Categorizing respondents as risk-takers or risk-averse

Relevant variables for risky behavior:

"Thinking about the last 7 days... how often have you taken the following measures to protect yourself or others from coronavirus? As a reminder, please exclude any measures that you have already taken for reasons other than coronavirus."

1. **washed_hands:** "... Washed hands with soap and water"
2. **hand_sanitizer:** "... Used hand sanitiser"
3. **avoid_contact_covid:** "... Avoided contact with people who have symptoms or you think may have been exposed to the coronavirus"
4. **avoid_public_transit:** "... Avoided taking public transport"
5. **avoid_guests:** "... Avoided having guests to your home"
6. **avoid_gatherings_small:** "... Avoided small social gatherings (not more than 2 people)"
7. **avoid_gatherings_medium:** "... Avoided medium-sized social gatherings (between 3 and 10 people)"
8. **avoid_gatherings_large:** "... Avoided large-sized social gatherings (more than 10 people)"
9. **avoid_shops:** "... Avoided going to shops"
10. **avoid_public_events:** "... Avoided attending public events, such as sports matches, festivals, theatres, clubs, or going to religious services"
11. **avoid_hh_indoors:** "... Avoided mixing with other households indoors"
12. **maskgrocery:** "... Worn a face mask inside a grocery store / supermarket"
13. **maskclotheshop:** "... Worn a face mask inside a clothing / footwear shop"
14. **masktransit:** "... Worn a face mask on public transportation"

The following answer choices are available for these questions:

- Always
- Frequently
- Sometimes
- Rarely
- Not at all

Our current method uses the last three variables (maskgrocery, maskclotheshop, and masktransit) to categorize respondents as risky or not risky. Ideally, if someone rarely or does not at all wear a mask in these settings, we would know they are risk-takers. However, we are concerned that someone might select these options because they haven't been to the setting. To give an example, if someone says they haven't at all worn a mask in public transport we want to know if this is because they are risky or because they simply haven't been on public transport. To figure this out, we use the question: *"Thinking about the last 7 days... how often have you [avoided taking public transport] to protect yourself or others from coronavirus? As a reminder, please exclude any measures that you have already taken for reasons other than coronavirus."*

Our goal is to use this question to figure out if someone has taken public transport the last 7 days. If we think they have, then they must indicate to have worn a mask at least some of the time. However, if they have not, then their mask-wearing behavior in the setting becomes irrelevant.

Responses 'always' and 'not at all' seem to me as the most straight-forward to interpret:

- Always
  - Person who has not taken public transport because they are being covid-cautious. This person would be considered risk-averse,
    - One issue you raised was that someone might "avoid" public transport for reasons other than covid (e.g., they have a car and drive everywhere) so they might *not* actually be risk averse even if they're never in the setting. Yet, I think the way the question is asked helps us avoid the problem by explicitly asking respondents to "exclude any measures that you have already taken for reasons other than coronavirus". So, if someone does not take public transport because they have a car, their response to the question "Avoided taking public transport" would be "not at all".
- Not at all
  - Person who has taken public transport with the same frequency as they would regardless of covid. This could mean:
    - They haven't taken any public transport for other reasons (e.g., they have a car)
    - They have taken public transport as they normally would

The responses "Frequently", "Sometimes", and "Rarely" are harder to interpret. The most straightforward example:

- Person who used to take the bus every single day but reduced their bus rides because of covid concerns. They didn't reduce bus rides fully or else they would have said "Always" so we know they spent at least some time on the bus.
  - For example, person X used to take the bus to work 5 days a week, but due to covid concerns they were able to get a ride with a friend for 4 out of the 5 days. Because they spent *some* time on the bus, we would expect them to have worn a mask.

But the question is, do we also believe this slightly more convoluted example:

- Person who never used the bus this week but only some of the times due to covid concerns
  - For example, person X was going to take the bus on Wednesday, but the bus was delayed and person X was forced to walk. On Thursday, person X was going to ride the bus, but they got nervous about covid and decided to walk instead.

