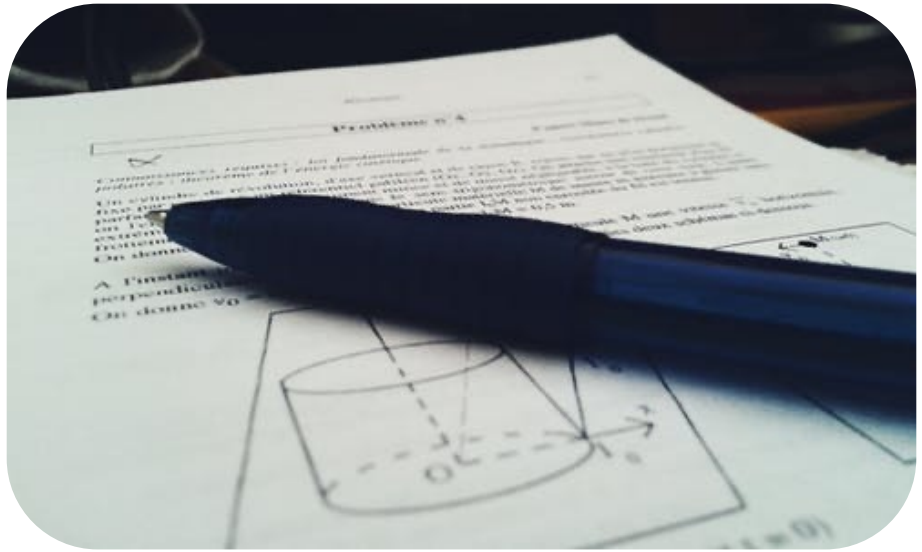
How We Manage Large-Scale Data Collection

Darryl Chamberlain, Ph.D.
Emily Faulconer, Ph.D.
Embry-Riddle Aeronautical University

As a mixed-method study, we are collecting a variety of data across 18 terms.

All data is collected in each term of MATH 111 and PHYS 102 between August 2021 and September 2023.

- Withdrawals in MATH 111 and PHYS 102
- Responses to voluntary, 8 Likert scale question survey
- Student performance data (discussion grades and final grade)
- Transcripts of discussion responses
- Focus-group interviews



For context, PHYS 102 AUG ‘21 had 182 students and 39,682 sentences in discussion responses.

Data cleaning involves standardizing data for analysis and removing incomplete, irrelevant, or erroneous data.

For our data, this primarily included:

- Standardizing withdrawal/drop reports
- Downloading discussion transcripts and parsing them by sentence for analysis
- Organizing all data for cross-semester analysis

Initially, we manually collected and cleaned discussion transcripts.

Create Word Document and list each instructor/student in the cours

akes

Replace names with IDs.

Copy/paste each sentence into an Excel sheet.

Now, we let technology do the heavy lifting.

Save the webpage as a PDF file.

Use an Adobe plugin to convert from a PDF to a Word file.

Run Python scripts on Word file.

Add results to master Excel file

A simple trick to improve your data organization is to put all data into columns.

	A	B	C	D	E	F	H
1	Speaker	Speaker Type	Class ID	Term	Cohort	Module Number	Sentence
13583	QVXCAP	Student	PHY	AUG21	2	8	In the principles of electromagnetism, a magnetic field is induced by a changing elect
13584	QVXCAP	Student	PHY	AUG21	2	8	When a magnet moves through a coil of wires, electric current is induced as long as t
13585	QVXCAP	Student	PHY	AUG21	2	8	Magnetic flux, a measurement of a magnetic field in a given area, depends on the size

I	J	K	L	M	N	O	P	Q	S
Word Count	Emily	Syaza	Abigai	Patric	Amina	Qaisar	Andre	Negotiated Code	COMMENTS
27				IS			IS		
34				IS			IS		
39				IS			IS		

