WHY RATINGS ON QUESTIONNAIRE MEASURES SHOULD NOT SERVE AS INDEPENDENT VARIABLES

(UNLESS YOU CORRECT FOR ENDOGENEITY)

John Antonakis
Faculty of Business and Economics
University of Lausanne

16 February 2024

Let’s begin with an example: The data are from Fischer, Dietz, and Antonakis (2024). The below is not reported in the paper (and we demonstrated another point); here I show why it is folly to use questionnaire ratings, which are endogenous, to predict anything. We have an experimental design, where:

1. We reproduce the “script” where we predict y from x as usually done in observational studies
2. But, we control the information environment perfectly
3. The outcome, y, is costly; yet x is not exogenous
4. We show causal illusions when we use x, from ratings, to predict y.

We proceed as follow:

- We use ratings of leadership at T1 to predict an objective outcome in T2 (typical in management journals). We emulate this situation experimentally.

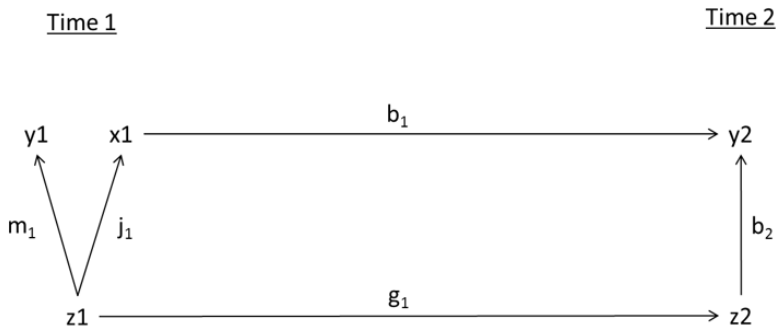
- Randomize participants (n = 409) to watch a video about a leader motivating workers in a mail sorting task to raise money for a charity

- We manipulate charisma (Antonakis, d'Adda, Weber, & Zehnder, 2022), and performance cues (Lord, Binning, Rush, & Thomas, 1978).

- Participants rate leader on various “styles” here, the TLI; the “vision” component (Podsakoff, MacKenzie, Moorman, & Fetter, 1990)—this measure closely models what charisma ratings should capture as outcomes (Banks et al., 2017).
- Participants are paid out, but also receive a bonus which they can keep or donate (for real) to a charity.

We have *full control* over the environment, and what causes the endogenous rating (i.e., the questionnaire measure).

Yet, in practice what do researchers do? They measure x, at time 1, then measure y later. They may even have an objective metric in y (i.e., sales, or a costly outcome); so time and method or source is separated. Yet, they (and reviewers and editors) consistent fall into the *post-hoc ergo propter hoc* fallacy.



Suppose x1 is LMX. Let's think though some causes of it? The only time b_1 will give you a true estimate of the effect on y is if either j_1, g_1, or b_2 are zero; these are rather heroic assumptions to make if x is endogenous.

What is endogenous and exogenous really mean?

- Exogenous (x): varies randomly in nature, is fixed, or is manipulated; is not determined by variables in, or omitted from, the model; does not correlate with the error term, thus the coefficient is consistently estimated.
- Endogenous (z); determined by variables in or omitted from the model

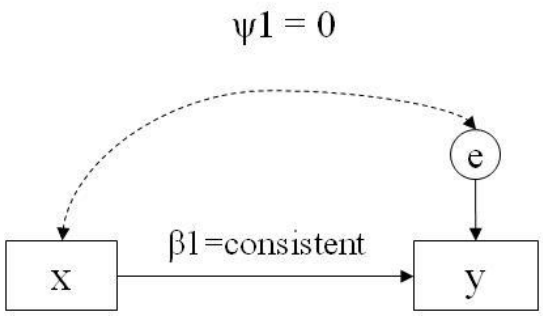


Figure 1A

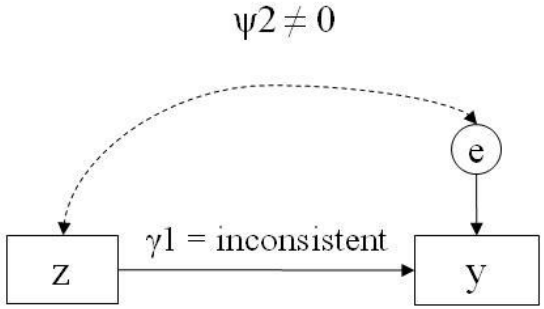


Figure 1B

If a shock in *e* also affects the predictor, then the predictor is endogenous.

What is the problem of using ratings of styles? They are not exogenous!