I honestly don't feel super worried about this kind of example. In general, my sense is that:

- someone who says they 'frequently', 'sometimes', or 'rarely' avoided taking public transport are people who spent at least some time on the bus.
- someone who says they 'always' avoided taking public transport did not spend any time on the bus because they were worried about covid
- someone who did 'not at all' avoid taking public transport might have spent a lot of time on public transport or spent no time on it for reasons other than covid concerns
  - this makes it the least informative response because we don't know what to expect for mask behavior, and we don't know whether they are risk-takers or not

Thus, for each setting we have three possibilities:

- Risk-averse: If they "Always", "Frequently", or "Sometimes" wear a mask and/or they "Always" avoided the setting

- Risk-taker: If they "Frequently", "Sometimes", or "Rarely" avoid the setting and "rarely" or "not at all" wear a mask
- Unknown: If they "Not at all" avoided the setting and "rarely" or "not at all" wear a mask

Now we want to know:

1. How do we categorize those "unknown" respondents?
2. Once we have risk-level for each setting, how do we categorize the respondent overall? If they are risky in one setting? If they are risky in two settings? Etc.

To do this, I checked the different permutations of risk by setting with the average of the following variables (where always = 5 and not at all = 1)

1. washed_hands: "... Washed hands with soap and water"
2. hand_sanitizer: "... Used hand sanitiser"
3. avoid_contact_covid: "... Avoided contact with people who have symptoms or you think may have been exposed to the coronavirus"
4. avoid_guests: "... Avoided having guests to your home"
5. avoid_gatherings_small: "... Avoided small social gatherings (not more than 2 people)"
6. avoid_gatherings_medium: "... Avoided medium-sized social gatherings (between 3 and 10 people)"
7. avoid_gatherings_large: "... Avoided large-sized social gatherings (more than 10 people)"
8. avoid_public_events: "... Avoided attending public events, such as sports matches, festivals, theatres, clubs, or going to religious services"
9. avoid_hh_indoors: "... Avoided mixing with other households indoors"

