Tidy data

flickr: Monika Hoinkis

Tidy data

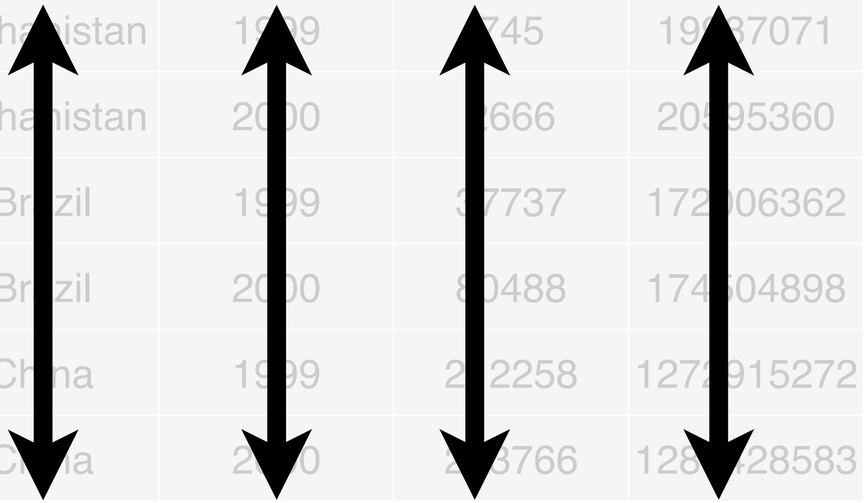
country	year	cases	pop
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

A data set is **tidy** iff:

Hadley Wickham. Tidy data. Journal of statistical software (2014).

Tidy data

country	year	cases	pop
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583



A data set is **tidy** iff:

1. Each **variable** is in its own **column**

Hadley Wickham. Tidy data. Journal of statistical software (2014).

Tidy data

A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**

Hadley Wickham. Tidy data. Journal of statistical software (2014).

Tidy data

A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**
3. Each **value** is in its own **cell**

Hadley Wickham. Tidy data. Journal of statistical software (2014).

Tidy data

country	year	cases	pop
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

A data set is **tidy** iff:

1. Each **variable** is in its own **column**
2. Each **case** is in its own **row**
3. Each **value** is in its own **cell**
4. (Every case is the same type of thing)

Hadley Wickham. Tidy data. Journal of statistical software (2014).

A data example

What variables are represented?

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

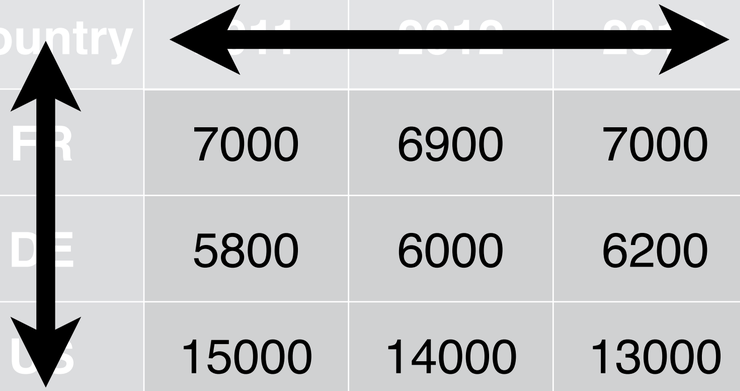
A data example

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

- Country

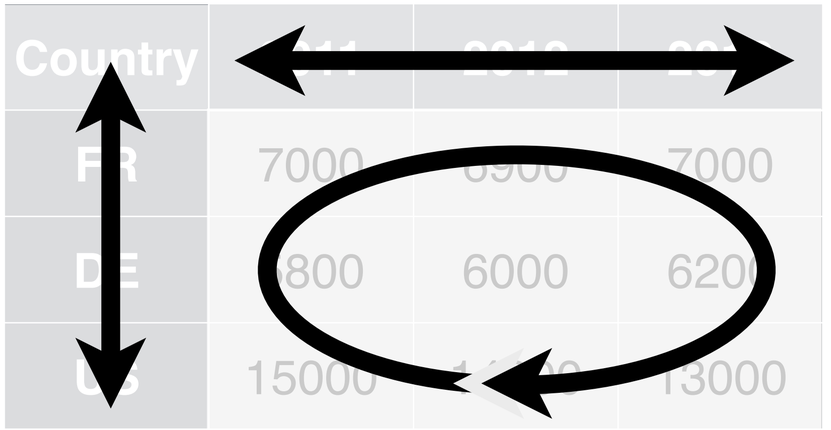
A data example

Country	[illegible]	[illegible]	[illegible]
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000



- Country
- Year

A data example



- Country
- Year
- Count

A data example

On a piece of paper, re-write this dataset with year, country, and count as variables

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

Another example

What are the variables in this dataset?