- The rating of the style is endogenous
- Omitted causes could drive the rating and whether subject donates, and for other reasons, including evaluative judgments due to how questionnaire measures are constructed.
- What are some of these possible omitted causes? Think about this in real-world data; why not use the perceptual rating to predict the outcome?

Yet, many studies in OB use perceptual ratings (from questionnaires) to predict an outcome.

Let's use the endogenous rating (called charisma_rating here) and a LPM. And, voilà; nice and significant results! The rating predicts the donation! YAY!

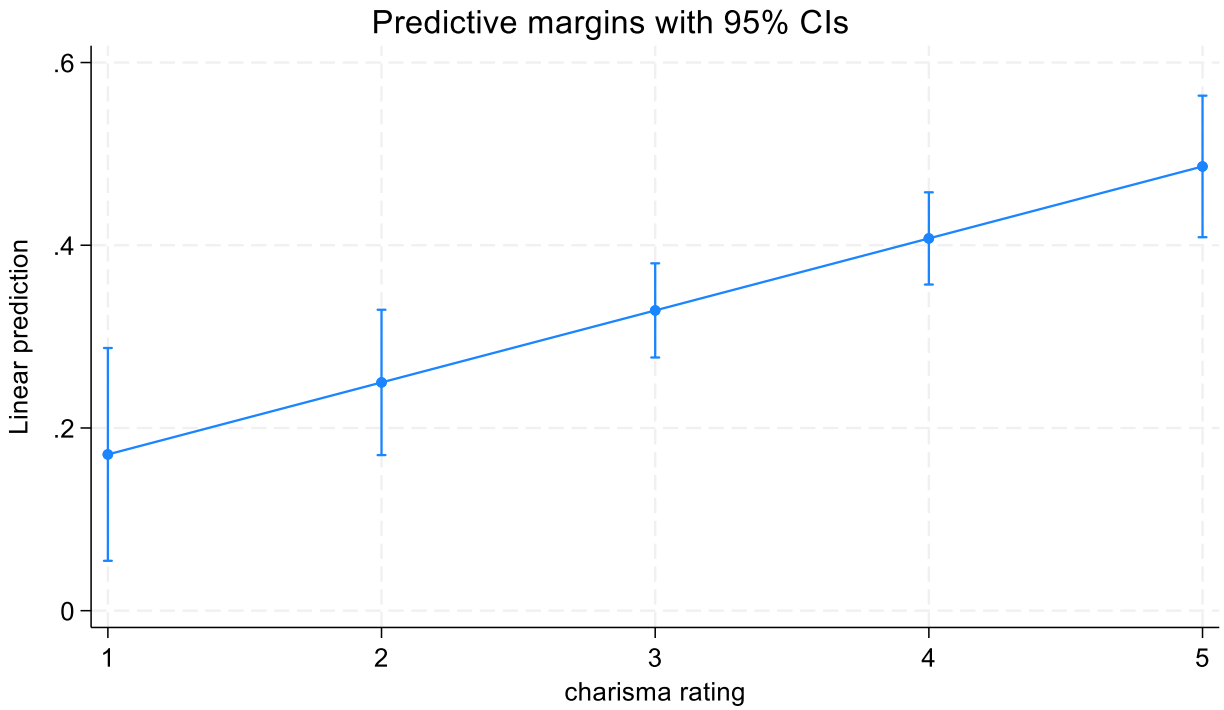
```
. reg donation charisma_rating panas_pos- openness, vce(robust)

Linear regression                               Number of obs     =        409
                                                F(8, 400)         =       2.58
                                                Prob > F          =     0.0094
                                                R-squared         =     0.0433
                                                Root MSE          =     .47982

------------------------------------------------------------------------------
             |               Robust
    donation | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
charisma_rating |   .0788035   .0214677     3.67   0.000     .0365998    .1210071
      panas_pos |   .0082289   .0369836     0.22   0.824    -.0644775    .0809354
      panas_neg |  -.0154792   .0469749    -0.33   0.742    -.1078276    .0768693
         extrav |  -.0085972   .0230565    -0.37   0.709    -.0539243    .0367299
        agreeab |   .0291385   .0256619     1.14   0.257    -.0213106    .0795875
      conscient |   .0415081   .0344327     1.21   0.229    -.0261836    .1091998
         neurot |   .0303796   .0268175     1.13   0.258    -.0223413    .0831005
       openness |  -.0310837   .0258881    -1.20   0.231    -.0819775      .01981
          _cons |  -.1183937   .2240456    -0.53   0.597    -.5588477    .3220603
------------------------------------------------------------------------------
```

Does it really?

