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 conscientiousness
"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 conscientiousness.

- **Findings:**
  - Similar to grit, **careless answering** correlates mostly with conscientiousness and neuroticism.
  - Not much on **non-response rates**.
Survey-effort measures can not only overcome some previous problems, but also proxy for character skills.
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 conscientiousness may...hastily rush through a survey

  - 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?
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 conscientiousness may ... hastily rush through a survey" → answering time between 2 questions
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 conscientiousness may ... hastily rush through a survey" → answering time between 2 questions

Mean:

Longer time to answer...more thoughtful VS more doubtful?
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 conscientiousness may ... hastily rush through a survey" → answering time between 2 questions

- **Mean:**
  - Longer time to answer...more thoughtful VS more doubtful?

- **Variance:**
  - Constant answering time...organized VS mechanical?
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 conscientiousness may ... hastily rush through a survey" → 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?
"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 conscientiousness"
"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 conscientiousness"

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

- 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?
"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 conscientiousness"

- 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!
"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
"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 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 wish I was dead" vs "Setbacks don’t discourage me"?

Could you videotape faces to identify mood swings?
"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
"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

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 wish I was dead" vs "Setbacks don't discourage me"?

Could you videotape faces to identify mood swings?
“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
"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?
"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"
“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 wish I was dead" vs "Setbacks don’t discourage me"?
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?
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"

→ Any measure that is robust to the **Lucas Critique**?
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.
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
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
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
"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.
"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 strengthen 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

<table>
<thead>
<tr>
<th>Item</th>
<th>Non-response</th>
<th>Careless Answering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-Reported Measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grit</td>
<td>-0.160</td>
<td>-0.110</td>
</tr>
<tr>
<td>Locus of Control</td>
<td>-0.104</td>
<td>-0.014</td>
</tr>
<tr>
<td>Self Control Combined</td>
<td>-0.162</td>
<td>-0.161</td>
</tr>
<tr>
<td>Self Control Work</td>
<td>-0.122</td>
<td>-0.161</td>
</tr>
<tr>
<td>Self Control Interpersonal</td>
<td>-0.154</td>
<td>-0.113</td>
</tr>
<tr>
<td><strong>Teachers Reported Measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher Reported Grit</td>
<td>-0.162</td>
<td>-0.183</td>
</tr>
<tr>
<td>Teacher Reported Work Self Control</td>
<td>-0.143</td>
<td>-0.182</td>
</tr>
<tr>
<td>Teacher Reported Interpersonal self control</td>
<td>-0.061</td>
<td>-0.095</td>
</tr>
<tr>
<td>Teacher Reported Redirection</td>
<td>0.055</td>
<td>0.145</td>
</tr>
<tr>
<td>Teacher Reported HW Completion</td>
<td>-0.086</td>
<td>-0.080</td>
</tr>
</tbody>
</table>