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

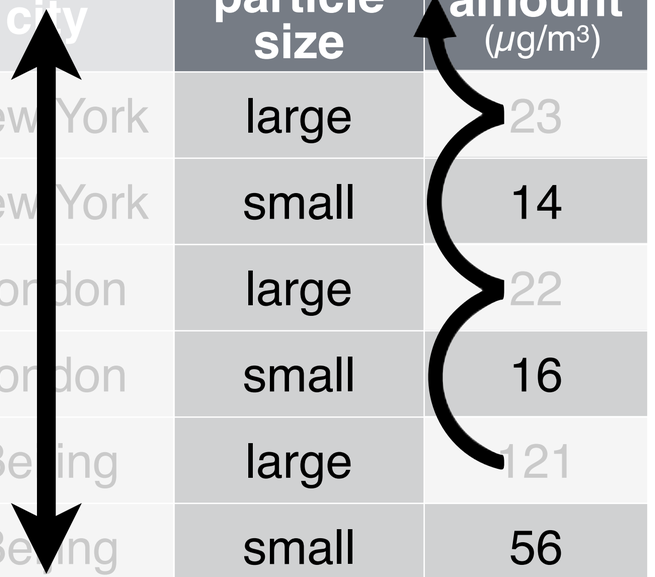
Another example

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

- City

Another example

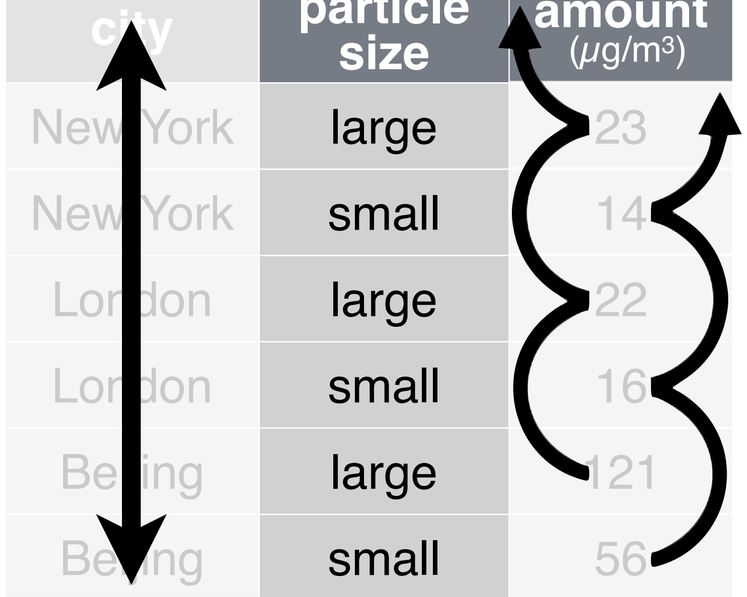
city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56



