Discussion - Gema Zamarro

Christian Fons-Rosen

Universitat Pompeu Fabra

March 4, 2017

"Comparing and Validating Measures of Character Skills: Performance Task Measures and Self-Reports from a Nationally Representative Internet Panel" (Zamarro, Cheng, Shakeel, and Hitt)

- *Problem:* Difficulty in reliably measuring non-cognitive character skills
- Previous approaches:
  - **Self-reported psychometric scales** biased: social desirability bias, reference group bias, etc.
  - Measures based on **performance tasks** that capture some *underlying* character skill are costly and difficult to obtain in large samples

"Comparing and Validating Measures of Character Skills: Performance Task Measures and Self-Reports from a Nationally Representative Internet Panel" (Zamarro, Cheng, Shakeel, and Hitt)

- *Problem:* Difficulty in reliably measuring non-cognitive character skills
- Previous approaches:
  - **Self-reported psychometric scales** biased: social desirability bias, reference group bias, etc.
  - Measures based on **performance tasks** that capture some *underlying* character skill are costly and difficult to obtain in large samples

- Internet panel from Understanding America Study (4k households)
- *Hypothesis:* Survey-effort measures of character skills reveal something about character skills, especially conscienciousness

"Comparing and Validating Measures of Character Skills: Performance Task Measures and Self-Reports from a Nationally Representative Internet Panel" (Zamarro, Cheng, Shakeel, and Hitt)

- *Problem:* Difficulty in reliably measuring non-cognitive character skills
- Previous approaches:
  - **Self-reported psychometric scales** biased: social desirability bias, reference group bias, etc.
  - Measures based on **performance tasks** that capture some *underlying* character skill are costly and difficult to obtain in large samples

- Internet panel from Understanding America Study (4k households)
- *Hypothesis:* Survey-effort measures of character skills reveal something about character skills, especially conscienciousness

- *Findings:*
  - Similar to grit, careless answering correlates mostly with conscienciousness and neuroticism
  - Not much on non-response rates

General Comment

- Survey-effort measures can not only overcome some previous problems, but also proxy for character skills

General Comment

- Survey-effort measures can not only overcome some previous problems, but also proxy for character skills
- But given the plethora of possible measures, which ones should I choose? Could theory guide us in choosing?
  - Medical research: Kitchen sink approach vs theoretical approach?

General Comment

- Survey-effort measures can not only overcome some previous problems, but also proxy for character skills
- But given the plethora of possible measures, which ones should I choose? Could theory guide us in choosing?
  - Medical research: Kitchen sink approach vs theoretical approach?
- *"Other respondents lacking conscienciousness may ... hastily rush through a survey"* $\longrightarrow$ answering time between 2 questions

General Comment

- Survey-effort measures can not only overcome some previous problems, but also proxy for character skills
- But given the plethora of possible measures, which ones should I choose? Could theory guide us in choosing?
  - Medical research: Kitchen sink approach vs theoretical approach?
- *"Other respondents lacking conscienciousness may ... hastily rush through a survey"* $\longrightarrow$ answering time between 2 questions
  - *Mean:*
  - Longer time to answer...more thoughtful VS more doubtful?

General Comment

- Survey-effort measures can not only overcome some previous problems, but also proxy for character skills
- But given the plethora of possible measures, which ones should I choose? Could theory guide us in choosing?
  - Medical research: Kitchen sink approach vs theoretical approach?
- *"Other respondents lacking conscienciousness may ... hastily rush through a survey"* $\longrightarrow$ answering time between 2 questions
  - *Mean:*
  - Longer time to answer...more thoughtful VS more doubtful?
  - *Variance:*
  - Constant answering time...organized VS mechanical?

General Comment

- Survey-effort measures can not only overcome some previous problems, but also proxy for character skills
- But given the plethora of possible measures, which ones should I choose? Could theory guide us in choosing?
  - Medical research: Kitchen sink approach vs theoretical approach?
- *"Other respondents lacking conscienciousness may ... hastily rush through a survey"* $\longrightarrow$ answering time between 2 questions
  - *Mean:*
  - Longer time to answer...more thoughtful VS more doubtful?
  - *Variance:*
  - Constant answering time...organized VS mechanical?
  - *Heterogeneity:*
  - Do previous conclusions differ across types of individuals? Which characteristics are relevant in splitting individuals into groups?

