The Role of Incentives in Measuring Cognitive and Non-cognitive Skills: Experimental Evidence from Primary Schools in Shanghai

Yuanyuan Chen, Shuaizhang Feng, James J. Heckman, and Tim Kautz

HCEO Measuring and Assessing Skills
March 3, 2017
This draft, March 2, 2017

Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

4 Main Results

5 Subgroup Analyses

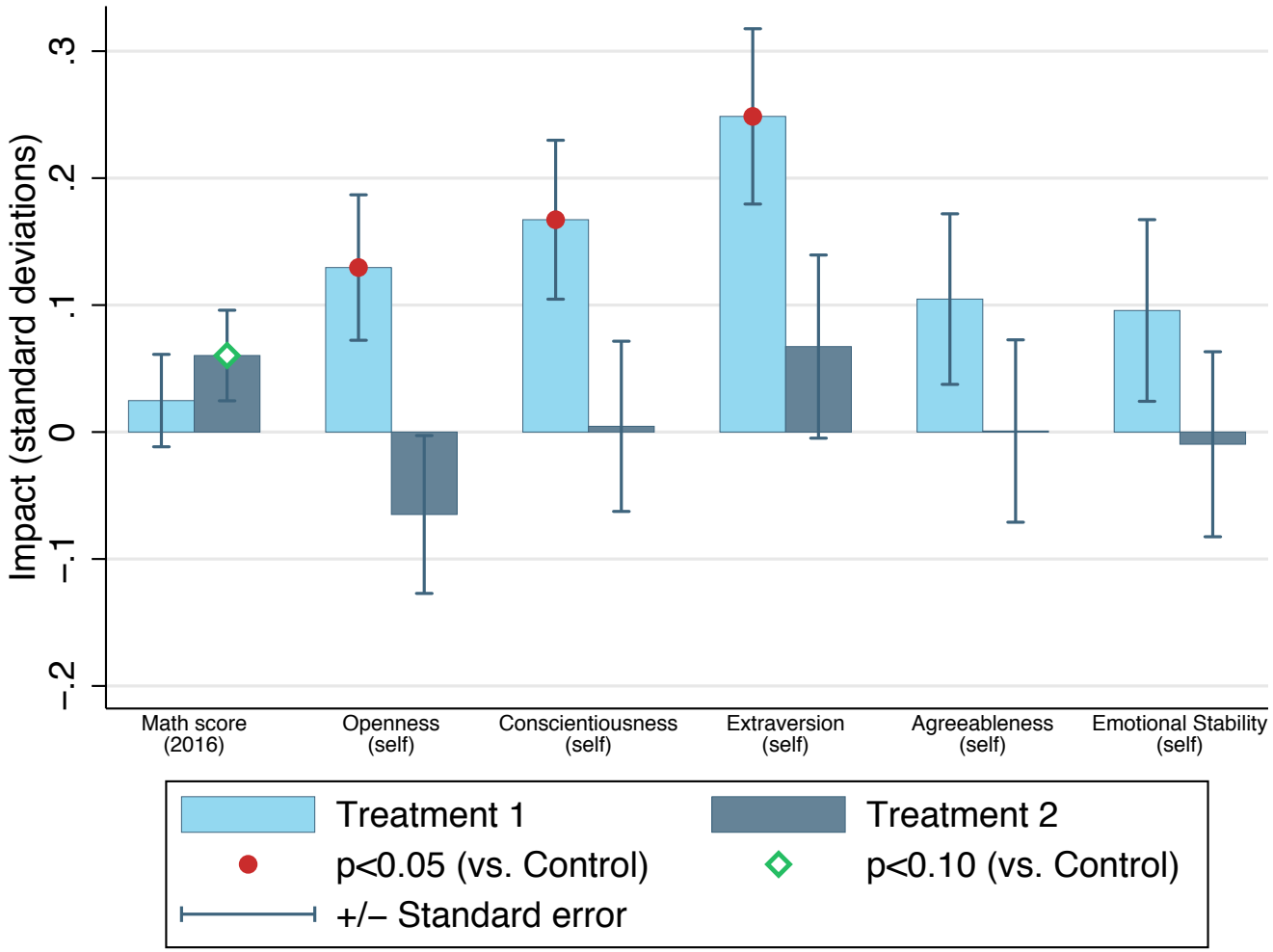
6 Conclusion

7 Appendix

There is a growing interest in non-cognitive skills but measurement remains an issue

1. A growing body of evidence shows the importance of non-cognitive skills in predicting life outcomes (Almlund et al., 2011)
2. Interventions have been shown to improve life outcomes through non-cognitive skills (Kautz et al., 2014)
3. Policy-makers are interested in expanding programs to develop non-cognitive skills but require reliable measures
4. One commonly used taxonomy is the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), often collected through self-reports

Impact of an "intervention" on math performance and self-reported Big Five



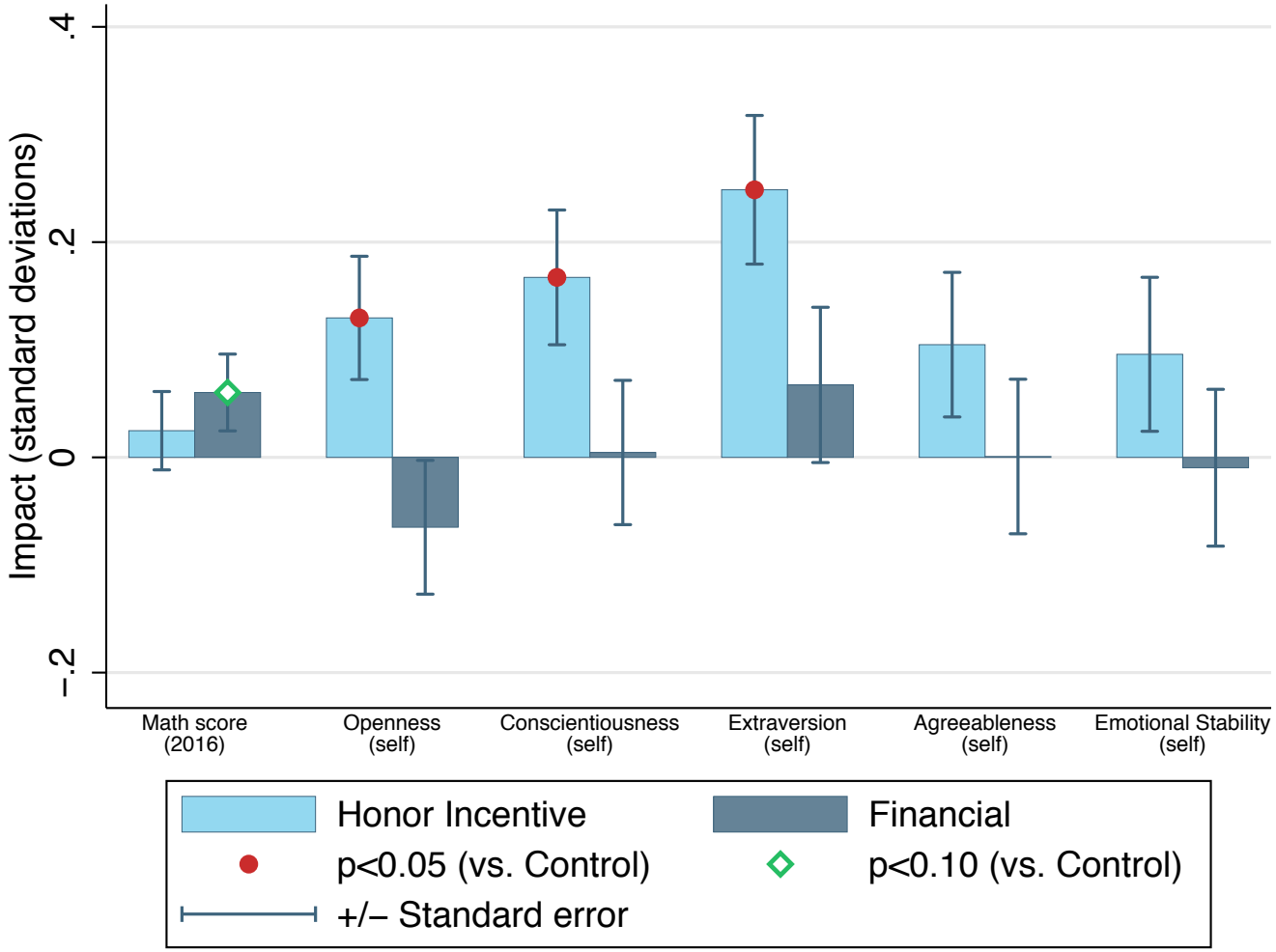
More about the "intervention"

- Both treatments cost less than $1 per student
- Took less than five minutes to deliver to a classroom
- The projected rate of return is huge

What is this intervention?

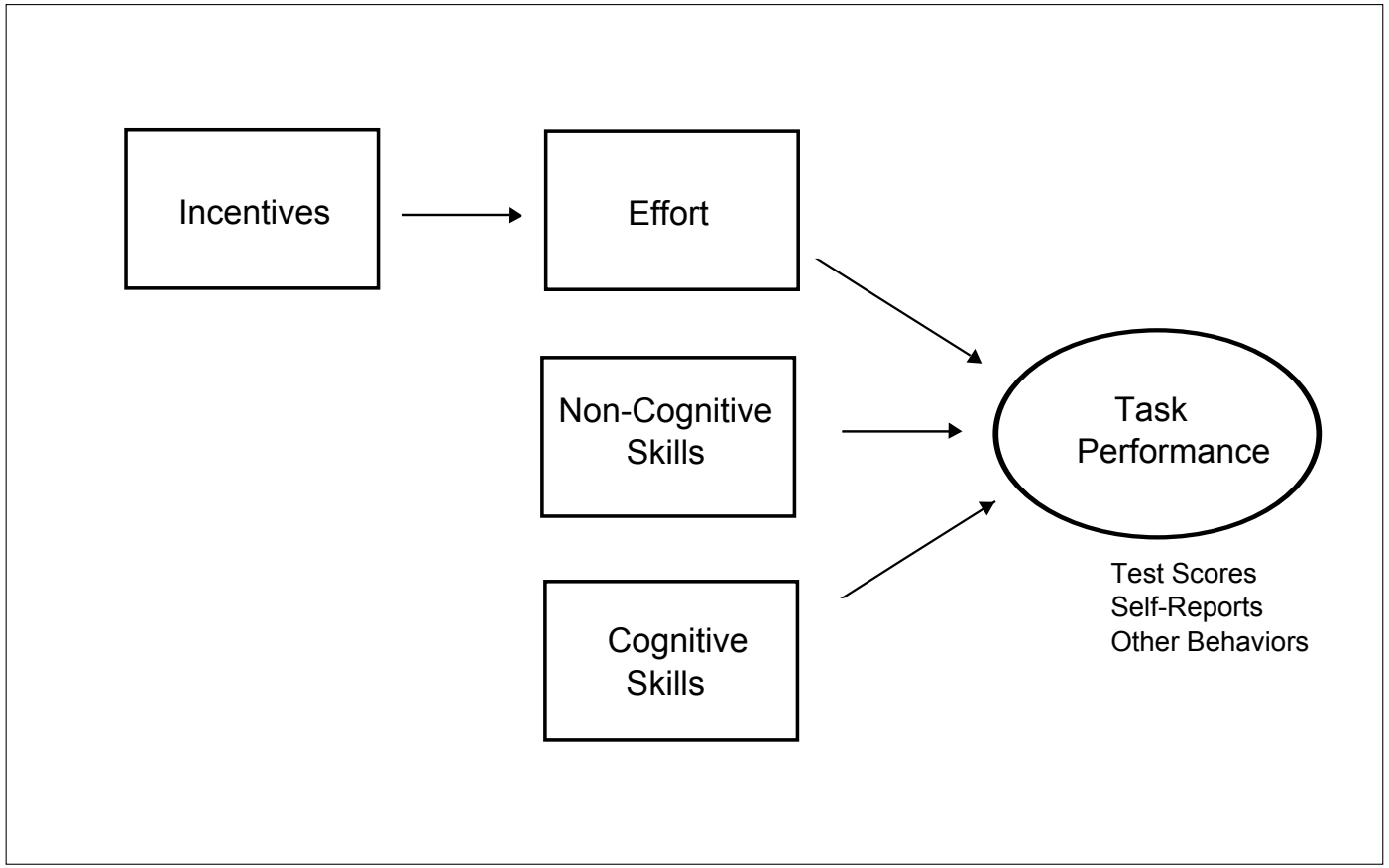
- Right before a math test and Big Five survey, fourth-graders received different instructions
- **Treatment 1 ("Honor incentive")**: Receive a certificate of honor if the math test score in the top 10% of the school in terms of overall performance or improvement
- **Treatment 2 ("Financial incentive")**: Receive 50 Yuan (≈\$7.5) if the math test score in the top 10% of the school in terms of overall performance or improvement
- Self-reported Big Five was administered directly after the math test

Impact of an "intervention" on math performance and self-reported Big Five



All measures are based on a behavior

Figure 1: Determinants of Task Performance



Source: Kautz et al. (2014).