- City
- Amount of large particulate

Another example

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56



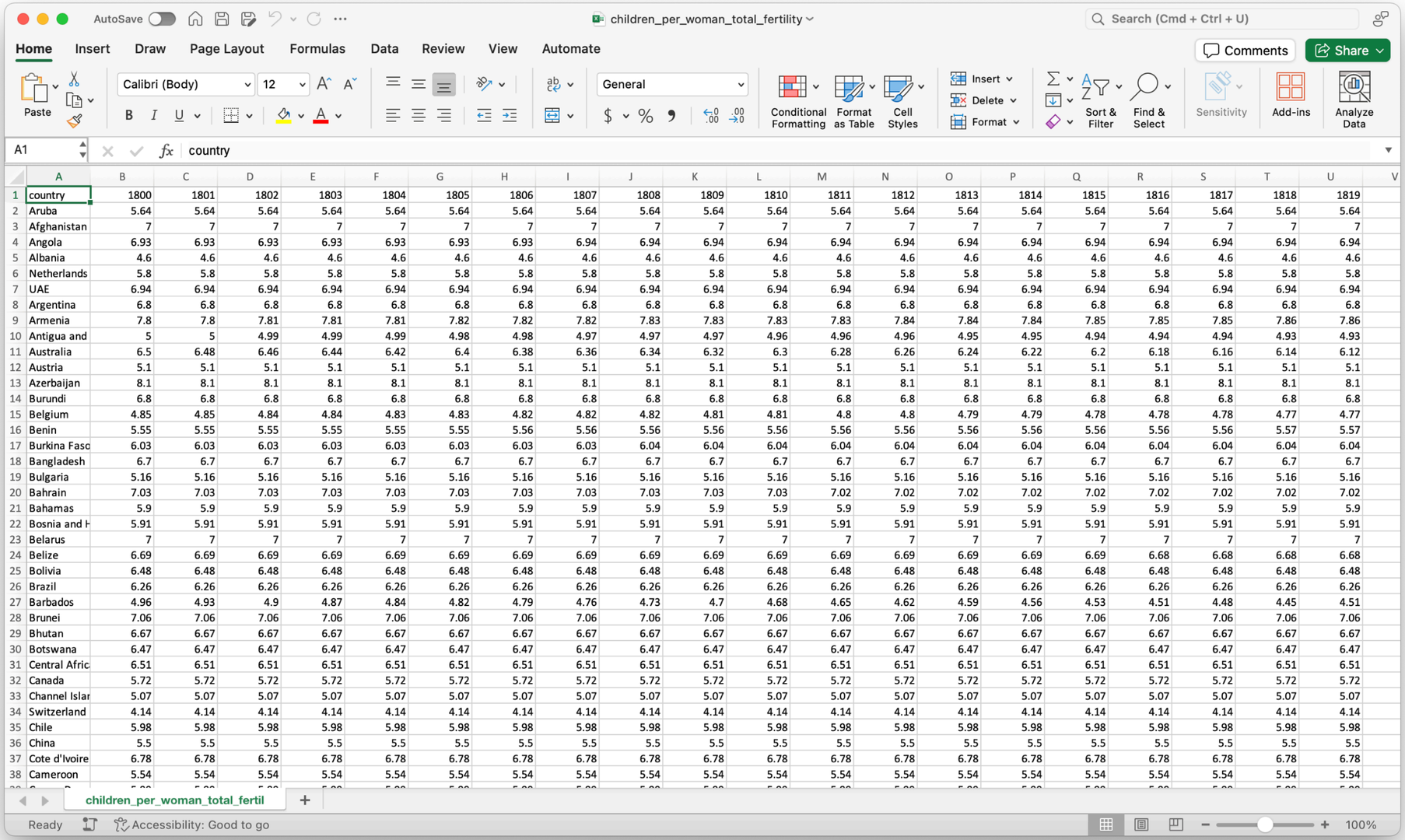
- City
- Amount of large particulate
- Amount of small particulate

Another example

On a piece of paper, re-write this dataset with city, small particulate, and large particulate as variables