It seems so; the marginal effect of charisma



Wow! A change from a -1 to +1 SD in charisma changes probability of donating by +57.20% (from .29 to 46).

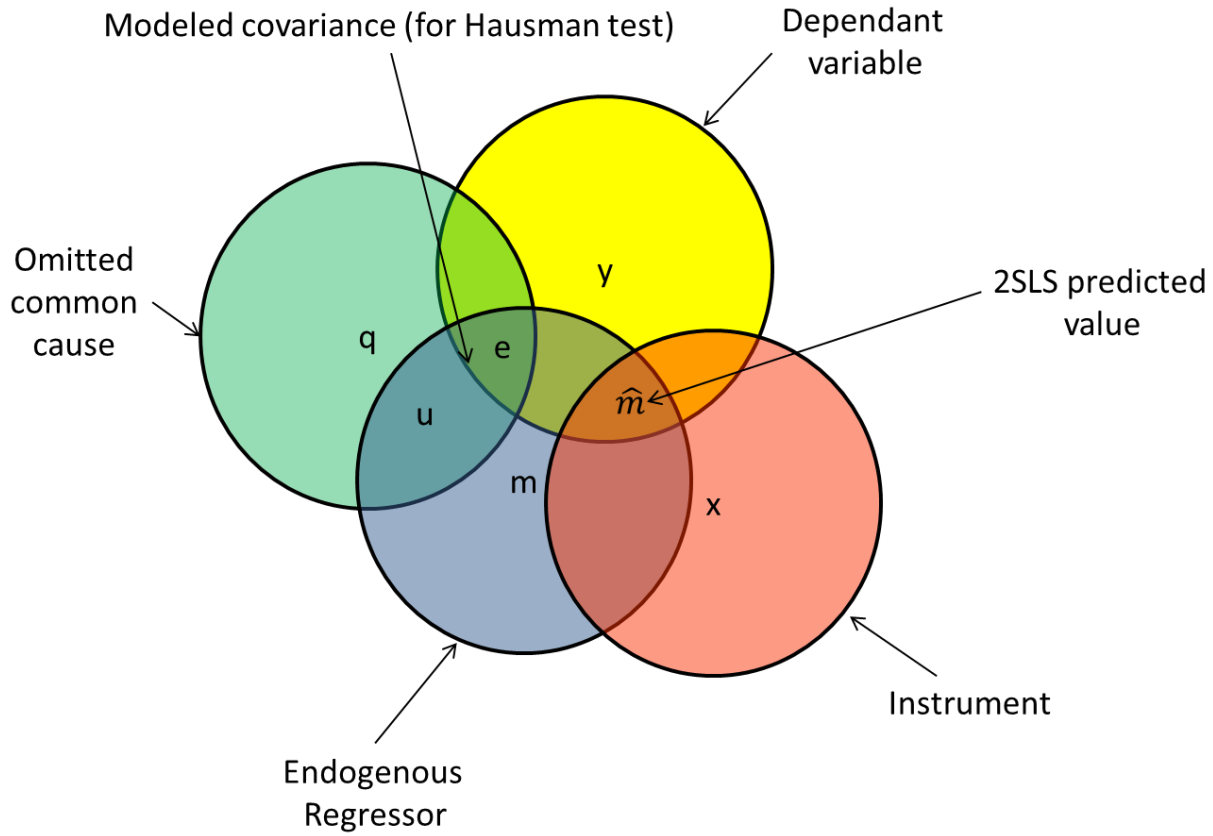
Yet, the average marginal effects of the manipulations show nothing! If the behavioral manipulation shows nothing how can the rating of the behavior show something?

```
reg donation manip_charisma manip_cue panas_pos- openness, vce(robust)

Linear regression                               Number of obs     =        409
                                                F(9, 399)         =       0.88
                                                Prob > F          =     0.5430
                                                R-squared         =     0.0173
                                                Root MSE          =      .4869

------------------------------------------------------------------------------
             |               Robust
    donation | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
manip_charisma |   .0496189   .0486637     1.02   0.309    -.0460503    .1452882
   manip_cue |   .0046598   .0486152     0.10   0.924    -.0909142    .1002339
   panas_pos |   .0288227   .0367977     0.78   0.434     -.043519    .1011644
   panas_neg |  -.0162122   .0465792    -0.35   0.728    -.1077836    .0753592
      extrav |  -.0103089   .0237724    -0.43   0.665    -.0570436    .0364259
     agreeab |   .0301093   .0266212     1.13   0.259     -.022226    .0824447
   conscient |   .0429167   .0349081     1.23   0.220      -.02571    .1115435
      neurot |   .0323433    .027287     1.19   0.237     -.021301    .0859875
    openness |  -.0261269   .0262516    -1.00   0.320    -.0777356    .0254818
       _cons |   .0503764   .2257777     0.22   0.824    -.3934861    .4942389
------------------------------------------------------------------------------
```