Past studies establish the importance of incentives in cognitive testing

- A series of studies show that IQ scores can be improved by giving children candy or other incentives (Almlund et al., 2011)
- Borghans et al. (2008) find that incentives affect the time spent on IQ tests and people with higher levels of Emotional Stability and Conscientiousness are less affected by incentives
- Segal (2012) shows that coding speed scores can be influenced by incentives and that people with higher levels of Conscientiousness are more intrinsically motivated
- We found no experimental studies that focus on the effect of the situation on measures of non-cognitive skills

Main research questions

1. To what extent do cognitive skill measures depend on incentives or other aspects of the situation in a school setting?
2. Do different types of students respond differently?
3. Do different incentives (monetary vs. non-monetary) work differently?
4. Could self-reports of non-cognitive skills be inadvertently affected by incentives?

Key findings

- The incentives had little effect on overall test scores but did improve scores for better students
- The honor treatment had a large and statistically significant effect on self-reported Big Five measures
- Students in the "honor" treatment also rated their peers better in terms of Big Five (particularly females)

Why did the honor incentive but not financial incentive affect reporting of the Big Five?

- Somehow shifted the students' frame of mind
- Elicited more social-desirability bias by causing them to think of public recognition
- Other ideas?

These results suggest caution in interpreting evaluations based on self-reports

- Self-reported non-cognitive skill measures are used in policy-evaluation and school accountability
- A meta-analysis of interventions with mostly short-term (less than 6 month) follow-ups found effect sizes of 0.22-0.27 across five domains and 0.57 for another (Durlak et al., 2011)
- The honor incentive had impacts of approximately 0.10-0.20 standard deviations
- Interventions could plausibly have a similar psychological effect as the honor incentive

Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

4 Main Results

5 Subgroup Analyses

6 Conclusion

7 Appendix

Designed to maximize power and minimize contamination

1. Within schools, ranked students by fall math test scores and randomized triplets of students with the same scores into the control group, the honor treatment, or the financial treatment
2. On test day, separated students into classrooms based on treatment status
3. Students completed the math test and self-report of Big Five
4. Students returned to original classroom and assessed the Big Five of a peer
5. Two weeks later, teachers assessed their own students' Big Five

Survey description

- Approximately 1,900 fourth-grade students from 19 different schools in Shanghai
- 1st wave of survey administered in Fall 2015
- 2nd wave of survey administered in Spring 2016
- 3rd wave of survey will be administered in Spring 2017

The data include a rich set of demographics, cognitive skills, and non-cognitive skills

Demographic variables: gender, parental education, family income, rural hukou, Shanghai hukuo, private school, age

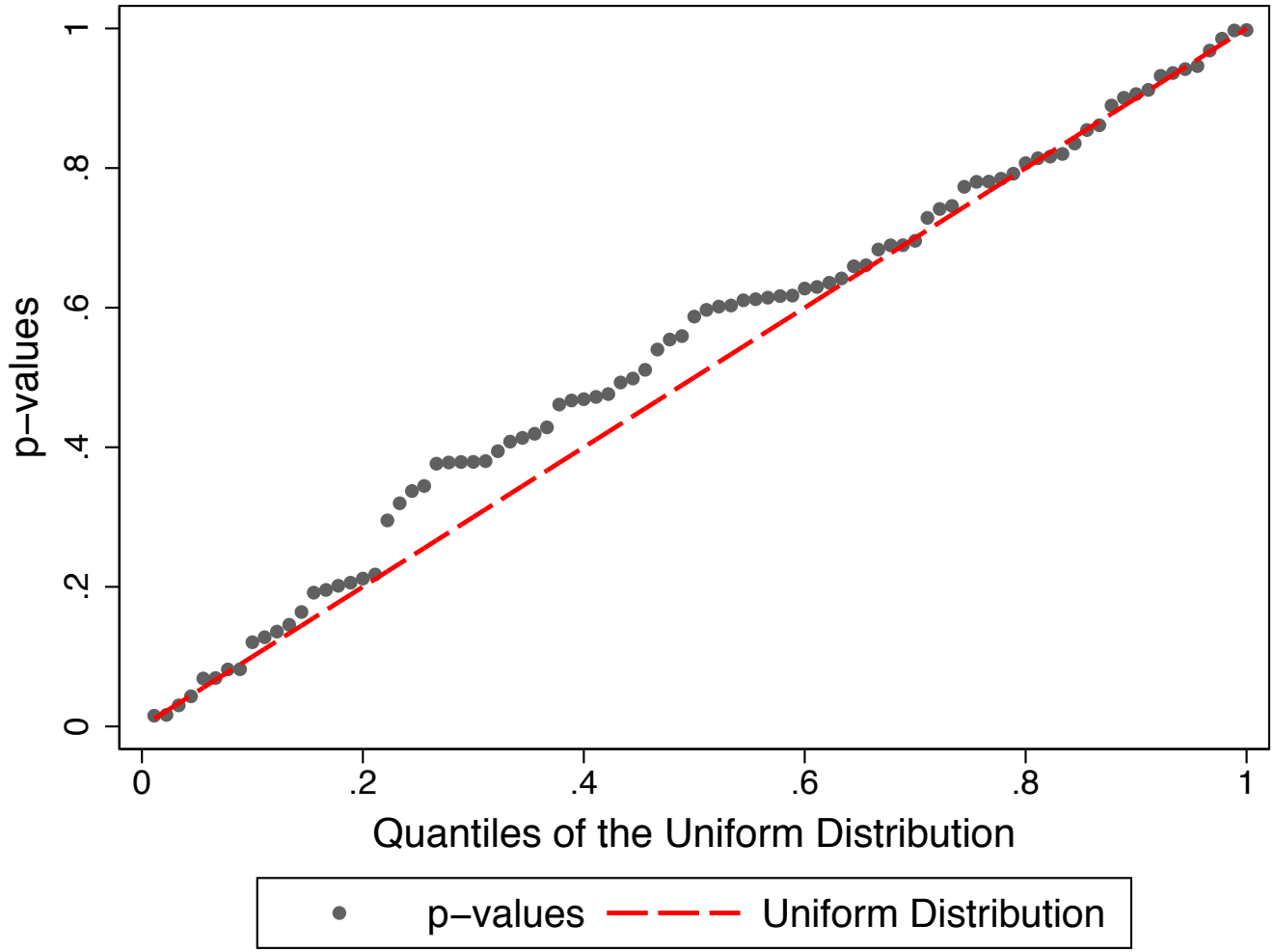
Cognitive skills: IQ, pre-intervention math test, math grades, Chinese grades, English grades

Non-cognitive skills : Big Five (self, peer, teacher reports), group-leader status, 1-3 rating of daily performance (teacher report), 1-3 rating of punctuality (teacher report), 1-3 rating of discipline (teacher report)

The experiment achieved baseline equivalence between the treatment and control groups

- Assess baseline equivalence using 30 different pre-program variables
- Of the 90 pairwise tests between groups, only 4 of them are statistically significant at the % level

The distribution of *p*-values follow a distribution consistent with baseline equivalence



Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

4 Main Results

5 Subgroup Analyses

6 Conclusion

7 Appendix

Estimation model:

$$Y_{is} = \alpha + \beta^{honor} T_{is}^{honor} + \beta^{financial} T_{is}^{financial} + \gamma X_{is} + \varepsilon_{is}.$$

Y_{is}: outcome for student i in school s
T_{is}^{honor}: indicator for honor treatment
$T_{is}^{financial}$: indicator for financial treatment
X_{is}: covariates (including school fixed effects)
ε_{is}: error term, allowing for heteroskedasticity

Main specification

- To increase precision, control for background and ability measures (and school fixed effects)
- **Demographic variables**: gender, parental education, family income, rural hukou, Shanghai hukuo, private school, age
- **Cognitive skills**: IQ, pre-intervention math test, math grades, Chinese grades, English grades
- **Non-cognitive skills** : Big Five (teacher reports), group-leader status, 1-3 rating of daily performance (teacher report), 1-3 rating of punctuality (teacher report), 1-3 rating of discipline (teacher report)
- Results are similar with no controls or various combinations of controls

Outcome measures

- To reduce measurement error, we apply a factor model to each grouping of items in the Big Five traits separately and predict factor scores (similar results if using means of items)
- All outcomes are standardized so that they are mean zero and have a standard deviation of one

Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

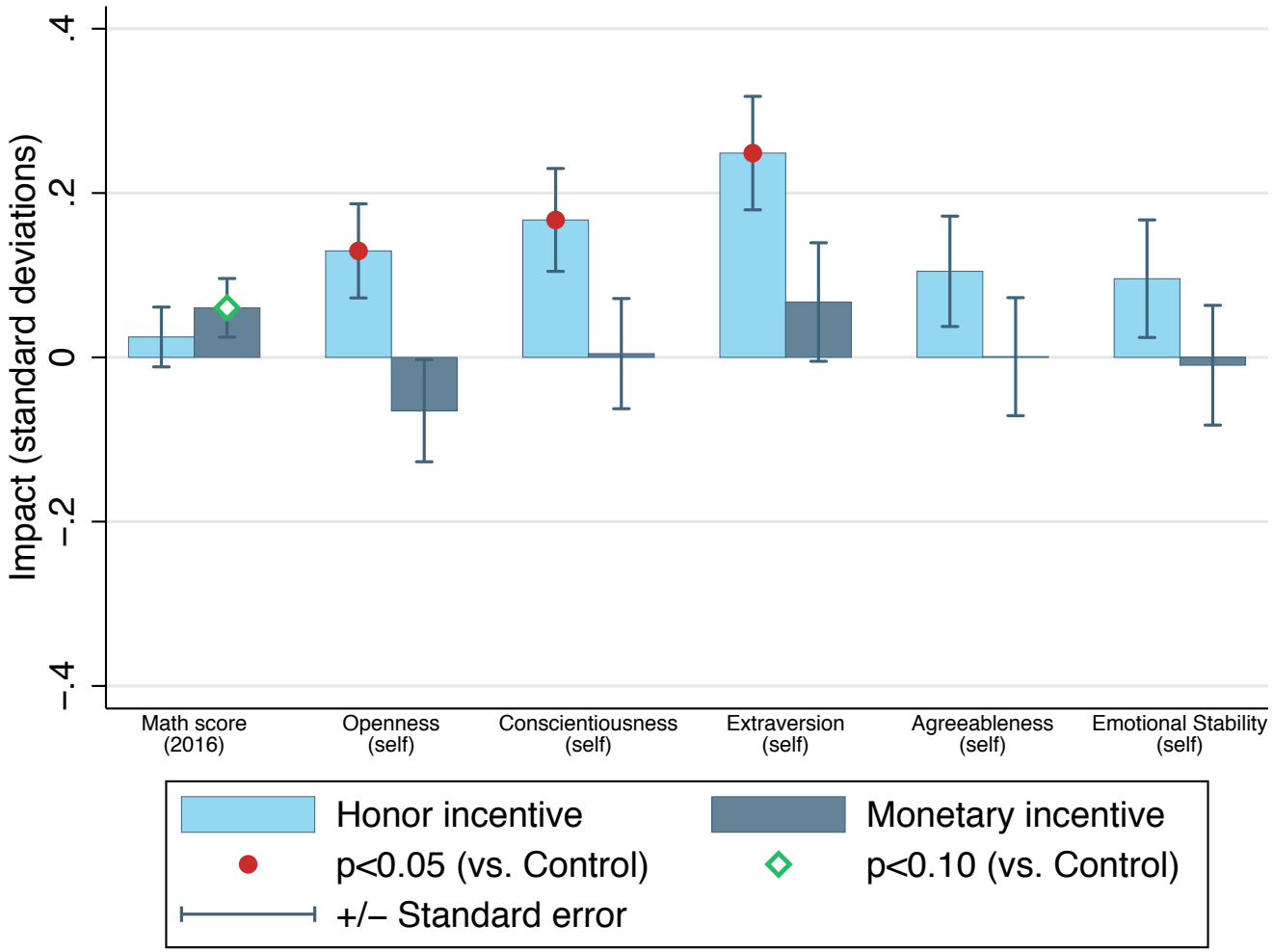
4 Main Results

5 Subgroup Analyses

6 Conclusion

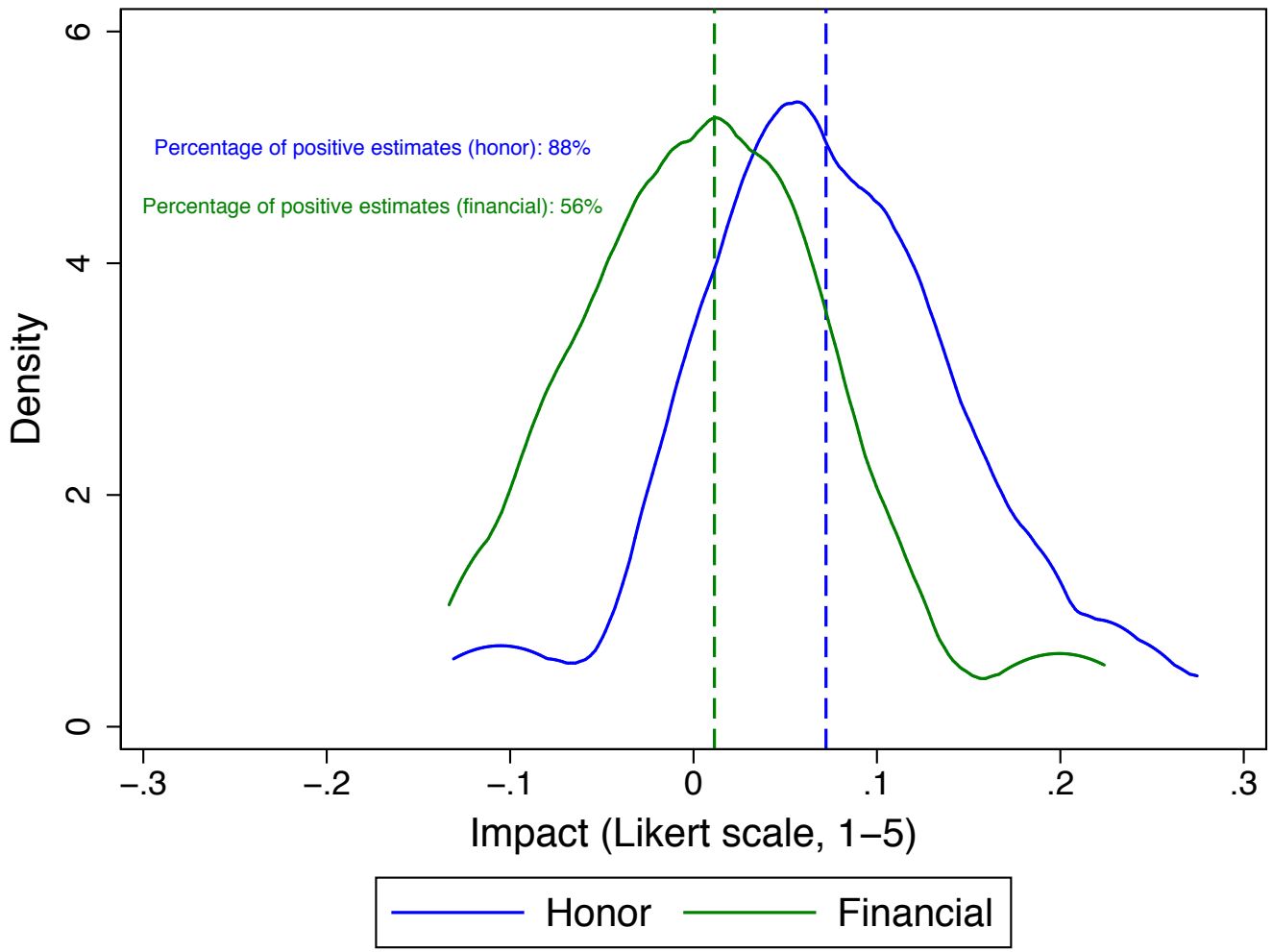
7 Appendix

Impact on math performance and self-reported Big Five (Full Sample)