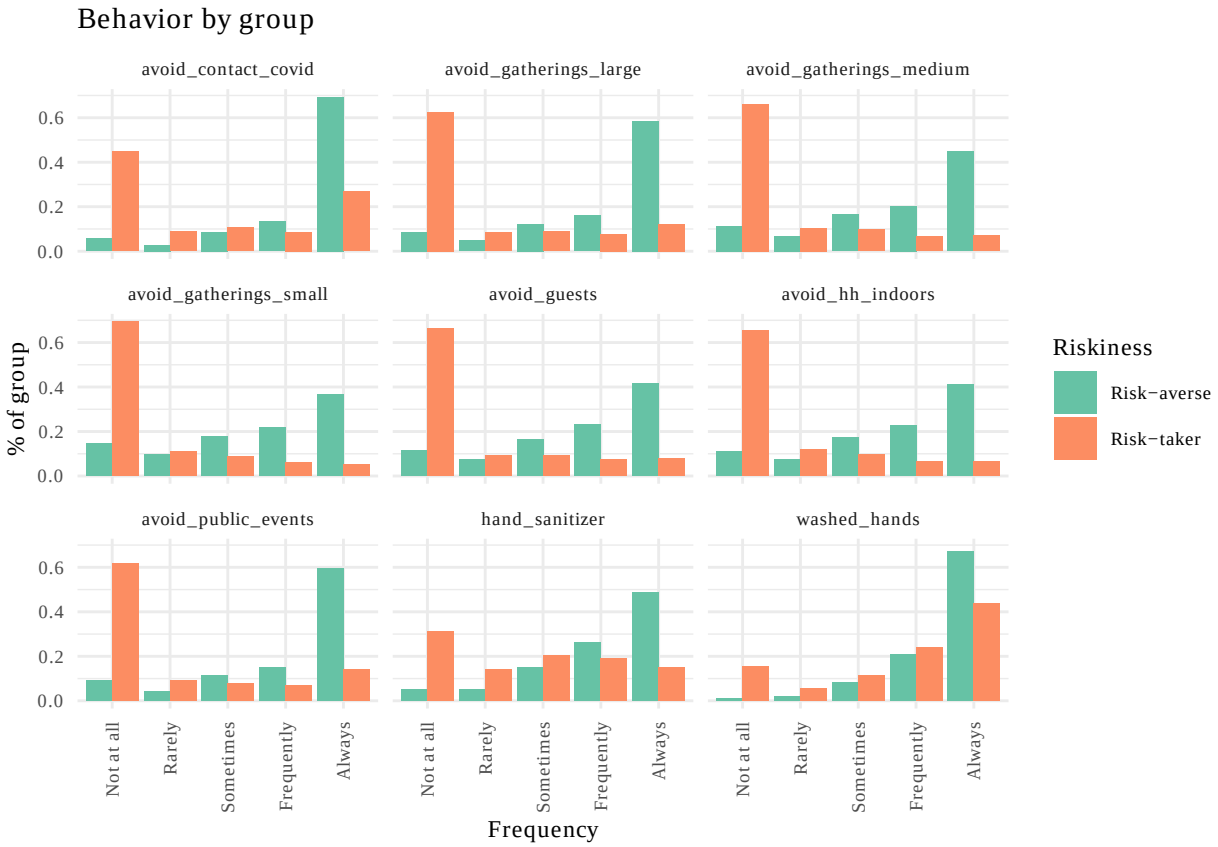
risky_shop	risky_grocery	risky_transport	risky_avg	total
Unknown	Unknown	Unknown	1.340642	1608
Risky	Unknown	Unknown	1.497619	20
Unknown	Risky	Unknown	1.546819	87
Unknown	Unknown	Risky	1.782668	95
Unknown	Risky	Risky	1.834842	17
Unknown	Not risky	Unknown	1.839015	279
Not risky	Unknown	Unknown	2.028571	5
Risky	Risky	Unknown	2.060637	186
Unknown	Unknown	Not risky	2.135494	573
Risky	Unknown	Risky	2.142857	3
Unknown	Risky	Not risky	2.168418	57
Not risky	Unknown	Risky	2.250000	6
Risky	Risky	Risky	2.327240	289
Not risky	Risky	Unknown	2.474490	7
Risky	Unknown	Not risky	2.528571	10
Unknown	Not risky	Risky	2.534203	20
Unknown	Not risky	Not risky	2.590763	239
Not risky	Unknown	Not risky	2.678571	16
Not risky	Not risky	Unknown	2.694022	643
Risky	Not risky	Unknown	2.697043	199
Not risky	Risky	Risky	2.769130	73
Risky	Not risky	Risky	2.796416	156
Risky	Risky	Not risky	3.013507	743
Not risky	Not risky	Risky	3.161895	456

risky_shop	risky_grocery	risky_transport	risky_avg	total
Not risky	Risky	Not risky	3.162235	268
Risky	Not risky	Not risky	3.513036	938
Not risky	Not risky	Not risky	4.190975	17000

Looking at the data, my instinct would be to recategorize "Unknown" as risky for each setting. Then, I would categorize any respondents with two or more risky settings as risky overall (eg, choose based on majority). Doing this we obtain the following totals:

is_risky	count	percent
Not Risky	19560	0.8152378
Risky	4433	0.1847622

Now, looking at different behaviors by group:



Collapsing to look at averages where "Always" = 5 and "Not at all" = 1

behavior	Risk-averse	Risk-taker
avoid_contact_covid	4.376380	2.634108
avoid_gatherings_large	4.109151	1.986691
avoid_gatherings_medium	3.805726	1.778705

behavior	Risk-averse	Risk-taker
avoid_gatherings_small	3.563804	1.664336
avoid_guests	3.757873	1.808933
avoid_hh_indoors	3.757371	1.768433
avoid_public_events	4.116722	2.026830
hand_sanitizer	4.081237	2.722310
washed_hands	4.511401	3.758628

We estimate that around 80% of Americans were risk-averse during the time of the survey. To sanity check our estimate, we use a different survey conducted by the NYTimes on mask wearing behavior, obtaining a similar result:

Risk Averse %	Risk Taker %
0.8193595	0.1805902