Common calculations such as AVERAGE, STDEV, and COUNTIF are far easier to complete when you can call an entire column.

```
=IF(AND(OR($C$1="Emily", $C$2="Emily"), OR($C$1="Syaza", $C$2="Syaza")), IF($C$3="ALL", CO
IF(AND(OR($C$1="Abigail", $C$2="Abigail"), OR($C$1="Syaza", $C$2="Syaza")), IF($C$3="ALL", C
IF(AND(OR($C$1="Amina", $C$2="Amina"), OR($C$1="Qaisara", $C$2="Qaisara")), IF($C$3="ALL
IF(AND(OR($C$1="Patrick", $C$2="Patrick"), OR($C$1="Andrea", $C$2="Andrea")), IF($C$3="ALL
```

```
COUNTIFS(DATA!$J:$J, IRR!$B22, DATA!$K:$K, IRR!N$5), COUNTIFS(DATA!$D:$D, $C$3, DATA!$J:$J, IRR!$B22, DATA!$K:$K, IRR!N$5)),
", COUNTIFS(DATA!$L:$L, IRR!$B22, DATA!$K:$K, IRR!N$5), COUNTIFS(DATA!$D:$D, $C$3, DATA!$L:$L, IRR!$B22, DATA!$K:$K, IRR!N$5)),
ALL", COUNTIFS(DATA!$N:$N, IRR!$B22, DATA!$O:$O, IRR!N$5), COUNTIFS(DATA!$D:$D, $C$3, DATA!$N:$N, IRR!$B22, DATA!$O:$O, IRR!N$5)),
ALL", COUNTIFS(DATA!$M:$M, IRR!$B22, DATA!$P:$P, IRR!N$5), COUNTIFS(DATA!$D:$D, $C$3, DATA!$M:$M, IRR!$B22, DATA!$P:$P, IRR!N$5)), "NA"))))
```

Strong data organization can allow seamless data analysis.

Term	Cohort		Sentence Count							
OCT20	ALL		25489							

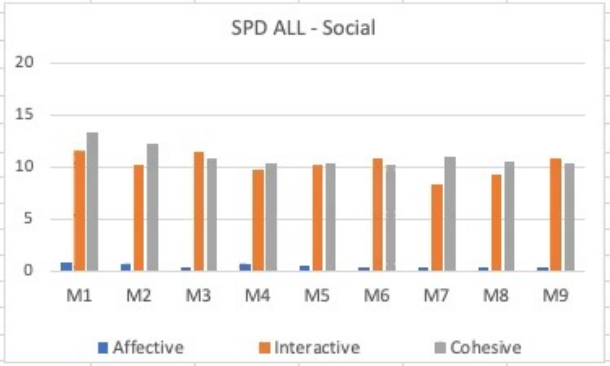
Student Presence Code Counts

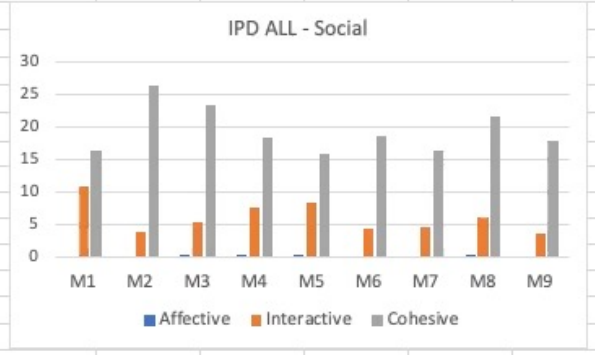
Code	M1	M2	M3	M4	M5	M6	M7	M8	M9	TOTAL
Affective	3%	3%	2%	4%	2%	2%	2%	2%	2%	3%
Interactive	45%	44%	50%	46%	48%	50%	42%	46%	50%	47%
Cohesive	52%	53%	48%	50%	49%	47%	55%	52%	48%	50%
Triggering Event	9%	7%	9%	6%	7%	5%	6%	4%	3%	6%
Exploration	75%	76%	76%	80%	80%	82%	81%	83%	82%	80%
Integration	14%	14%	13%	13%	12%	12%	12%	12%	13%	13%
Resolution	3%	3%	2%	1%	1%	1%	1%	1%	1%	2%