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

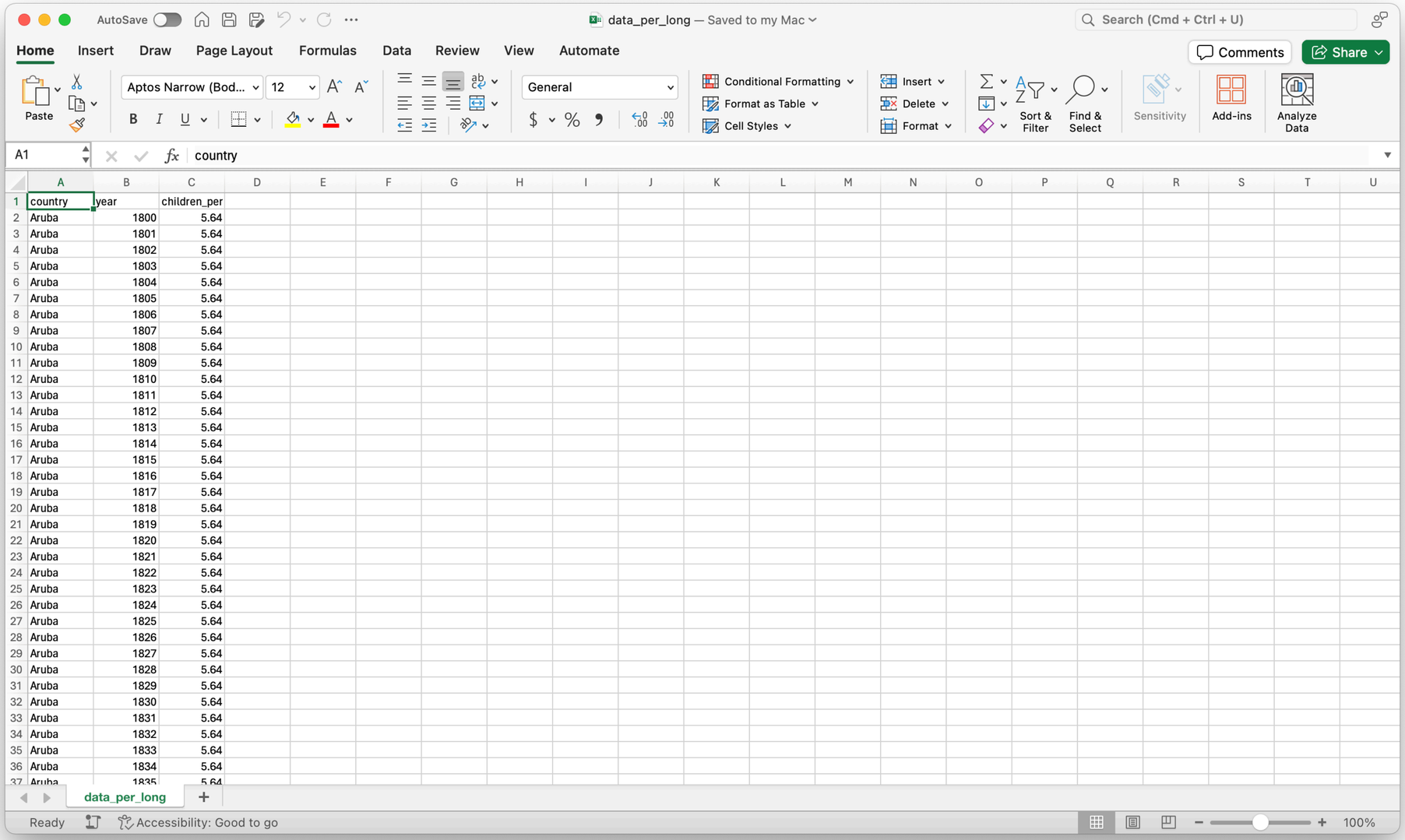
How else could this data be organized?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819
2	Aruba	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64	5.64
3	Afghanistan	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
4	Angola	6.93	6.93	6.93	6.93	6.93	6.93	6.93	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94
5	Albania	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6	4.6
6	Netherlands	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.8
7	UAE	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94	6.94
8	Argentina	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8
9	Armenia	7.8	7.8	7.81	7.81	7.81	7.82	7.82	7.82	7.83	7.83	7.83	7.83	7.84	7.84	7.84	7.85	7.85	7.85	7.86	7.86
10	Antigua and	5	5	4.99	4.99	4.99	4.98	4.98	4.97	4.97	4.97	4.96	4.96	4.96	4.95	4.95	4.94	4.94	4.94	4.93	4.93
11	Australia	6.5	6.48	6.46	6.44	6.42	6.4	6.38	6.36	6.34	6.32	6.3	6.28	6.26	6.24	6.22	6.2	6.18	6.16	6.14	6.12
12	Austria	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1	5.1
13	Azerbaijan	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1	8.1
14	Burundi	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8	6.8
15	Belgium	4.85	4.85	4.84	4.84	4.83	4.83	4.82	4.82	4.82	4.81	4.81	4.8	4.8	4.79	4.79	4.78	4.78	4.78	4.77	4.77
16	Benin	5.55	5.55	5.55	5.55	5.55	5.55	5.56	5.56	5.56	5.56	5.56	5.56	5.56	5.56	5.56	5.56	5.56	5.56	5.57	5.57
17	Burkina Faso	6.03	6.03	6.03	6.03	6.03	6.03	6.03	6.03	6.04	6.04	6.04	6.04	6.04	6.04	6.04	6.04	6.04	6.04	6.04	6.04
18	Bangladesh	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7
19	Bulgaria	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16	5.16
20	Bahrain	7.03	7.03	7.03	7.03	7.03	7.03	7.03	7.03	7.03	7.03	7.03	7.02	7.02	7.02	7.02	7.02	7.02	7.02	7.02	7.02
21	Bahamas	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9	5.9
22	Bosnia and H	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91	5.91
23	Belarus	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
24	Belize	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.69	6.68	6.68	6.68	6.68
25	Bolivia	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48	6.48
26	Brazil	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26
27	Barbados	4.96	4.93	4.9	4.87	4.84	4.82	4.79	4.76	4.73	4.7	4.68	4.65	4.62	4.59	4.56	4.53	4.51	4.48	4.45	4.51
28	Brunei	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06	7.06
29	Bhutan	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67
30	Botswana	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47
31	Central Afric	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51
32	Canada	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72
33	Channel Islar	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07	5.07
34	Switzerland	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14
35	Chile	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98
36	China	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5
37	Cote d'Ivoire	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78	6.78
38	Cameroon	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54	5.54

