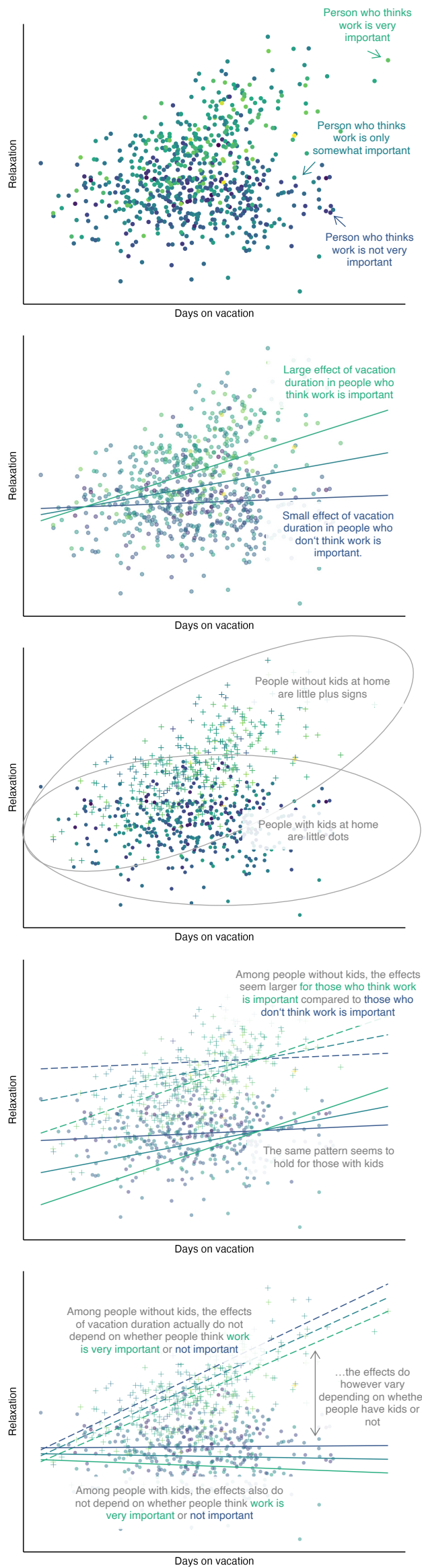


# Controlling for confounding when analyzing interactions



Imagine you are interested in how the duration of vacation affects subsequent relaxation. Let's pretend you had infinite resources, so you actually randomize how long people go on vacation (all expenses paid) and measure relaxation afterwards. This means you don't need to worry about confounding between vacation duration and relaxation.

More importantly, you are interested in whether the effects of vacation duration vary depending on the importance that people assign to work, measured pre-vacation. Maybe the effects differ between people who assign a **very high importance** vs. people who assign only **medium importance** vs. people who assign only **low importance**.

So, you fit the following regression model:

```
lm(relaxation ~ vacation duration + importance of work + vacation duration:importance of work)
```

Success! Your significant interaction indicates that for **people who assign a very high importance to work**, vacation duration has a big effect on relaxation. Maybe they are particularly likely to profit from increased mental distance to work. In contrast, **people who assign low importance to work** seem to profit only little from vacation duration. Maybe because they didn't care about work to begin with?

At this point you realize that you may be missing part of the picture. Some people in your study have kids at home – which, of course, they have to take with them as they go on vacation! – and others have no kids at home.

And, quite reasonably, the experience of going on vacation may vary depending on whether kids are involved.

You assume that kids are a confounder – that is, they may affect both whether people say that work is important and the effects of vacation duration.

So, you additionally adjust for whether or not people have kids at home:

```
lm(relaxation ~ vacation duration + importance of work + vacation duration:importance of work + kids)
```

Indeed, kids make a difference: People without kids at home report a lot more relaxation.

But you still find a significant interaction in which **people who assign a very high importance to work** profit more from vacation duration than **people who assign low importance to work**. So this seems to hold regardless of kids.

Notice that your model so far only allows for a main effect of kids. But your concern is that relaxation might not just depend on kids or vacation duration alone — instead, the two might combine in a way that affects how relaxing the vacation is. You adjust your model to reflect this possibility:

```
lm(relaxation ~ vacation duration + importance of work + vacation duration:importance of work + kids + vacation duration:kids)
```

Now, you find an interaction between kids and vacation duration. For people who travel without kids, the longer the vacation, the more relaxation. This isn't the case for those who travel with kids, who actually do not profit from increased vacation duration at all.

Importantly, it now no longer looks as if **people who assign a very high importance to work** profit more from vacation duration than **people who assign low importance to work**. The regression lines are now parallel. You initial finding was confounded: It looked like **people who assign a very high importance to work** profited more only because those people happened to be the ones who don't have kids, which renders their vacation more relaxing.

Figure by Julia M. Rohrer  
All data are simulated, see code at <https://juliarohrer.com/resources/>