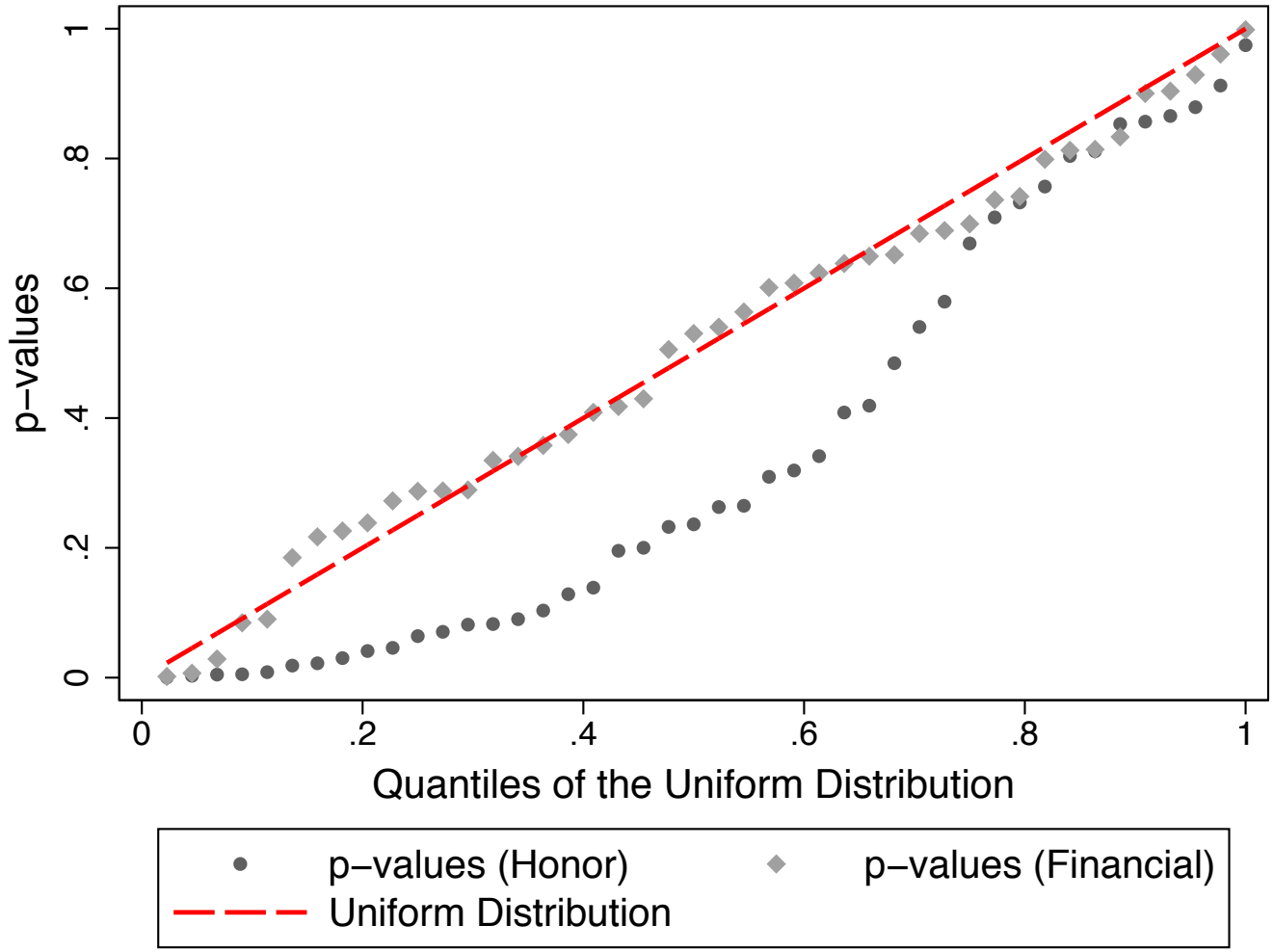


- No one item within the Big Five drove the results

Distribution of impact on individual Big Five items

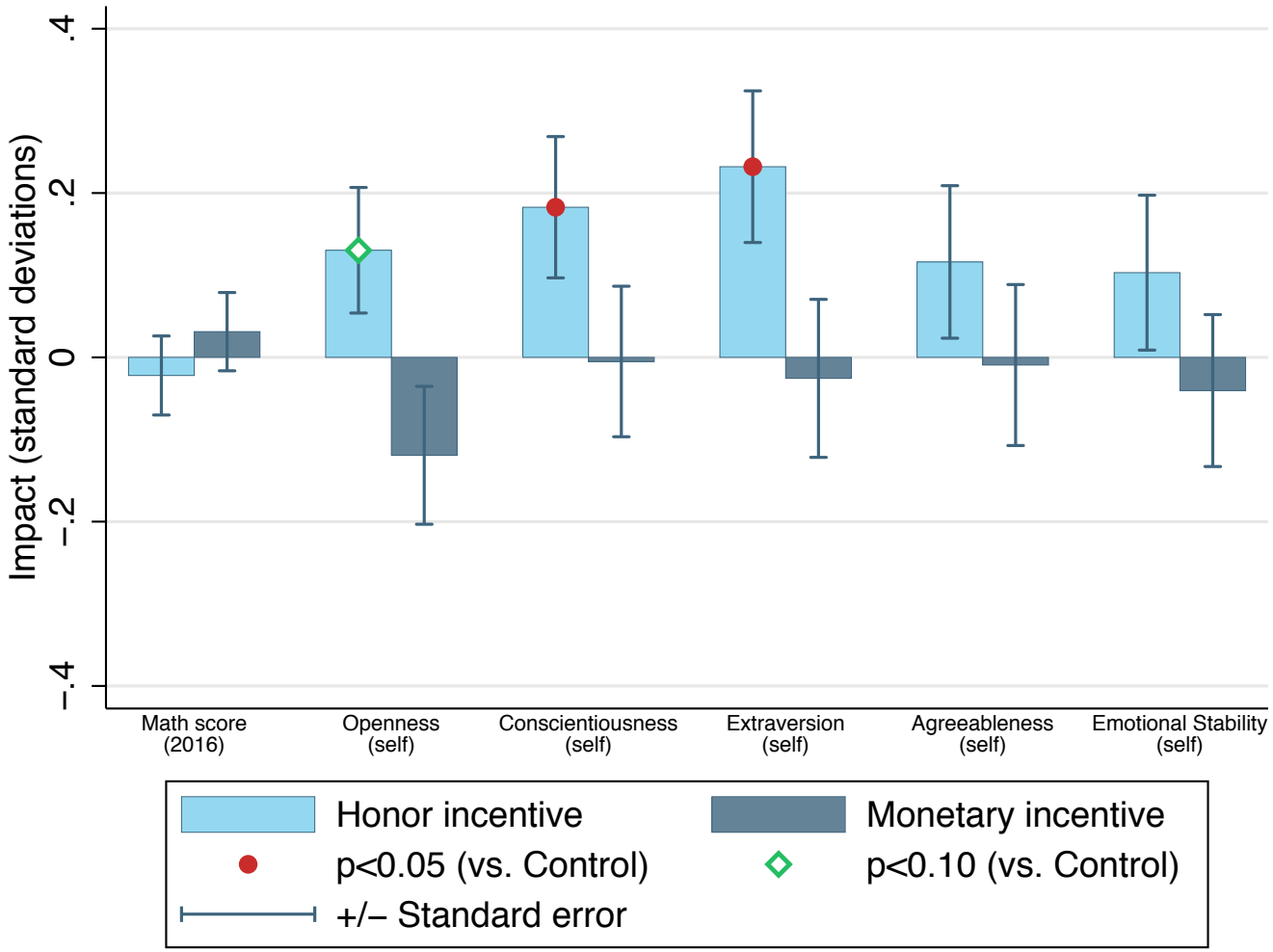


Distribution of *p*-values associated with impacts on individual Big Five items

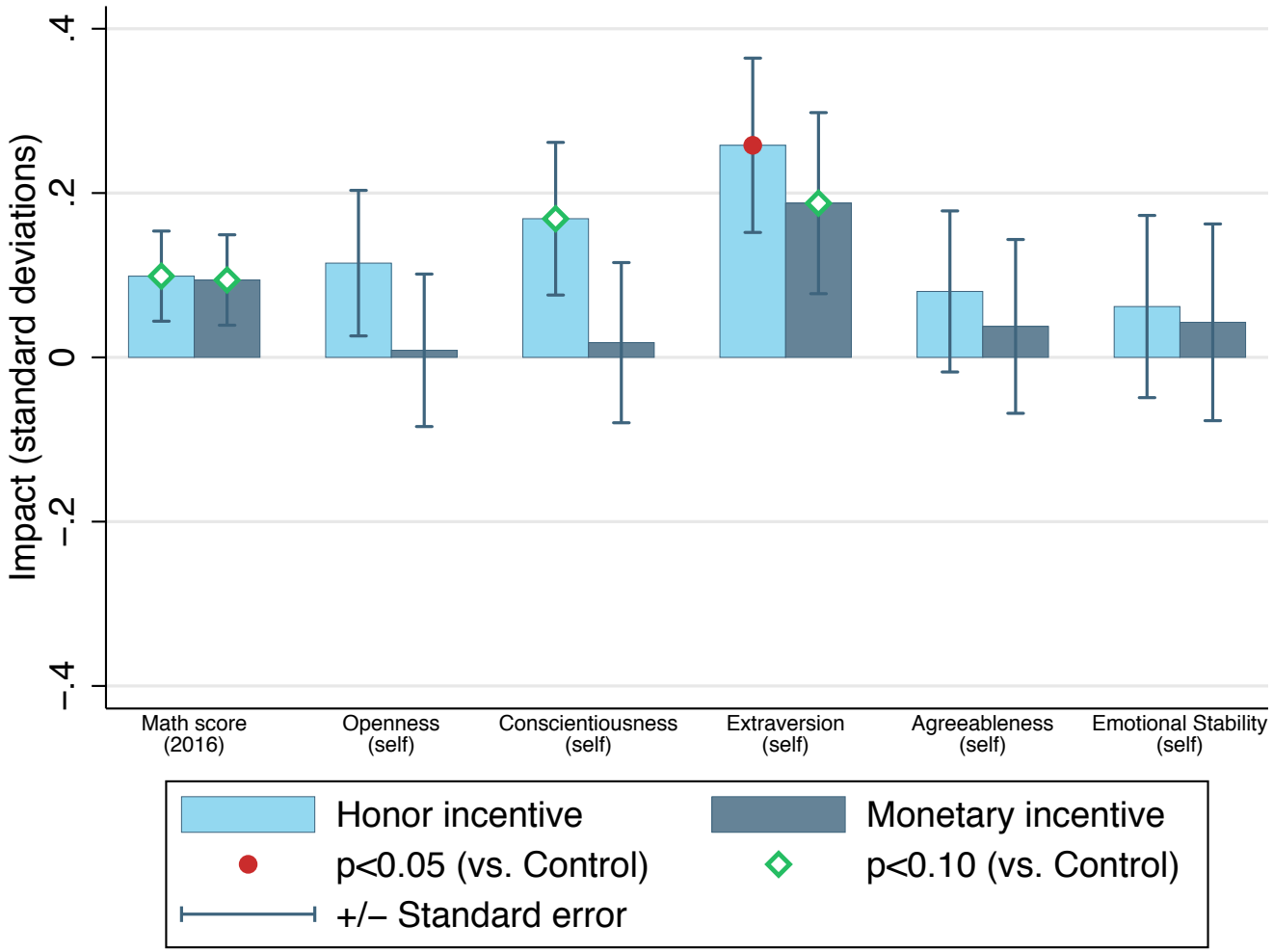


- There are some gender differences in the impacts on test scores

Impact on math performance and self-reported Big Five (Males)

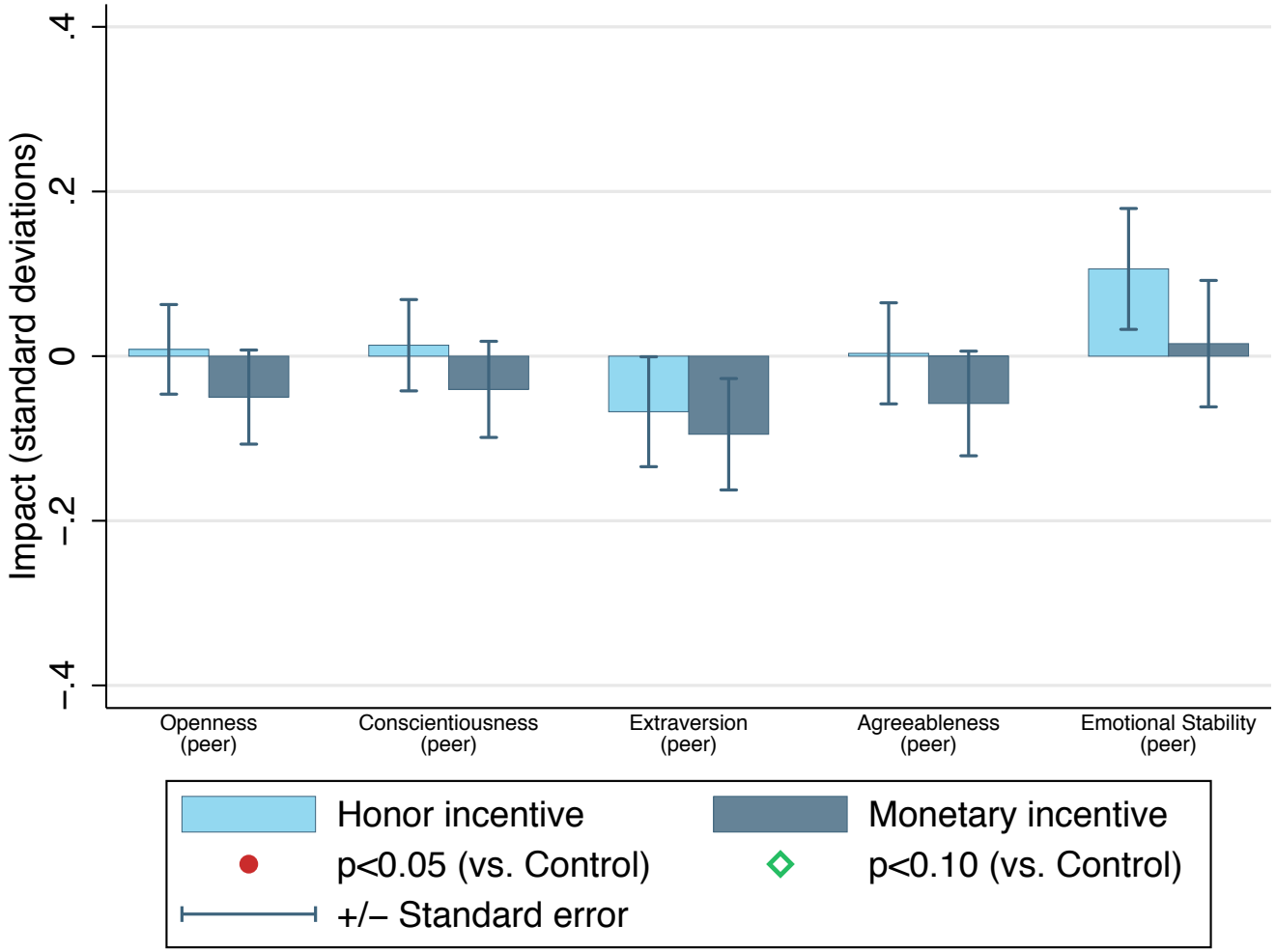


Impact on math performance and self-reported Big Five (Females)