Babies per woman data from Gapminder https://www.gapminder.org/data/

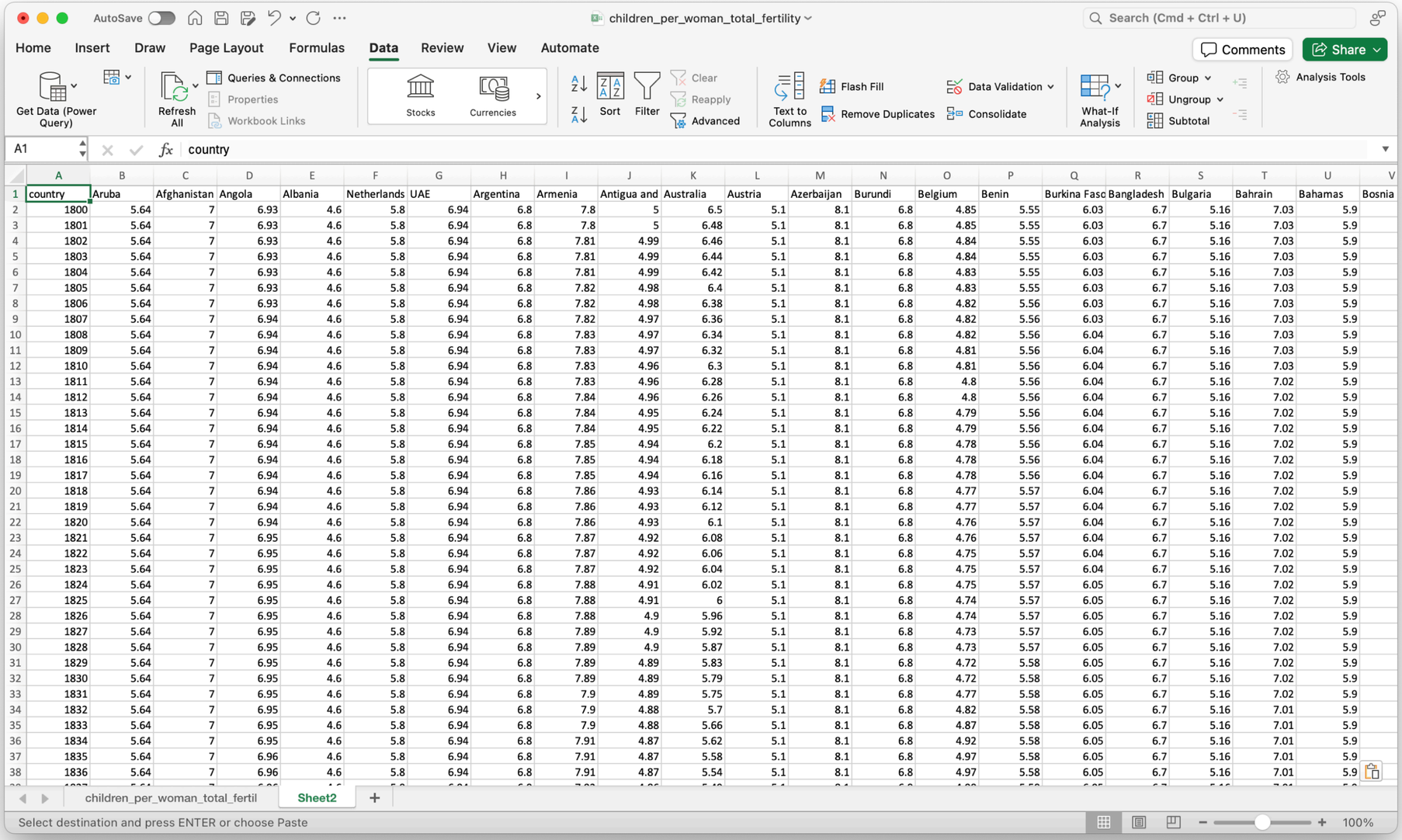
Longer— R would like this



data_per_long — Saved to my Mac

	A	B	C
1	country	year	children_per
2	Aruba	1800	5.64
3	Aruba	1801	5.64
4	Aruba	1802	5.64
5	Aruba	1803	5.64
6	Aruba	1804	5.64
7	Aruba	1805	5.64
8	Aruba	1806	5.64
9	Aruba	1807	5.64
10	Aruba	1808	5.64
11	Aruba	1809	5.64
12	Aruba	1810	5.64
13	Aruba	1811	5.64
14	Aruba	1812	5.64
15	Aruba	1813	5.64
16	Aruba	1814	5.64
17	Aruba	1815	5.64
18	Aruba	1816	5.64
19	Aruba	1817	5.64
20	Aruba	1818	5.64
21	Aruba	1819	5.64
22	Aruba	1820	5.64
23	Aruba	1821	5.64
24	Aruba	1822	5.64
25	Aruba	1823	5.64
26	Aruba	1824	5.64
27	Aruba	1825	5.64
28	Aruba	1826	5.64
29	Aruba	1827	5.64
30	Aruba	1828	5.64
31	Aruba	1829	5.64
32	Aruba	1830	5.64
33	Aruba	1831	5.64
34	Aruba	1832	5.64
35	Aruba	1833	5.64
36	Aruba	1834	5.64
37	Aruba	1835	5.64