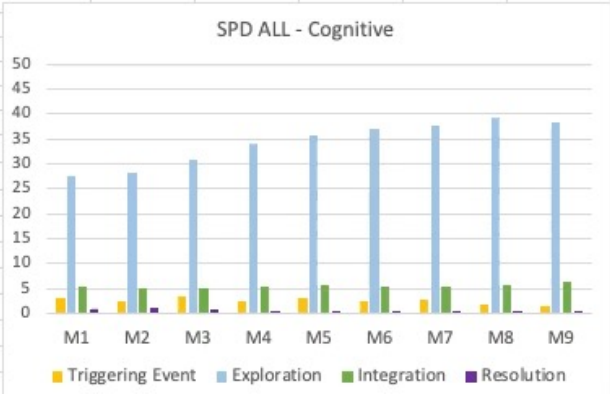
Student Presence Density Aggregated

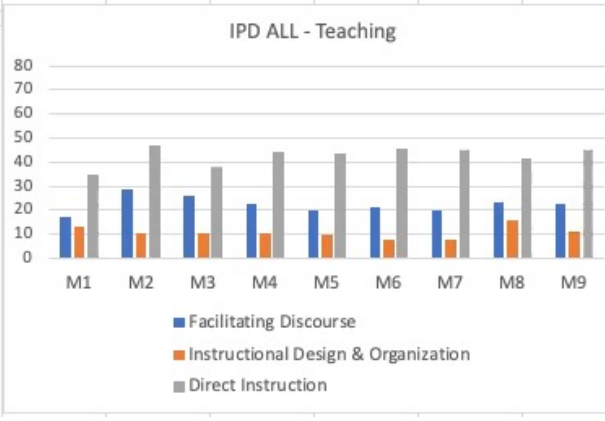
Code	M1	M2	M3	M4	M5	M6	M7	M8	M9	TOTAL
Affective	0.90	0.71	0.49	0.76	0.53	0.50	0.47	0.50	0.51	0.60
Interactive	11.61	10.25	11.51	9.79	10.33	10.95	8.45	9.27	10.86	10.35
Cohesive	13.43	12.28	10.87	10.50	10.49	10.35	11.08	10.55	10.44	11.14
Triggering Event	3.21	2.55	3.56	2.57	3.12	2.45	2.92	1.72	1.59	2.66
Exploration	27.53	28.34	30.93	34.22	35.74	37.05	37.62	39.17	38.32	34.12
Integration	5.22	5.15	5.15	5.34	5.55	5.53	5.53	5.83	6.28	5.49
Resolution	0.97	1.03	0.88	0.58	0.42	0.24	0.45	0.58	0.57	0.64