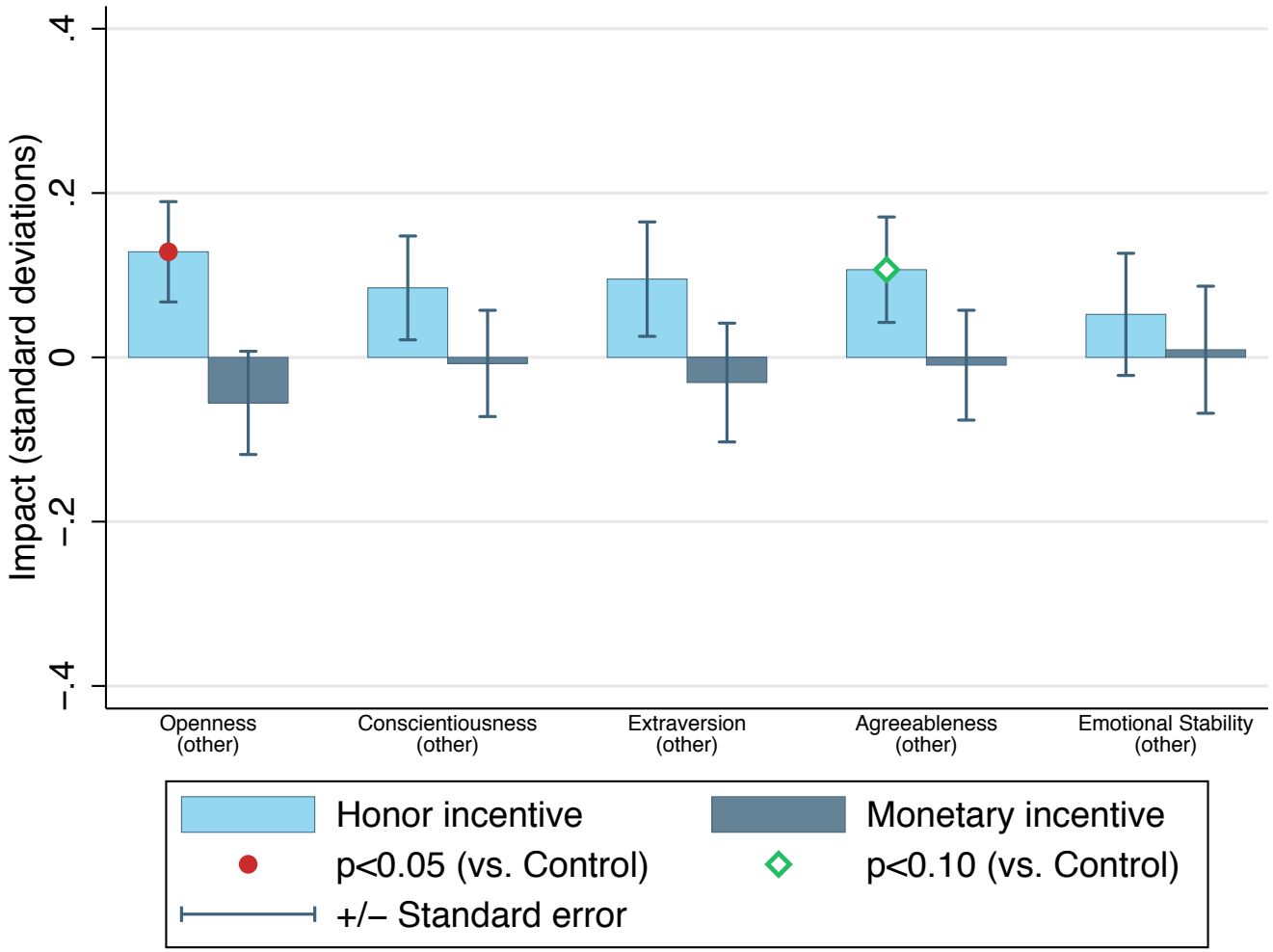


- Consider the impact of the treatments on how students rated their peers in two ways

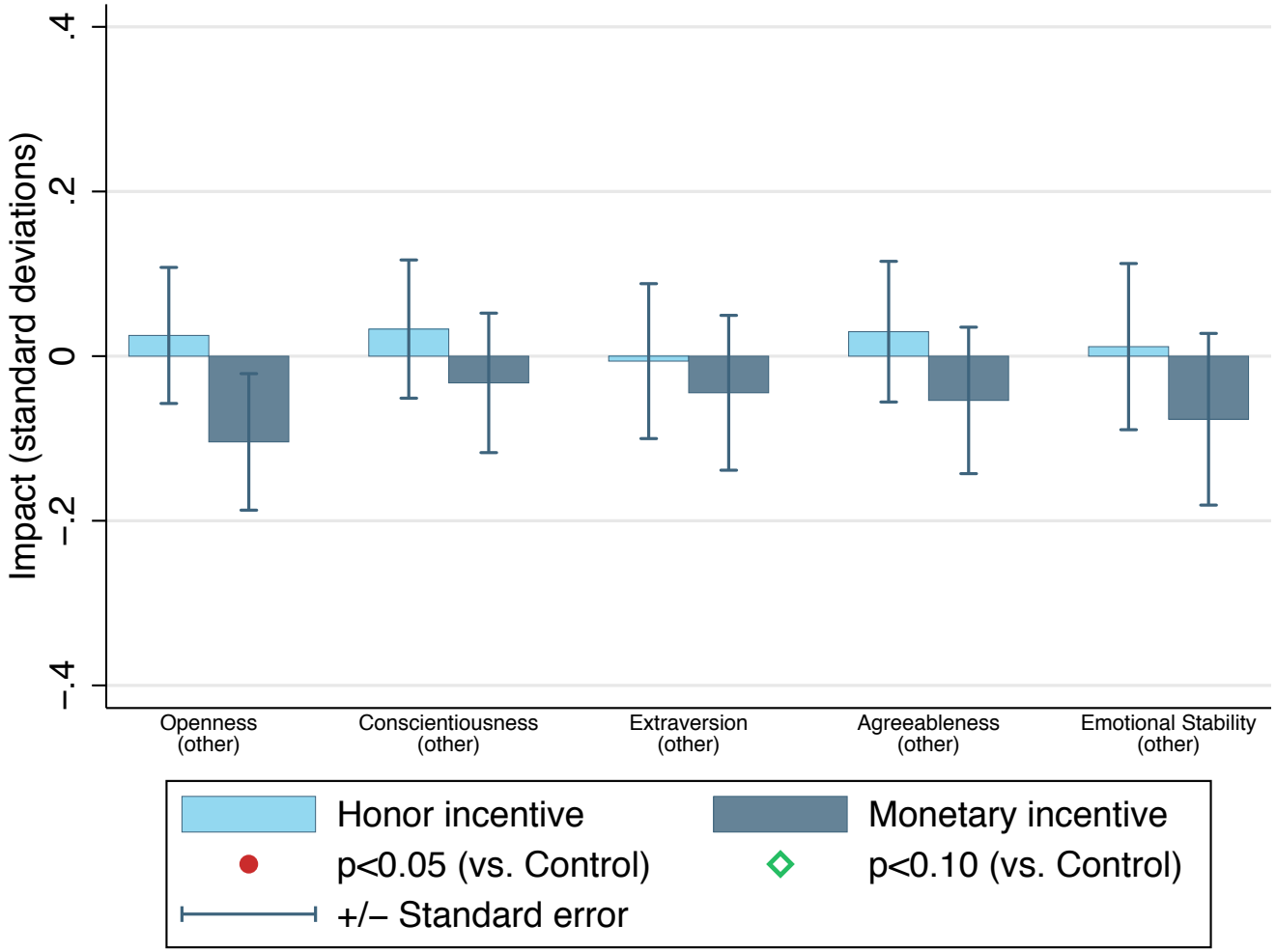
Impact on how treatment groups are assessed by peers (Full Sample)



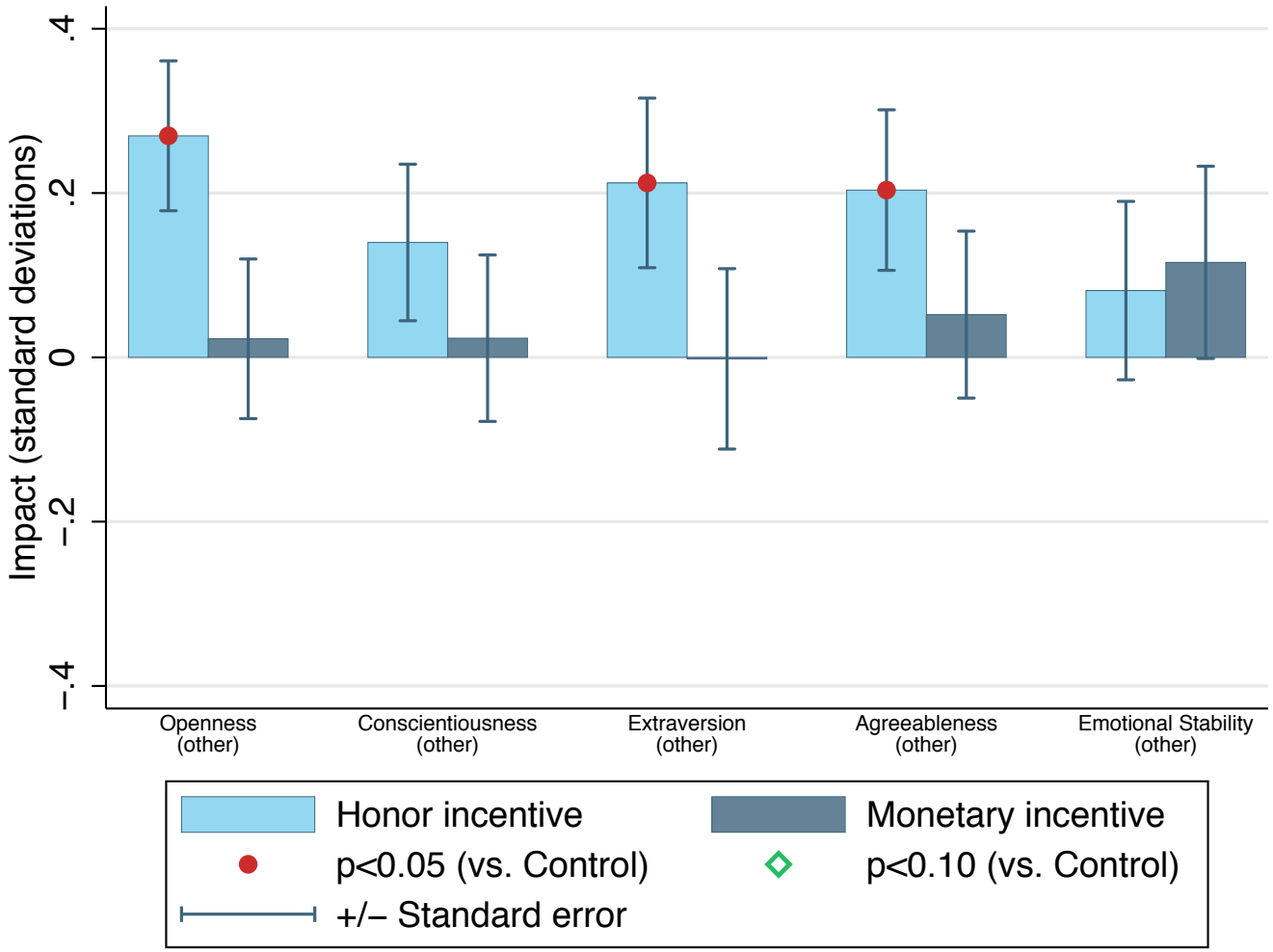
Impact on how treatment groups assess peers (Full Sample)



Impact on how treatment groups assess peers (Males)



Impact on how treatment groups assess peers (Females)



Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

4 Main Results

5 Subgroup Analyses

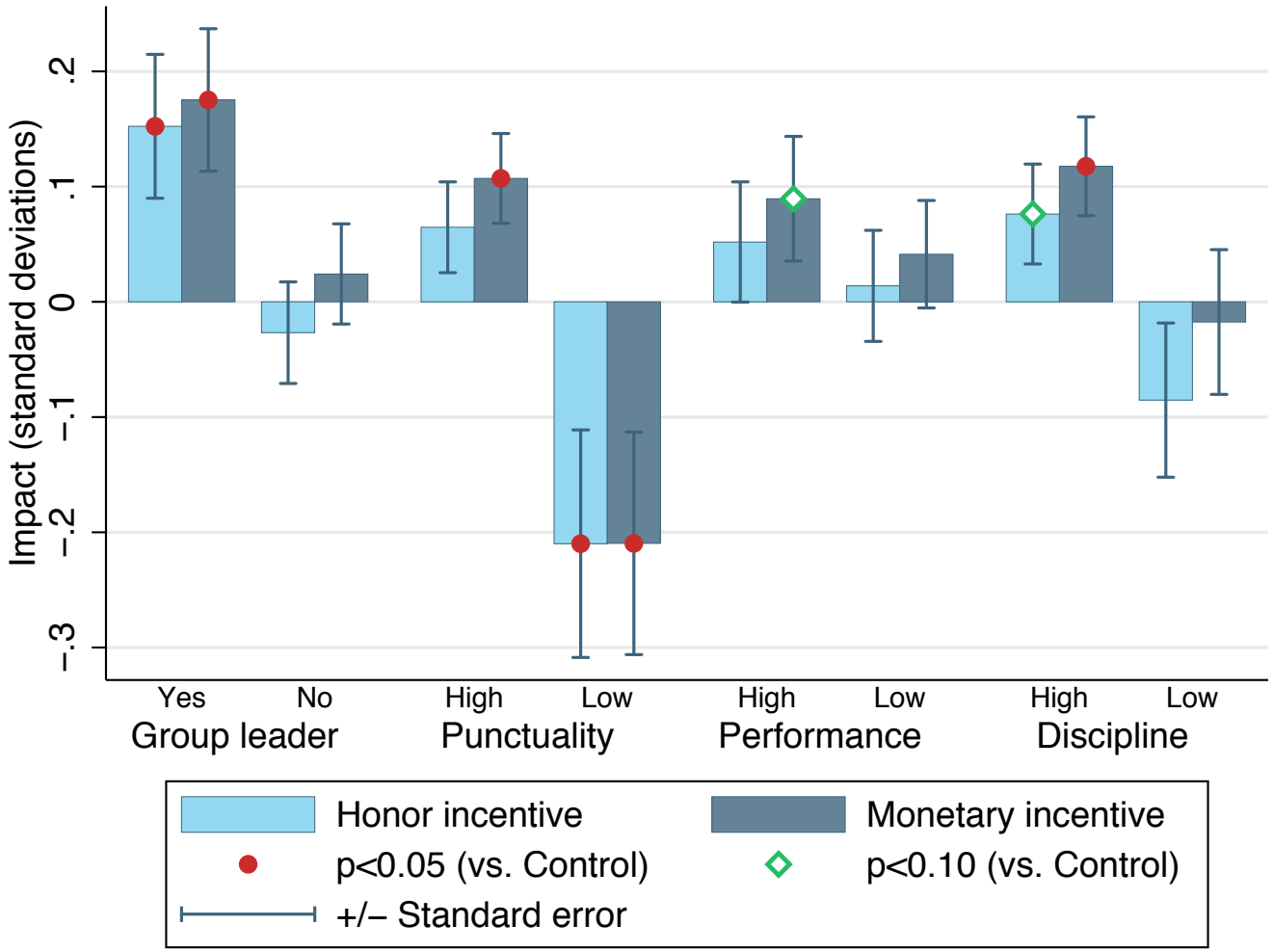
6 Conclusion

7 Appendix

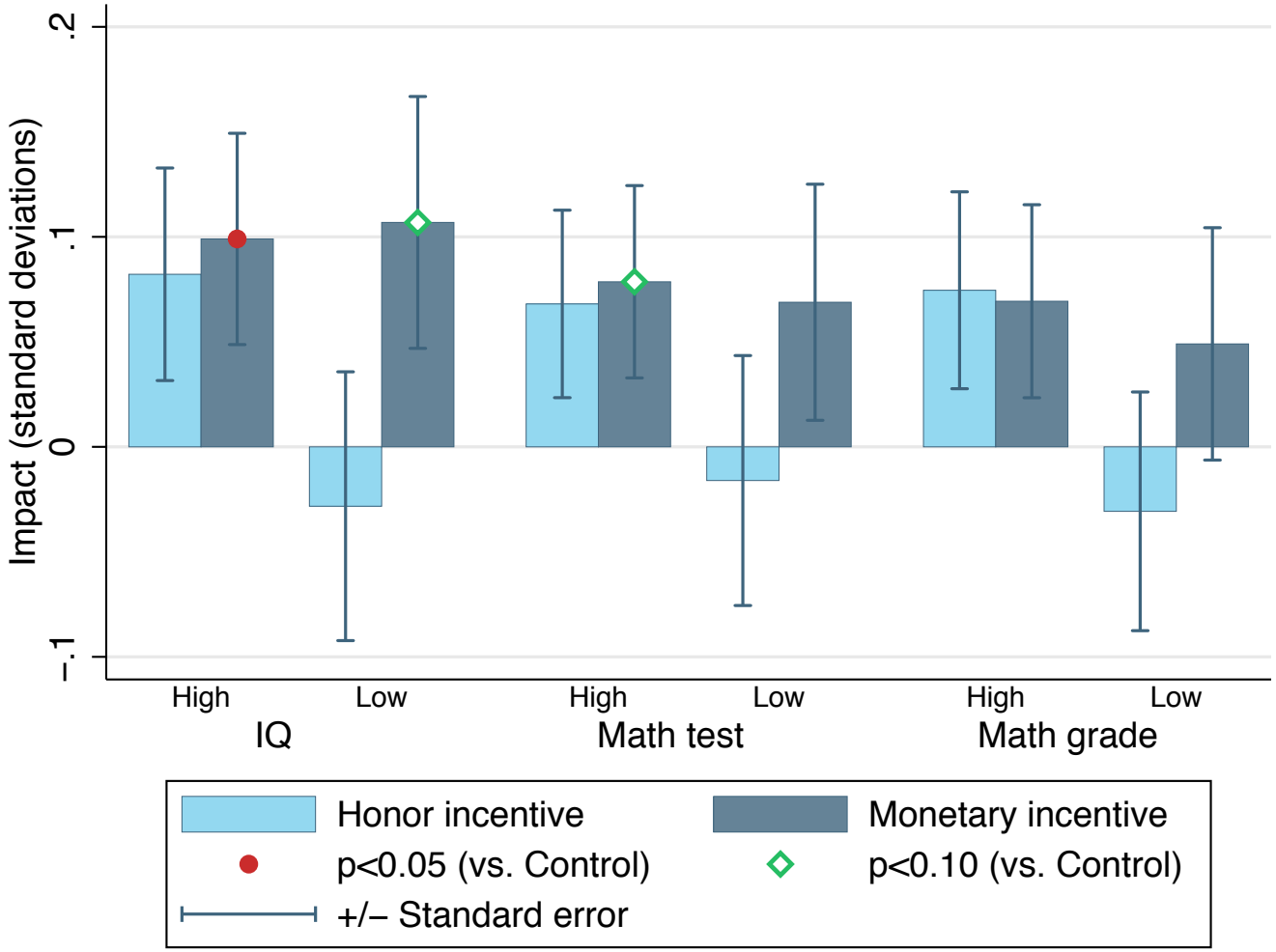
Some subgroups responded differently to incentives

- Better performing and better behaved students performed better on the math test in response to incentives
- The patterns were less consistent when examining self-reported Big Five outcomes

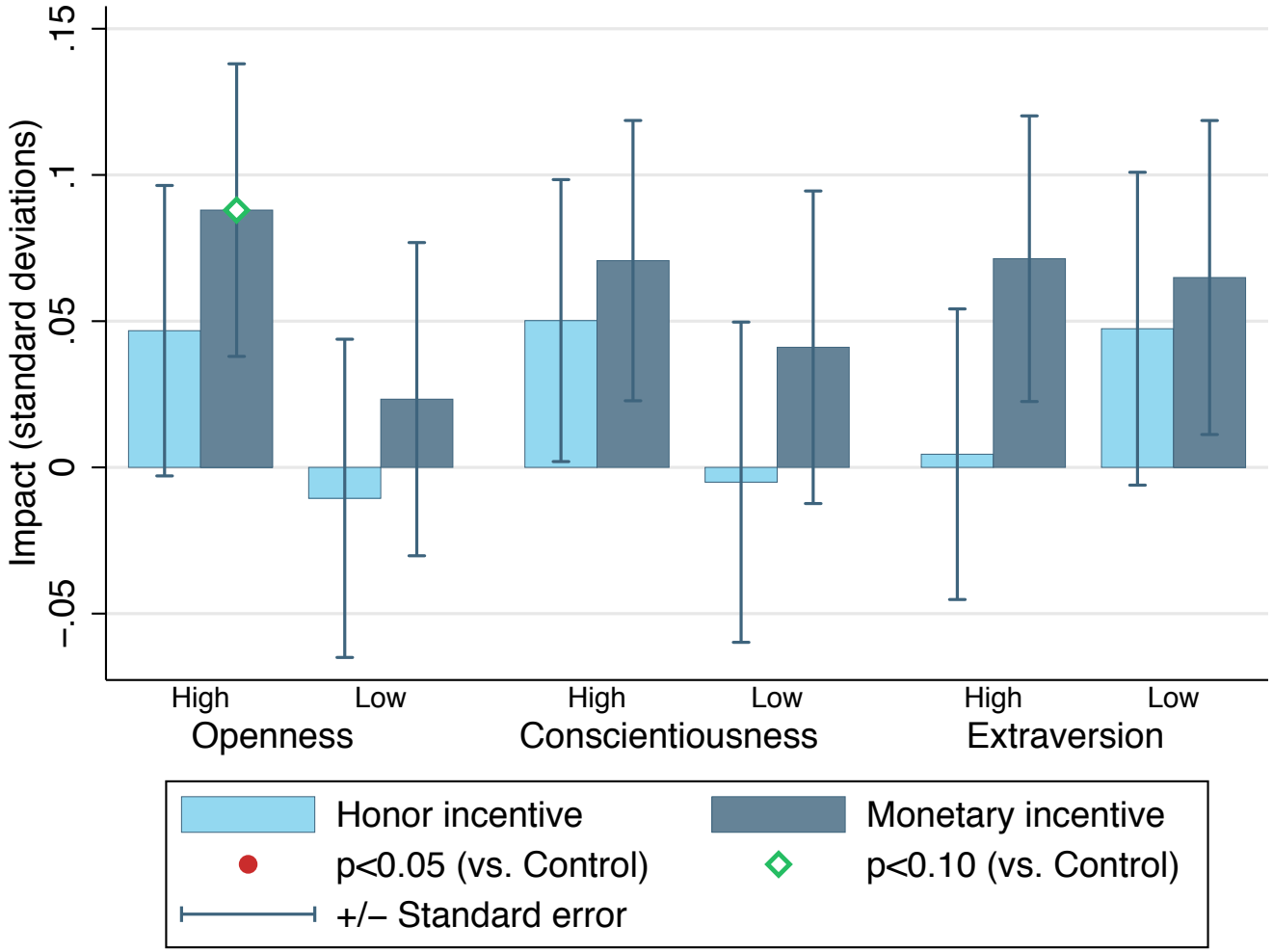
Impact on math test scores by subgroup based on non-cognitive measures