Transposed— Datawrapper would like this



country	Aruba	Afghanistan	Angola	Albania	Netherlands	UAE	Argentina	Armenia	Antigua and	Australia	Austria	Azerbaijan	Burundi	Belgium	Benin	Burkina Fasc	Bangladesh	Bulgaria	Bahrain	Bahamas	Bosnia
1800	5.64	7	6.93	4.6	5.8	6.94	6.8	7.8	5	6.5	5.1	8.1	6.8	4.85	5.55	6.03	6.7	5.16	7.03	5.9	
1801	5.64	7	6.93	4.6	5.8	6.94	6.8	7.8	5	6.48	5.1	8.1	6.8	4.85	5.55	6.03	6.7	5.16	7.03	5.9	
1802	5.64	7	6.93	4.6	5.8	6.94	6.8	7.81	4.99	6.46	5.1	8.1	6.8	4.84	5.55	6.03	6.7	5.16	7.03	5.9	
1803	5.64	7	6.93	4.6	5.8	6.94	6.8	7.81	4.99	6.44	5.1	8.1	6.8	4.84	5.55	6.03	6.7	5.16	7.03	5.9	
1804	5.64	7	6.93	4.6	5.8	6.94	6.8	7.81	4.99	6.42	5.1	8.1	6.8	4.83	5.55	6.03	6.7	5.16	7.03	5.9	
1805	5.64	7	6.93	4.6	5.8	6.94	6.8	7.82	4.98	6.4	5.1	8.1	6.8	4.83	5.55	6.03	6.7	5.16	7.03	5.9	
1806	5.64	7	6.93	4.6	5.8	6.94	6.8	7.82	4.98	6.38	5.1	8.1	6.8	4.82	5.56	6.03	6.7	5.16	7.03	5.9	
1807	5.64	7	6.94	4.6	5.8	6.94	6.8	7.82	4.97	6.36	5.1	8.1	6.8	4.82	5.56	6.03	6.7	5.16	7.03	5.9	
1808	5.64	7	6.94	4.6	5.8	6.94	6.8	7.83	4.97	6.34	5.1	8.1	6.8	4.82	5.56	6.04	6.7	5.16	7.03	5.9	
1809	5.64	7	6.94	4.6	5.8	6.94	6.8	7.83	4.97	6.32	5.1	8.1	6.8	4.81	5.56	6.04	6.7	5.16	7.03	5.9	
1810	5.64	7	6.94	4.6	5.8	6.94	6.8	7.83	4.96	6.3	5.1	8.1	6.8	4.81	5.56	6.04	6.7	5.16	7.03	5.9	
1811	5.64	7	6.94	4.6	5.8	6.94	6.8	7.83	4.96	6.28	5.1	8.1	6.8	4.8	5.56	6.04	6.7	5.16	7.02	5.9	
1812	5.64	7	6.94	4.6	5.8	6.94	6.8	7.84	4.96	6.26	5.1	8.1	6.8	4.8	5.56	6.04	6.7	5.16	7.02	5.9	
1813	5.64	7	6.94	4.6	5.8	6.94	6.8	7.84	4.95	6.24	5.1	8.1	6.8	4.79	5.56	6.04	6.7	5.16	7.02	5.9	
1814	5.64	7	6.94	4.6	5.8	6.94	6.8	7.84	4.95	6.22	5.1	8.1	6.8	4.79	5.56	6.04	6.7	5.16	7.02	5.9	
1815	5.64	7	6.94	4.6	5.8	6.94	6.8	7.85	4.94	6.2	5.1	8.1	6.8	4.78	5.56	6.04	6.7	5.16	7.02	5.9	
1816	5.64	7	6.94	4.6	5.8	6.94	6.8	7.85	4.94	6.18	5.1	8.1	6.8	4.78	5.56	6.04	6.7	5.16	7.02	5.9	