Student Presence Density by Type

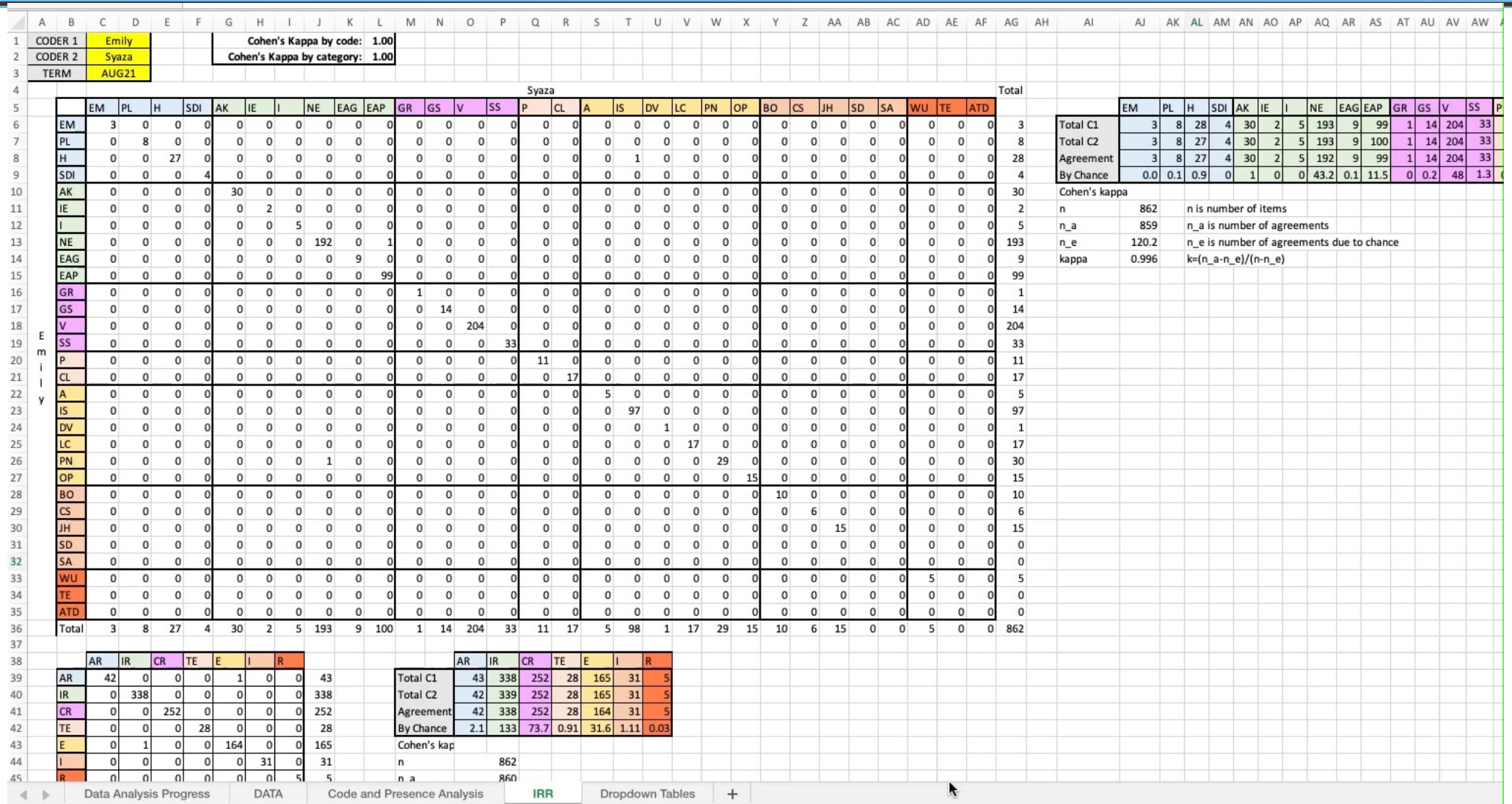








Data cleaning has a direct impact on data analysis.



CODER 1	Emily
CODER 2	Syaza
TERM	AUG21

Cohen's Kappa by code: 1.00
Cohen's Kappa by category: 1.00

Emily \ Syaza

	EM	PL	H	SDI	AK	IE	I	NE	EAG	EAP	GR	GS	V	SS	P	CL	A	IS	DV	LC	PN	OP	BO	CS	JH	SD	SA	WU	TE	ATD	Total
EM	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
PL	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
H	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	28
SDI	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
AK	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30
IE	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
I	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
NE	0	0	0	0	0	0	0	192	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	193
EAG	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
EAP	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99
GR	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
GS	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
V	0	0	0	0	0	0	0	0	0	0	0	0	204	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	204
SS	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
CL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	5
IS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0	0	0	0	0	0	97
DV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
LC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	17
PN	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	30
OP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	15
BO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	10
CS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	6
JH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	15
SD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
TE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ATD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	3	8	27	4	30	2	5	193	9	100	1	14	204	33	11	17	5	98	1	17	29	15	10	6	15	0	0	5	0	0	862

	EM	PL	H	SDI	AK	IE	I	NE	EAG	EAP	GR	GS	V	SS	P	CL	A	IS	DV	LC	PN	OP	BO	CS	JH	SD	SA	WU	TE	ATD
Total C1	3	8	28	4	30	2	5	193	9	99	1	14	204	33	11	17	5	98	1	17	29	15	10	6	15	0	0	5	0	0
Total C2	3	8	27	4	30	2	5	193	9	100	1	14	204	33	11	17	5	98	1	17	29	15	10	6	15	0	0	5	0	0
Agreement	3	8	27	4	30	2	5	192	9	99	1	14	204	33	11	17	5	98	1	17	29	15	10	6	15	0	0	5	0	0
By Chance	0.0	0.1	0.9	0	1	0	0	43.2	0.1	11.5	0	0.2	48	1.3	0.4	0.7	0.3	39.2	0	0.7	1.4	0.6	0.4	0.2	0.6	0	0	0.2	0	0

Cohen's kappa

n	862
n_a	859
n_e	120.2
kappa	0.996

n is number of items
n_a is number of agreements
n_e is number of agreements due to chance
k=(n_a-n_e)/(n-n_e)

	AR	IR	CR	TE	E	I	R
AR	42	0	0	0	1	0	0	43
IR	0	338	0	0	0	0	0	338
CR	0	0	252	0	0	0	0	252
TE	0	0	0	28	0	0	0	28
E	0	1	0	0	164	0	0	165
I	0	0	0	0	0	31	0	31
R	0	0	0	0	0	0	5	5

	AR	IR	CR	TE	E	I	R
Total C1	43	338	252	28	165	31	5
Total C2	42	339	252	28	165	31	5
Agreement	42	338	252	28	164	31	5
By Chance	2.1	133	73.7	0.91	31.6	1.11	0.03

Cohen's kap
n 862
n_a 860

Data Analysis Progress | DATA | Code and Presence Analysis | IRR | Dropdown Tables | +