Comment - Item Non-Response

"respondents were asked an average of 115 questions in each of these five waves. We then take the average item non-response rate across waves and within each respondent, hypothesizing that higher non-response rates indicate lower levels of conscienciousness"

Comment - Item Non-Response

"respondents were asked an average of 115 questions in each of these five waves. We then take the average item non-response rate across waves and within each respondent, hypothesizing that higher non-response rates indicate lower levels of conscienciousness"

- Not answering questions 6-10 probably means a different thing than not answering 100-105. Weighting schemes?

Comment - Item Non-Response

"respondents were asked an average of 115 questions in each of these five waves. We then take the average item non-response rate across waves and within each respondent, hypothesizing that higher non-response rates indicate lower levels of conscienciousness"

- Not answering questions 6-10 probably means a different thing than not answering 100-105. Weighting schemes?
- Could you tell them the expected time needed to answer the questions and see how it changes both *level* and *composition* of non-response items?

Comment - Item Non-Response

"respondents were asked an average of 115 questions in each of these five waves. We then take the average item non-response rate across waves and within each respondent, hypothesizing that higher non-response rates indicate lower levels of conscienciousness"

- Not answering questions 6-10 probably means a different thing than not answering 100-105. Weighting schemes?
- Could you tell them the expected time needed to answer the questions and see how it changes both *level* and *composition* of non-response items?
- Could you randomize the sequence of questions?
  - If pattern of non-response items change, it can help to understand whether main reason for non-response is the respondent or rather the question!

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead
- Are you careless or do you care too much?

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead

- Are you careless or do you care too much?
  - The Smiths (1985): *"That joke isn't funny anymore, it's too close to home and it's too near the bone*

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead

- Are you careless or do you care too much?
  - The Smiths (1985): *"That joke isn't funny anymore, it's too close to home and it's too near the bone ...more than you'll ever know"*

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead

- Are you careless or do you care too much?
  - The Smiths (1985): *"That joke isn't funny anymore, it's too close to home and it's too near the bone ...more than you'll ever know"*
  - Do we know how much courage it takes to answer one of these questions...let alone a battery of them!

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead

- Are you careless or do you care too much?
  - The Smiths (1985): *"That joke isn't funny anymore, it's too close to home and it's too near the bone ...more than you'll ever know"*
  - Do we know how much courage it takes to answer one of these questions...let alone a battery of them!
  - What does it mean to leave blank *"I with I was dead"* vs *"Setbacks don't discourage me"*?

Comment - Careless Answering

"following three scales to build our careless answering measure: a life satisfaction scale, a well-being scale and a depression scale."

- Depression Scale:
  - I feel sad or depressed
  - I feel guilty
  - The future looks hopeless
  - I thought about killing myself
  - I wish I was dead
- Are you careless or do you care too much?
  - The Smiths (1985): *"That joke isn't funny anymore, it's too close to home and it's too near the bone ...more than you'll ever know"*
  - Do we know how much courage it takes to answer one of these questions...let alone a battery of them!
  - What does it mean to leave blank *"I with I was dead"* vs *"Setbacks don't discourage me"*?
- Could you videotape faces to identify mood swings?

Comment - Careless Answering

- Should we think of careless responding as a person-specific or rather a scale-specific attribute?
  - If you run a regression with individual fixed effects using multiple scales per individual, how much variation is explained by fixed effects?
  - Is there a discontinuity in the degree of carelessness when you start questions on the new scale?

Comment - Careless Answering

- Should we think of careless responding as a person-specific or rather a scale-specific attribute?
  - If you run a regression with individual fixed effects using multiple scales per individual, how much variation is explained by fixed effects?
  - Is there a discontinuity in the degree of carelessness when you start questions on the new scale?