Now, the donation model correctly done: Instrumental-variable regression:



In the above case, suppose the instrument/manipulation did cause y. Then $\widehat{m}$ captures the causal effect of x on y. The IV formula is: cov(x,y)/cov(m,x). With the real data we have, however, x does not overlap with y, hence the null result.

```
ivreg2 donation (charisma_rating = i.manip_charisma##i.manip_cue) panas_pos- openness,
robust endog(charisma_rating)

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =       409
                                                      F(  8,   400) =      0.87
                                                      Prob > F      =    0.5404
Total (centered) SS     =   96.2591687                Centered R2   =    0.0229
Total (uncentered) SS   =          155                Uncentered R2 =    0.3932
Residual SS             =  94.05468512                Root MSE      =     .4795

------------------------------------------------------------------------------
             |               Robust
    donation | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
charisma_rating |   .0122155    .037661     0.32   0.746    -.0615987    .0860298
   panas_pos |   .0260532   .0377748     0.69   0.490     -.047984    .1000904
   panas_neg |   -.016802   .0459514    -0.37   0.715     -.106865    .0732611
      extrav |  -.0088497   .0233044    -0.38   0.704    -.0545255    .0368261
     agreeab |   .0290263   .0260708     1.11   0.266    -.0220714    .0801241
   conscient |   .0400632   .0342514     1.17   0.242    -.0270683    .1071948
      neurot |   .0322454   .0268252     1.20   0.229    -.0203311    .0848219
    openness |  -.0263917   .0261269    -1.01   0.312    -.0775996    .0248161
       _cons |   .0543241   .2387671     0.23   0.820    -.4136509     .522299
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):           149.802
                                                      Chi-sq(3) P-val =    0.0000
```

Diagnostics for instrumental variable regression are good.

```
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                75.346
                         (Kleibergen-Paap rk Wald F statistic):          83.733
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias    13.91
                                         10% maximal IV relative bias     9.08
                                         20% maximal IV relative bias     6.46
                                         30% maximal IV relative bias     5.39
                                         10% maximal IV size             22.30
                                         15% maximal IV size             12.83
                                         20% maximal IV size              9.54
                                         25% maximal IV size              7.80
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         1.462
                                                    Chi-sq(2) P-val =    0.4815
-endog- option:
Endogeneity test of endogenous regressors:                               5.098
                                                    Chi-sq(1) P-val =    0.0240
Regressors tested:    charisma_rating
------------------------------------------------------------------------------
Instrumented:         charisma_rating
Included instruments: panas_pos panas_neg extrav agreeab conscient neurot
                      openness
Excluded instruments: 1.manip_charisma 1.manip_cue 1.manip_charisma#1.manip_cue
```

See Bastardoz et al. (2023).

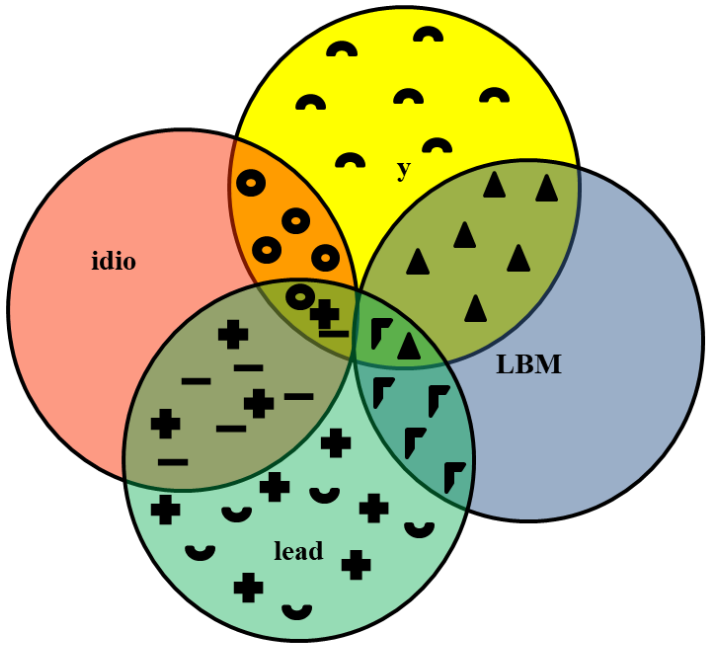
Another way to understand the problem (Fischer et al., 2024).