1817	5.64	7	6.94	4.6	5.8	6.94	6.8	7.85	4.94	6.16	5.1	8.1	6.8	4.78	5.56	6.04	6.7	5.16	7.02	5.9	
1818	5.64	7	6.94	4.6	5.8	6.94	6.8	7.86	4.93	6.14	5.1	8.1	6.8	4.77	5.57	6.04	6.7	5.16	7.02	5.9	
1819	5.64	7	6.94	4.6	5.8	6.94	6.8	7.86	4.93	6.12	5.1	8.1	6.8	4.77	5.57	6.04	6.7	5.16	7.02	5.9	
1820	5.64	7	6.94	4.6	5.8	6.94	6.8	7.86	4.93	6.1	5.1	8.1	6.8	4.76	5.57	6.04	6.7	5.16	7.02	5.9	
1821	5.64	7	6.95	4.6	5.8	6.94	6.8	7.87	4.92	6.08	5.1	8.1	6.8	4.76	5.57	6.04	6.7	5.16	7.02	5.9	
1822	5.64	7	6.95	4.6	5.8	6.94	6.8	7.87	4.92	6.06	5.1	8.1	6.8	4.75	5.57	6.04	6.7	5.16	7.02	5.9	
1823	5.64	7	6.95	4.6	5.8	6.94	6.8	7.87	4.92	6.04	5.1	8.1	6.8	4.75	5.57	6.04	6.7	5.16	7.02	5.9	
1824	5.64	7	6.95	4.6	5.8	6.94	6.8	7.88	4.91	6.02	5.1	8.1	6.8	4.75	5.57	6.05	6.7	5.16	7.02	5.9	
1825	5.64	7	6.95	4.6	5.8	6.94	6.8	7.88	4.91	6	5.1	8.1	6.8	4.74	5.57	6.05	6.7	5.16	7.02	5.9	
1826	5.64	7	6.95	4.6	5.8	6.94	6.8	7.88	4.9	5.96	5.1	8.1	6.8	4.74	5.57	6.05	6.7	5.16	7.02	5.9	
1827	5.64	7	6.95	4.6	5.8	6.94	6.8	7.89	4.9	5.92	5.1	8.1	6.8	4.73	5.57	6.05	6.7	5.16	7.02	5.9	
1828	5.64	7	6.95	4.6	5.8	6.94	6.8	7.89	4.9	5.87	5.1	8.1	6.8	4.73	5.57	6.05	6.7	5.16	7.02	5.9	
1829	5.64	7	6.95	4.6	5.8	6.94	6.8	7.89	4.89	5.83	5.1	8.1	6.8	4.72	5.58	6.05	6.7	5.16	7.02	5.9	
1830	5.64	7	6.95	4.6	5.8	6.94	6.8	7.89	4.89	5.79	5.1	8.1	6.8	4.72	5.58	6.05	6.7	5.16	7.02	5.9	
1831	5.64	7	6.95	4.6	5.8	6.94	6.8	7.9	4.89	5.75	5.1	8.1	6.8	4.77	5.58	6.05	6.7	5.16	7.02	5.9	
1832	5.64	7	6.95	4.6	5.8	6.94	6.8	7.9	4.88	5.7	5.1	8.1	6.8	4.82	5.58	6.05	6.7	5.16	7.01	5.9	
1833	5.64	7	6.95	4.6	5.8	6.94	6.8	7.9	4.88	5.66	5.1	8.1	6.8	4.87	5.58	6.05	6.7	5.16	7.01	5.9	
1834	5.64	7	6.95	4.6	5.8	6.94	6.8	7.91	4.87	5.62	5.1	8.1	6.8	4.92	5.58	6.05	6.7	5.16	7.01	5.9	
1835	5.64	7	6.96	4.6	5.8	6.94	6.8	7.91	4.87	5.58	5.1	8.1	6.8	4.97	5.58	6.05	6.7	5.16	7.01	5.9	
1836	5.64	7	6.96	4.6	5.8	6.94	6.8	7.91	4.87	5.54	5.1	8.1	6.8	4.97	5.58	6.05	6.7	5.16	7.01	5.9	