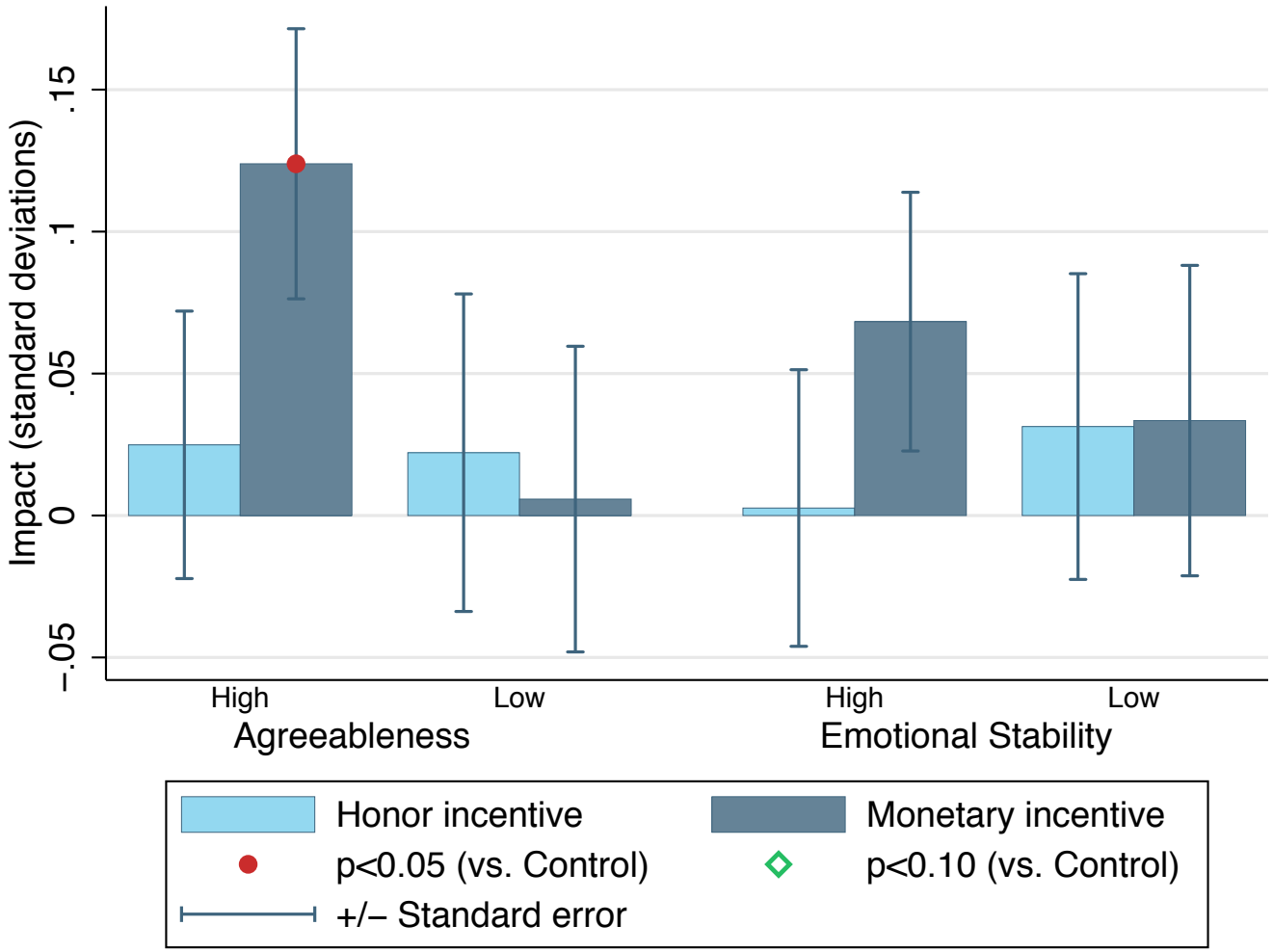
Impact on math test scores by subgroup based on cognitive measures



Impact on math test scores by subgroup based on Big Five (part 1)



Impact on math test scores by subgroup based on Big Five (part 2)



Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

4 Main Results

5 Subgroup Analyses

6 Conclusion

7 Appendix

Conclusion

- Self-reported measures are susceptible to unintended biases
- Standardizing for aspects of the situation will be important for policy evaluation and school accountability