- Regress charisma ratings on the manipulations and controls; save the residuals.
- What to the residuals capture? All idiosyncratic causes of the charisma rating not due to the manipulations and controls
- If the residuals predict the donation, then we know it was the idiosyncratic variation provided by the rater that correlates with y. Is that a problem? Yes! Because it is not the behavior that is being rated!
- Where is “behavior” in organizational behavior? (Banks, Woznyj, & Mansfield, 2021). See also Fischer (2023)—a great piece.

Assume: LBM = Leadership behavior Measure; Lead = leadership style measure

Idio = Idiosyncratic variation

Case 1: LBM is a cause of *y* and of *lead*



Note:

✚ *Residual* (from regressing *lead* on *LBM*)

– Overlap of Idiosyncratic Variation with Rated Behavior

o Variation in y due to Idiosyncratic variation

▲ Variation in y due to LBM

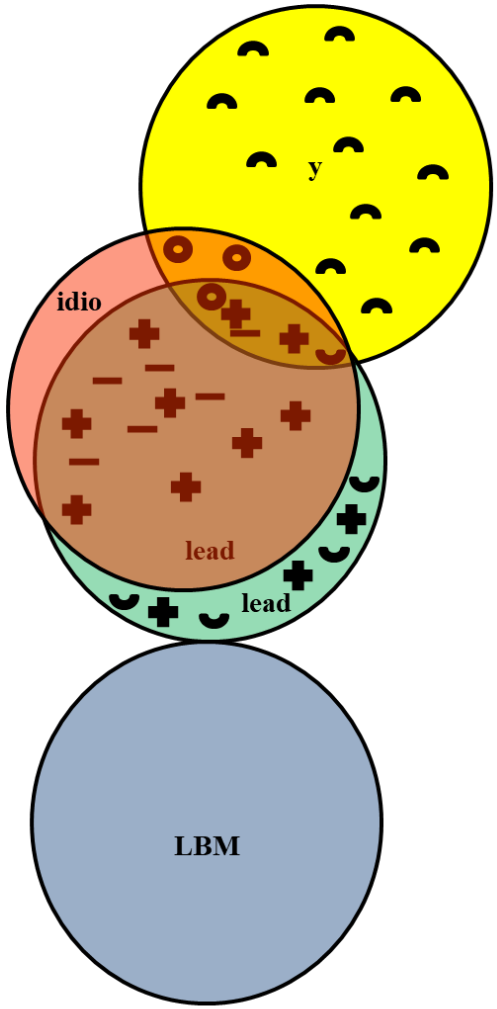
⌜ Variation in lead due to LBM

✚o – Information used to estimate the relation between residual of rated leader behavior and y

⌜▲ Information used by instrumental-variable estimator

⌒⌣ Measurement error

Case 2: LBM is neither a cause of *y* nor of *lead* (what I just showed you with our data)



Note:

✚ *Residual* (from regressing *lead* on *LBM*)

▬ Overlap of Idiosyncratic Variation with Rated Behavior

o Variation in y due to Idiosyncratic variation

▲ Variation in y due to LBM*

◤ Variation in lead due to LBM*

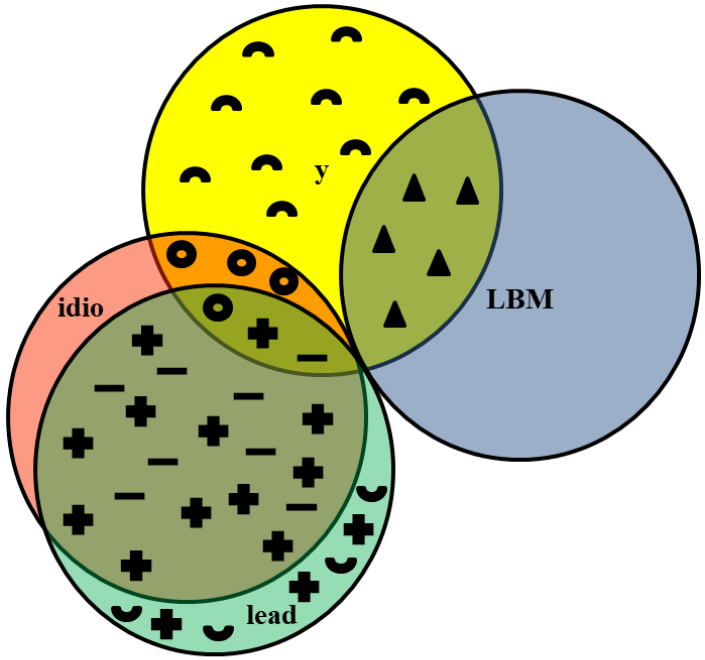
✚ o ▬ ◡ Information used to estimate the relation between residual of rated leader behavior and y