- Be consistent.
- Write dates as YYYY-MM-DD.
- Fill in all of the cells.
- Put just one thing in a cell.
- Make it a rectangle.
- Create a data dictionary.
- No calculations in the raw data files.
- Don’t use font color or highlighting as data.
- Choose good names for things.
- Make backups.
- Use data validation to avoid data entry mistakes.
- Save the data in plain text files.

Data Organization in Spreadsheets. Karl Broman and Kara Woo.
https://peerj.com/preprints/3183/
https://doi-org/10.1080/00031305.2017.1375989

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

1. Introduction

Spreadsheets, for all of their mundane rectangularness, have been the subject of angst and controversy for decades. Some writers have admonished that “real programmers don’t use spreadsheets” and that we must “stop that subversive spreadsheet” (Casimir 1992; Chadwick 2003). Others have advised researchers on how to use spreadsheets to improve their productivity (Wagner and Keisler 2006). Amid this debate, spreadsheets have continued to play a significant role in researchers’ workflows, and it is clear that they are a valuable tool that researchers are unlikely to abandon completely.

The dangers of spreadsheets are real, however—so much so that the European Spreadsheet Risks Interest Group keeps a public archive of spreadsheet “horror stories” (*http://www.eusprig.org/horror-stories.htm*). Many researchers have examined error rates in spreadsheets, and Panko (2008) reported that in 13 audits of real-world spreadsheets, an average of 88% contained errors. Popular spreadsheet programs also make certain types of errors easy to commit and difficult to rectify. Microsoft Excel converts some gene names to dates and stores dates differently between operating systems, which can cause problems in downstream analyses (Zeeberg et al. 2004; Woo 2014). Researchers who use spreadsheets should be aware of these common errors and design spreadsheets that are tidy, consistent, and as resistant to mistakes as possible.

Spreadsheets are often used as a multipurpose tool for data entry, storage, analysis, and visualization. Most spreadsheet programs allow users to perform all of these tasks, however we believe that spreadsheets are best suited to data entry and storage, and that analysis and visualization should happen separately. Analyzing and visualizing data in a separate program, or at least in a separate copy of the data file, reduces the risk of contaminating or destroying the raw data in the spreadsheet.

Murrell (2013) contrasted data that are formatted for humans to view by eye with data that are formatted for a computer. He provided an extended example of computer code to extract data from a set of files with complex arrangements. It is important that data analysts be able to work with such complex data files. But if the initial arrangement of the data files is planned with the computer in mind, the later analysis process is simplified.

In this article, we offer practical recommendations for organizing spreadsheet data in a way that both humans and computer programs can read. By following this advice, researchers will create spreadsheets that are less error-prone, easier for computers to process, and easier to share with collaborators and the public. Spreadsheets that adhere to our recommendations will work well with the tidy tools and reproducible methods described elsewhere in this collection and will form the basis of a robust and reproducible analytic workflow.

For an existing dataset whose arrangement could be improved, we recommend against applying tedious and potentially error-prone hand-editing to revise the arrangement. Rather, we hope that the reader might apply these principles when designing the layout for future datasets.

2. Be Consistent

The first rule of data organization is *be consistent*. Whatever you do, do it consistently. Entering and organizing your data in a consistent way from the start will prevent you and your collaborators from having to spend time harmonizing the data later.

Use consistent codes for categorical variables. For a categorical variable like the sex of a mouse in a genetics study, use a single common value for males (e.g., `"male"`), and a single common value for females (e.g., `"female"`). Do not sometimes write “M,”

CONTACT Karl W. Broman kbroman@biostat.wisc.edu Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.