"advantage of careless answering...not affected by social desirability bias, experimenter demand effects...as long as participants are unaware that their effort on surveys is being closely observed"

$\longrightarrow$ Any measure that is robust to the Lucas Critique?

Comment - Careless Answering

If the scale is reliable, each question will consistently measure the same underlying construct. Individual responses to each question then would be well predicted by the responses to other questions in this same reliable scale.

Comment - Careless Answering

If the scale is reliable, each question will consistently measure the same underlying construct. Individual responses to each question then would be well predicted by the responses to other questions in this same reliable scale.

- On the one hand, the point of having multiple questions is that each of them provides a new angle
  - Variation is good

Comment - Careless Answering

If the scale is reliable, each question will consistently measure the same underlying construct. Individual responses to each question then would be well predicted by the responses to other questions in this same reliable scale.

- On the one hand, the point of having multiple questions is that each of them provides a new angle
  - Variation is good
- But on the other hand, they are penalized when providing differing answers across questions
  - Too much variation is bad

Comment - Careless Answering

If the scale is reliable, each question will consistently measure the same underlying construct. Individual responses to each question then would be well predicted by the responses to other questions in this same reliable scale.

- On the one hand, the point of having multiple questions is that each of them provides a new angle
  - Variation is good
- But on the other hand, they are penalized when providing differing answers across questions
  - Too much variation is bad
- How much is too much? Guidelines based on theory, data analysis, or intuition?

"Further Validation of Survey-Effort Measures of Conscientiousness: Results from a Sample of High School Students" (Zamarro, Nichols, Duckworth, and D'Mello)

- *Problem:* Difficulty in reliably measuring non-cognitive character skills

"Further Validation of Survey-Effort Measures of Conscientiousness: Results from a Sample of High School Students" (Zamarro, Nichols, Duckworth, and D'Mello)

- *Problem:* Difficulty in reliably measuring non-cognitive character skills

- *Contribution:*
  - Validity of survey-effort measures by studying the correlation with teacher external reports and with student performance
  - Relationship between survey-effort measures and other performance task measures designed to capture diligence and tolerance of frustration

"Further Validation of Survey-Effort Measures of Conscientiousness: Results from a Sample of High School Students" (Zamarro, Nichols, Duckworth, and D'Mello)

- *Problem:* Difficulty in reliably measuring non-cognitive character skills

- *Contribution:*
  - Validity of survey-effort measures by studying the correlation with teacher external reports and with student performance
  - Relationship between survey-effort measures and other performance task measures designed to capture diligence and tolerance of frustration

- *Findings* using longitudinal data on high school seniors (2014):
  - survey-effort measures (item non-response and careless answering) promising proxies for character skills related to grit and self-control

Comment

"As several teachers reported on a single child, each teacher's score was averaged for each student to give that student a unique score."

- Explore the *variance* of teacher reports as a proxy for how confident an averaged score is

Comment

"As several teachers reported on a single child, each teacher's score was averaged for each student to give that student a unique score."

- Explore the *variance* of teacher reports as a proxy for how confident an averaged score is

How are "left blank" questions accounted for in the "careless answering scale?

- Based on Table 3, column (1), it might even strenghen your message if we consider "left blank" as a proxy for "careless answering" instead of ignoring them

Table 3. Spearman Correlations between Survey Effort Measures, Self-reports and Teacher Reports

	Item Non-response	Careless Answering
Self-Reported Measures		
Grit	**-0.160**	-0.110
Locus of Control	-0.104	-0.014
Self Control Combined	**-0.162**	**-0.161**
Self Contol Work	-0.122	**-0.161**
Self Control Interpersonal	-0.154	-0.113
Teachers Reported Measures		
Teacher Reported Grit	**-0.162**	**-0.183**
Teacher Reported Work Self Control	-0.143	**-0.182**
Teacher Reported Interpersonal self control	-0.061	-0.095
Teacher Reported Redirection	0.055	**0.145**
Teacher Reported HW Completion	-0.086	-0.080