◤ ▲ Information used by instrumental-variable estimator*

◠ ◡ Measurement error

*Absent

Case 3: LBM is a cause of *y* and not of *lead*



Note:

✚ *Residual* (from regressing *lead* on *LBM*)

▬ Overlap of Idiosyncratic Variation with Rated Behavior

o Variation in y due to Idiosyncratic variation

▲ Variation in y due to LBM

◤ Variation in lead due to LBM*

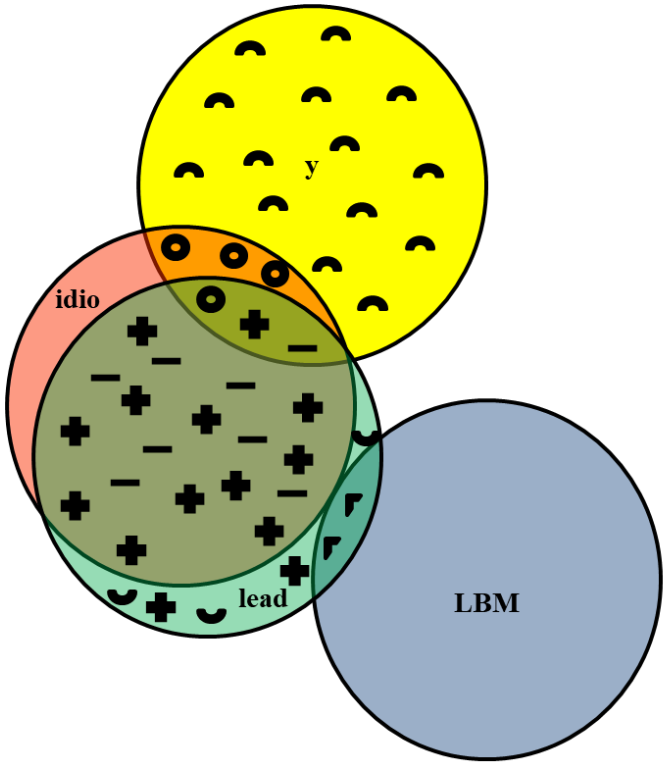
✚o ▬ Information used to estimate the relation between residual of rated leader behavior and y

◤▲ Information used by instrumental-variable estimator*

◠◡ Measurement error

*Absent

Case 4: LBM is a cause of lead but not *y*



Note:

✚ *Residual* (from regressing *lead* on *LBM*)

— Overlap of Idiosyncratic Variation with Rated Behavior

o Variation in y due to Idiosyncratic variation

▲ Variation in y due to LBM*

◤ Variation in lead due to LBM

✚o — Information used to estimate the relation between residual of rated leader behavior and y

◤▲ Information used by instrumental-variable estimator*

◠◡ Measurement error

*Absent

Let’s use this insight to do an example with Case 1: Where the manipulation is a cause of the leader rating and y. Data is from S2, Meslec, Curseu, Fodor, and Kenda (2020). We manipulate:

1. Charisma (leader manipulation)
2. Incentives (money manipulation)
3. We measure a costly outcome.

And also elicit measures charisma. Do these measures measure behavior?

The reduced form effect:

```
. reg performance leader money

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(2, 271)       =     94.53
       Model |  492884.567         2  246442.284   Prob > F        =    0.0000
    Residual |  706474.484       271  2606.91691   R-squared       =    0.4110
-------------+----------------------------------   Adj R-squared   =    0.4066
       Total |  1199359.05       273   4393.2566   Root MSE        =    51.058

------------------------------------------------------------------------------
 performance | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      leader |   54.56704   6.172284     8.84   0.000     42.41532    66.71876
       money |   64.22236   6.171132    10.41   0.000      52.0729    76.37181
       _cons |   132.8715   5.229363    25.41   0.000     122.5761    143.1668
------------------------------------------------------------------------------
```

So we know that the leader (charisma) manipulation has a cause effect.

Compared to the control treatment, the leader treatment induces 54.57 higher performance.

Now, let's regress the leader rating on performance (and control for the money manipulation):

```
. reg performance charisma_rating money

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(2, 271)       =     45.36
       Model |  300798.824         2  150399.412   Prob > F        =    0.0000
    Residual |  898560.228       271   3315.7204   R-squared       =    0.2508
-------------+----------------------------------   Adj R-squared   =    0.2453
       Total |  1199359.05       273   4393.2566   Root MSE        =    57.582

------------------------------------------------------------------------------
    performance | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
----------------+-------------------------------------------------------------
charisma_rating |   8.542106   4.554463     1.88   0.062     -.424521    17.50873
          money |   64.87133   6.959277     9.32   0.000     51.17021    78.57245
          _cons |   137.6671   12.36463    11.13   0.000     113.3241      162.01
------------------------------------------------------------------------------
```

Whoops. The effect should be 54.57!

We redo with IV-regression!

```
reg3 (perf = charisma_rating money) ( charisma_rating = i.leader i.leader#i.money
i.money) , 2sls

Two-stage least-squares regression
----------------------------------------------------------------------------
Equation             Obs    Params         RMSE  "R-squared"         F      P>F
----------------------------------------------------------------------------
performance          274         2     101.9751     -1.3497      23.69   0.0000
charisma_r~g         274         3      .733478      0.0913       9.05   0.0000
----------------------------------------------------------------------------

------------------------------------------------------------------------------
                 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-----------------+------------------------------------------------------------
performance      |
charisma_rating |   118.1262   26.69781     4.42   0.000     65.68213    170.5703
           money |   63.42033    12.3291     5.14   0.000     39.20155    87.63912
           _cons |  -135.8235   67.18519    -2.02   0.044    -267.7993   -3.847744
-----------------+------------------------------------------------------------
charisma_rating |
       1.leader |   .4633408   .1240944     3.73   0.000     .2195749    .7071068
                |
   leader#money |
          1 1   |  -.0032695   .1773765    -0.02   0.985    -.3517005    .3451615
                |
         1.money |   .0083803   .1236177     0.07   0.946    -.2344492    .2512099
           _cons |   2.273973   .0858471    26.49   0.000     2.105338    2.442607
------------------------------------------------------------------------------
Note: Small-sample degrees-of-freedom adjustment applied when estimating
      covariance matrix of residuals.
Endogenous: performance charisma_rating
Exogenous:  money 0.leader 1.leader 0.leader#0.money 0.leader#1.money
            1.leader#0.money 1.leader#1.money 0.money 1.money
```

And the non-linear combination of estimators, the “indirect effect” gives the correct response!

```
. nlcom _b[ charisma_rating :1.leader]* _b[performance: charisma_rating ]

       _nl_1: _b[ charisma_rating :1.leader]* _b[performance: charisma_rating ]

------------------------------------------------------------------------------
             | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _nl_1 |   54.73269   19.18077     2.85   0.004     17.13907    92.32631
------------------------------------------------------------------------------
```

Let's examine Meslec et al. using the residualization procedure from Fischer et al., (2024). Remember the reduced form showed an effect

```
. reg charisma_rating i.leader i.money

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(2, 271)       =     13.62
       Model |  14.6012032         2   7.3006016   Prob > F        =    0.0000
    Residual |  145.257481       271  .536005464   R-squared       =    0.0913
-------------+----------------------------------   Adj R-squared   =    0.0846
       Total |  159.858684       273  .585562945   Root MSE        =    .73212

------------------------------------------------------------------------------
charisma_r~g | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.leader |   .4617405   .0885048     5.22   0.000     .2874961    .6359849
     1.money |   .0067923   .0884883     0.08   0.939    -.1674196    .1810042
       _cons |   2.274738   .0749842    30.34   0.000     2.127113    2.422364
------------------------------------------------------------------------------

. predict charisma_resid, resi
```

As we see below, in this case, the residuals show nothing! There is no idiosyncratic variance (though remember the OLS estimate on p. 21 gave a wrong estimate)—we just cannot trust estimators using only observed ratings:

```
. reg perf charisma_resid i.leader i.money

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(3, 270)       =     62.98
       Model |  493770.237         3  164590.079   Prob > F        =    0.0000
    Residual |  705588.814       270   2613.2919   R-squared       =    0.4117
-------------+----------------------------------   Adj R-squared   =    0.4052
       Total |  1199359.05       273   4393.2566   Root MSE        =     51.12

------------------------------------------------------------------------------
   performance | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
---------------+--------------------------------------------------------------
charisma_resid |  -2.469259   4.241551    -0.58   0.561    -10.81998     5.88146
      1.leader |   54.56704   6.179826     8.83   0.000     42.40027    66.73382
       1.money |   64.22236   6.178673    10.39   0.000     52.05785    76.38686
         _cons |   132.8715   5.235753    25.38   0.000     122.5634    143.1796
------------------------------------------------------------------------------
```

The behavioral manipulation drives y; there is nothing, in this case from the rating, that correlates with y.

If we do the residualization for Fischer et al., (2024). We have Case 4:

```
. reg donation charisma_resid i.manip_charisma i.manip_cue panas_pos -openness

      Source |       SS           df       MS      Number of obs   =       409
-------------+----------------------------------   F(10, 398)      =      2.42
       Model |  5.51090437         10  .551090437   Prob > F        =    0.0084
    Residual |  90.7482643        398  .228010714   R-squared       =    0.0573
-------------+----------------------------------   Adj R-squared   =    0.0336
       Total |  96.2591687        408  .235929335   Root MSE        =     .4775

------------------------------------------------------------------------------
       donation | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
----------------+-------------------------------------------------------------
  charisma_resid |   .1166213   .0283992     4.11   0.000     .0607901    .1724524
1.manip_charisma |   .0496189    .047604     1.04   0.298    -.0439678    .1432057
     1.manip_cue |   .0046598   .0476132     0.10   0.922     -.088945    .0982647
       panas_pos |   .0288227   .0358291     0.80   0.422    -.0416152    .0992606
       panas_neg |  -.0162122   .0490468    -0.33   0.741    -.1126354     .080211
          extrav |  -.0103089   .0230892    -0.45   0.655    -.0557008    .0350831
         agreeab |   .0301093   .0264122     1.14   0.255    -.0218155    .0820342
       conscient |   .0429167   .0351839     1.22   0.223    -.0262528    .1120863
          neurot |   .0323433   .0268155     1.21   0.228    -.0203744     .085061
        openness |  -.0261269   .0256621    -1.02   0.309    -.0765771    .0243233
           _cons |   .0503764   .2285957     0.22   0.826    -.3990296    .4997824
------------------------------------------------------------------------------
```

What is the moral of the story?

- Do not use rated leadership measures, or rated measures of any construct to predict anything, unless you control the information environment and use IV regression.

- If you cannot manipulate then measure the behavior objectively (Emrich, Brower, Feldman, & Garland, 2001; Jacquart & Antonakis, 2015; Jensen et al., 2023; Tur, Harstad, & Antonakis, 2022).

References:

Antonakis, J., d'Adda, G., Weber, R. A., & Zehnder, C. 2022. "Just Words? Just Speeches?" On the Economic Value of Charismatic Leadership. ***Management Science***, 68(9): 6355-6381.

Banks, G. C., Engemann, K. N., Williams, C. E., Gooty, J., McCauley, K. D., & Medaugh, M. R. 2017. A meta-analytic review and future research agenda of charismatic leadership. ***The Leadership Quarterly***, 28(4): 508-529.

Banks, G. C., Woznyj, H. M., & Mansfield, C. A. 2021. Where is "behavior" in organizational behavior? A call for a revolution in leadership research and beyond. ***The Leadership Quarterly***, 34: 101581.

Bastardoz, N., Matthews, M. J., Sajons, G. B., Ransom, T., Kelemen, T. K., & Matthews, S. H. 2023. Instrumental variables estimation: Assumptions, pitfalls, and guidelines. ***The Leadership Quarterly***, 34(1): 101673.

Emrich, C. G., Brower, H. H., Feldman, J. M., & Garland, H. 2001. Images in words: Presidential rhetoric, charisma, and greatness. ***Administrative Science Quarterly***, 46: 527–557.

Fischer, T. 2023. Measuring behaviors counterfactually. ***The Leadership Quarterly***: 101750.

Fischer, T., Dietz, J., & Antonakis, J. 2024. A fatal flaw: Positive leadership style research creates causal illusions. ***The Leadership Quarterly***: 101771.

Jacquart, P., & Antonakis, J. 2015. When does charisma matter for top-level leaders? Effect of attributional ambiguity. ***Academy of Management Journal***, 58: 1051-1074.

Jensen, U., Rohner, D., Bornet, O., Carron, D., Garner, P. N., Loupi, D., & Antonakis, J. 2023. Combating COVID-19 with charisma: Evidence on Governor Speeches in the United States. ***The Leadership Quarterly***: 101702.

Lord, R. G., Binning, J. F., Rush, M. C., & Thomas, J. C. 1978. The effect of performance cues and leader behavior on questionnaire ratings of leadership behavior. ***Organizational Behavior and Human Performance***, 21(1): 27-39.

Meslec, N., Curseu, P. L., Fodor, O. C., & Kenda, R. 2020. Effects of charismatic leadership and rewards on individual performance. ***The Leadership Quarterly***, 31(6): 101423.

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. 1990. Transformational leader behaviors and their effects on follower's trust in leader, satisfaction, and organizational citizenship behaviors. ***The Leadership Quarterly***, 1(2): 107-142.

Tur, B., Harstad, J., & Antonakis, J. 2022. Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. ***The Leadership Quarterly***, 33(5): 101476.