Outline

1 Introduction

2 Experimental Design and Data

3 Analytical Methods

4 Main Results

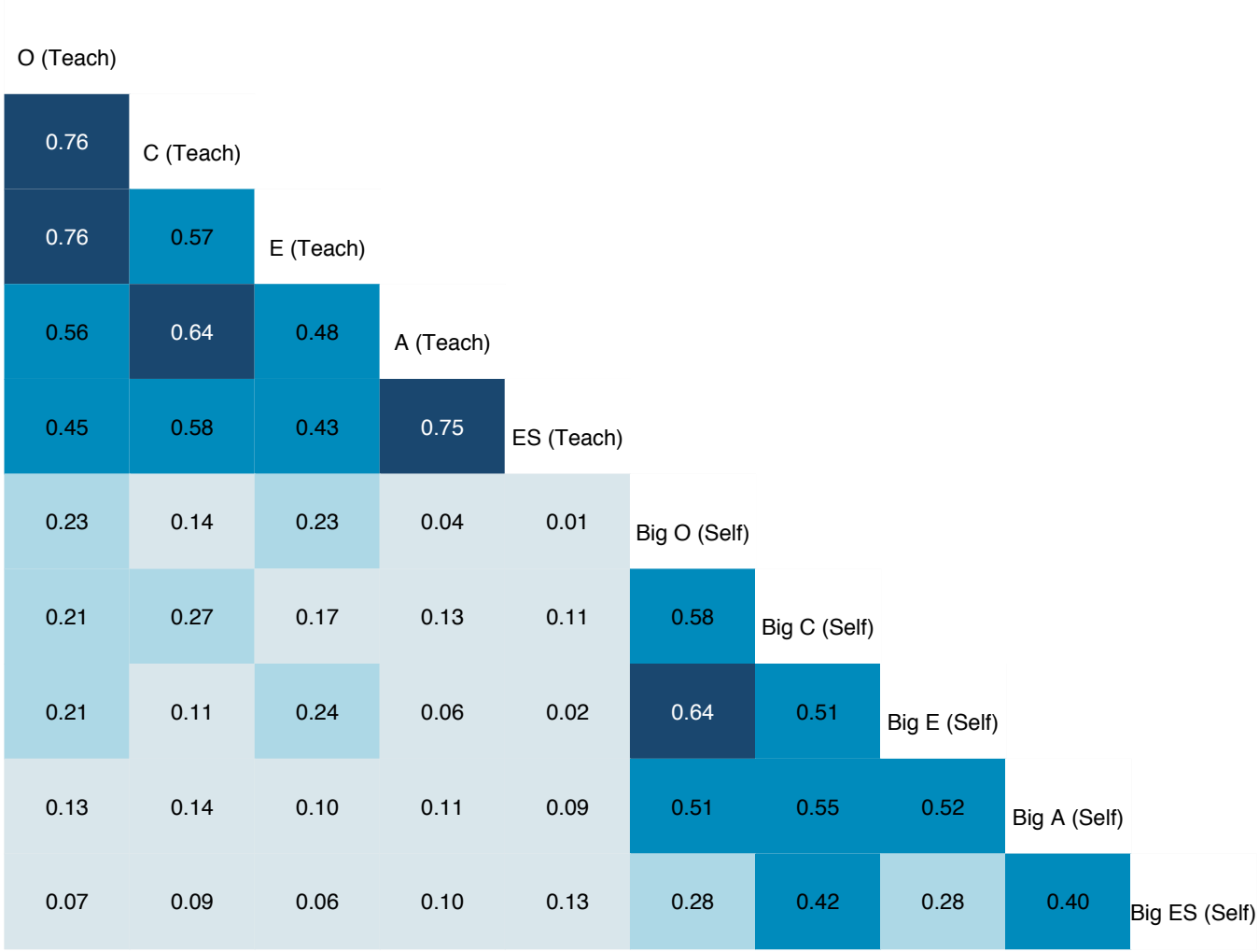
5 Subgroup Analyses

6 Conclusion

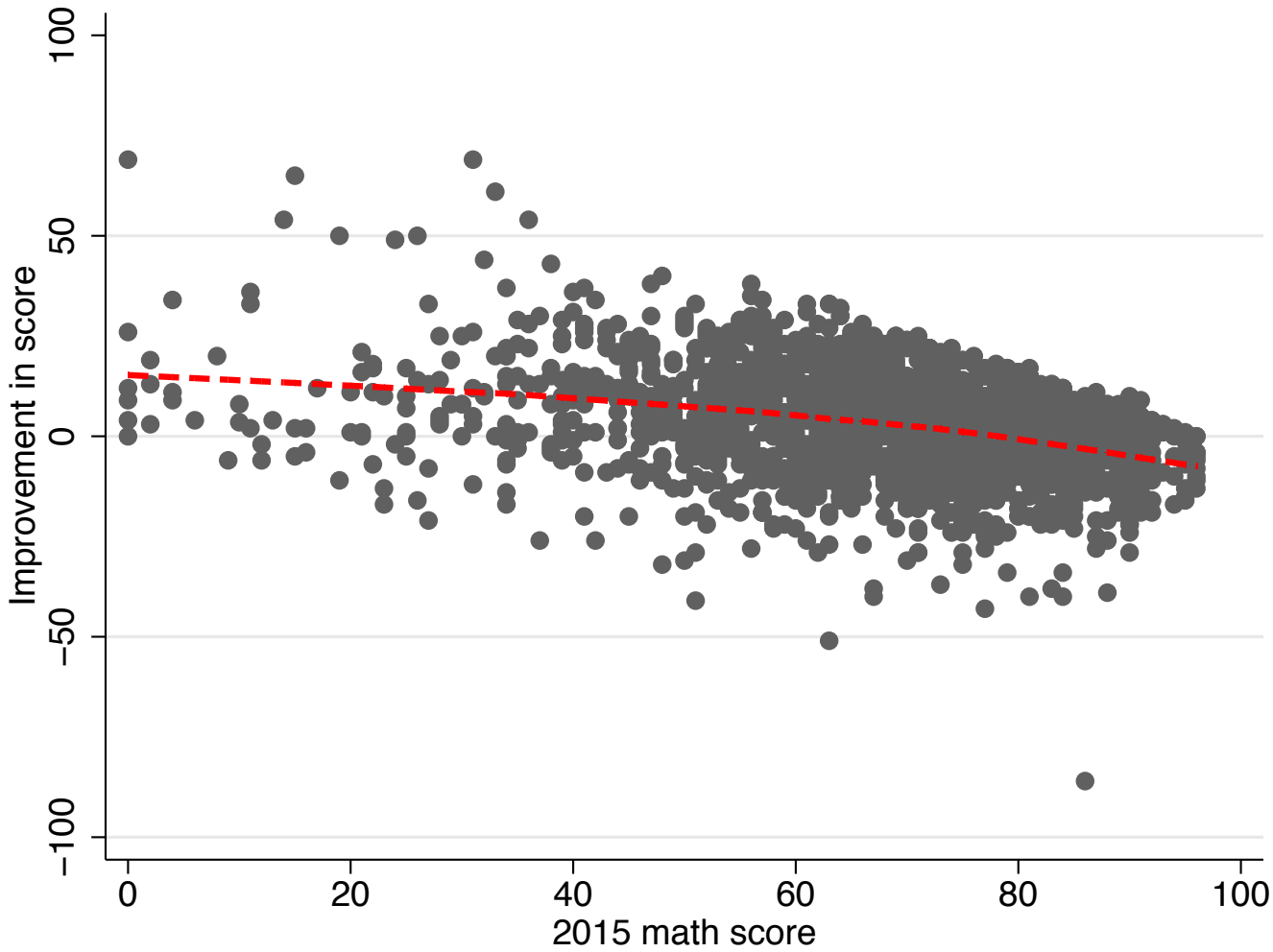
7 Appendix

Correlations between cognitive and non-cognitive measures

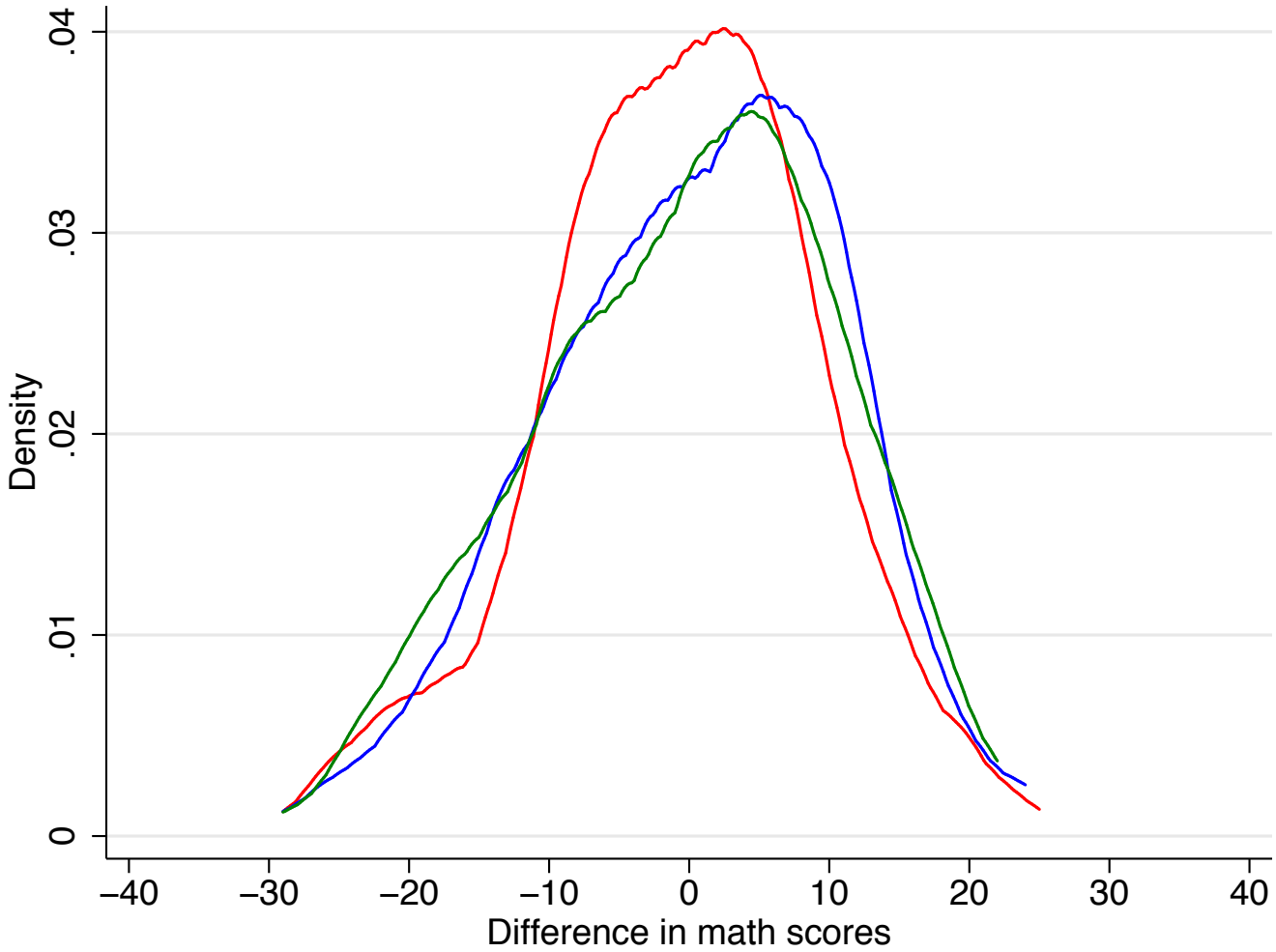
Correlations between Big Five from teacher- and self-reports



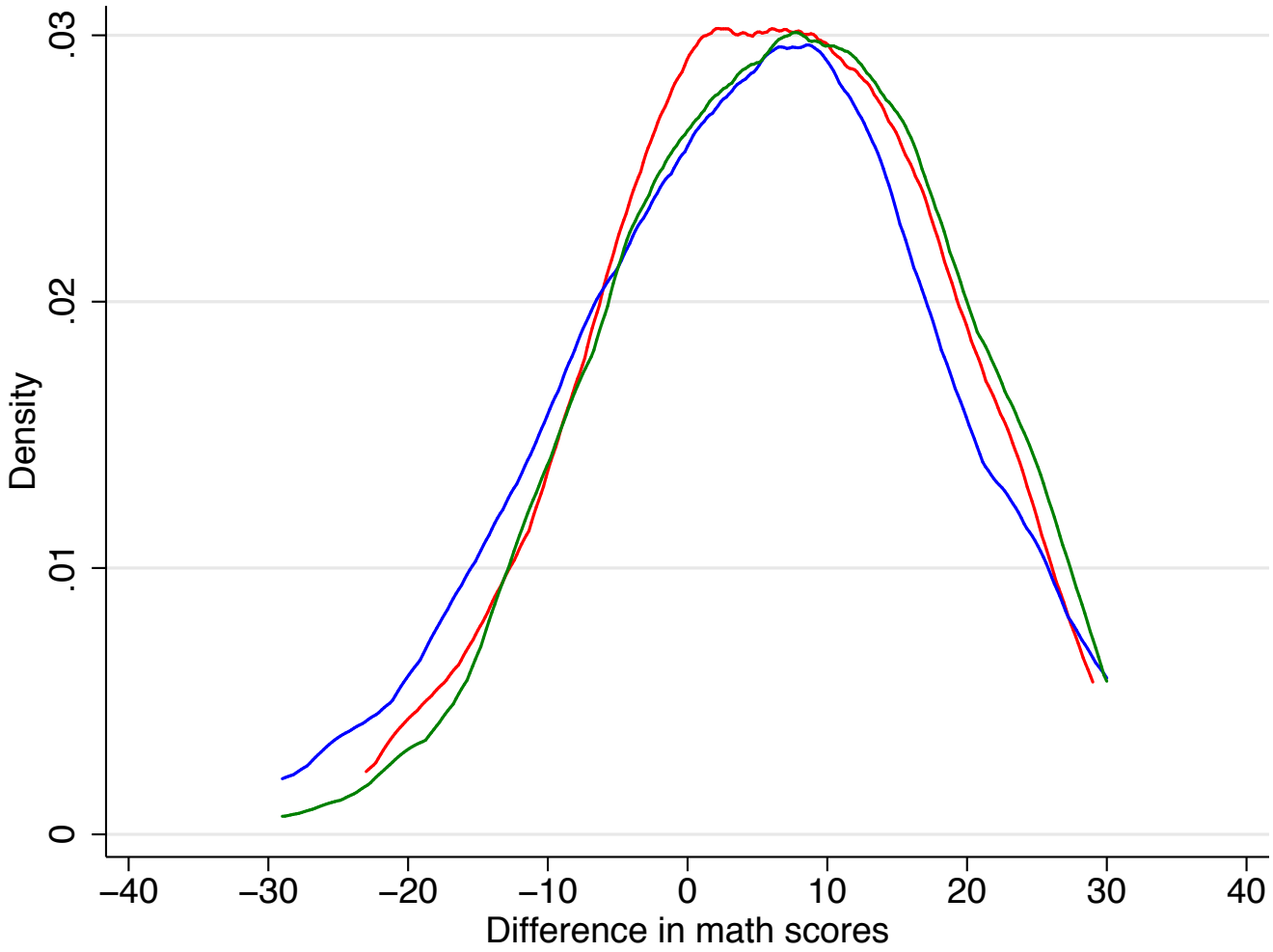
Low-performing students showed the most improvement



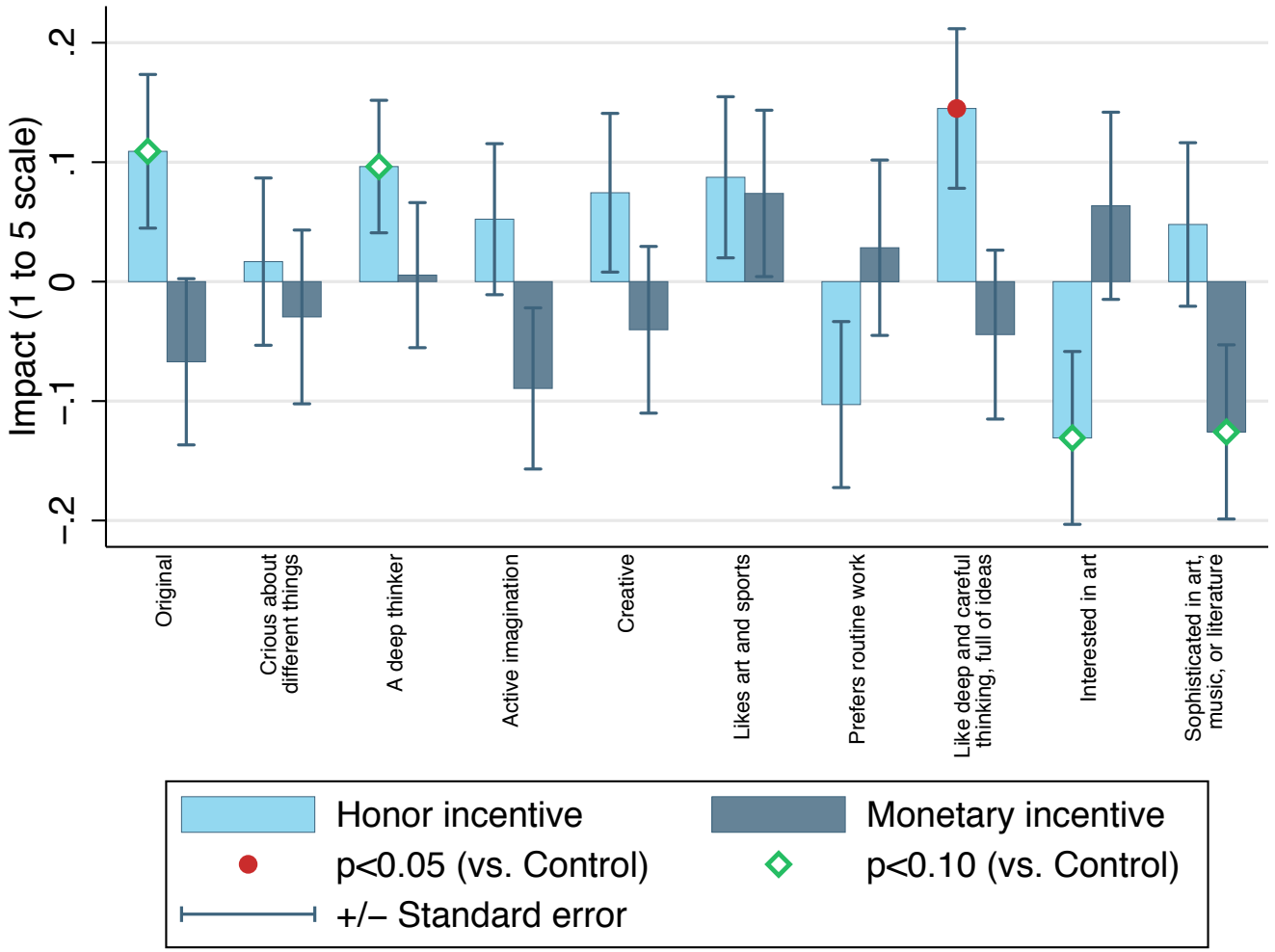
Distribution of differences in math test scores for students above the median on the 2015 test



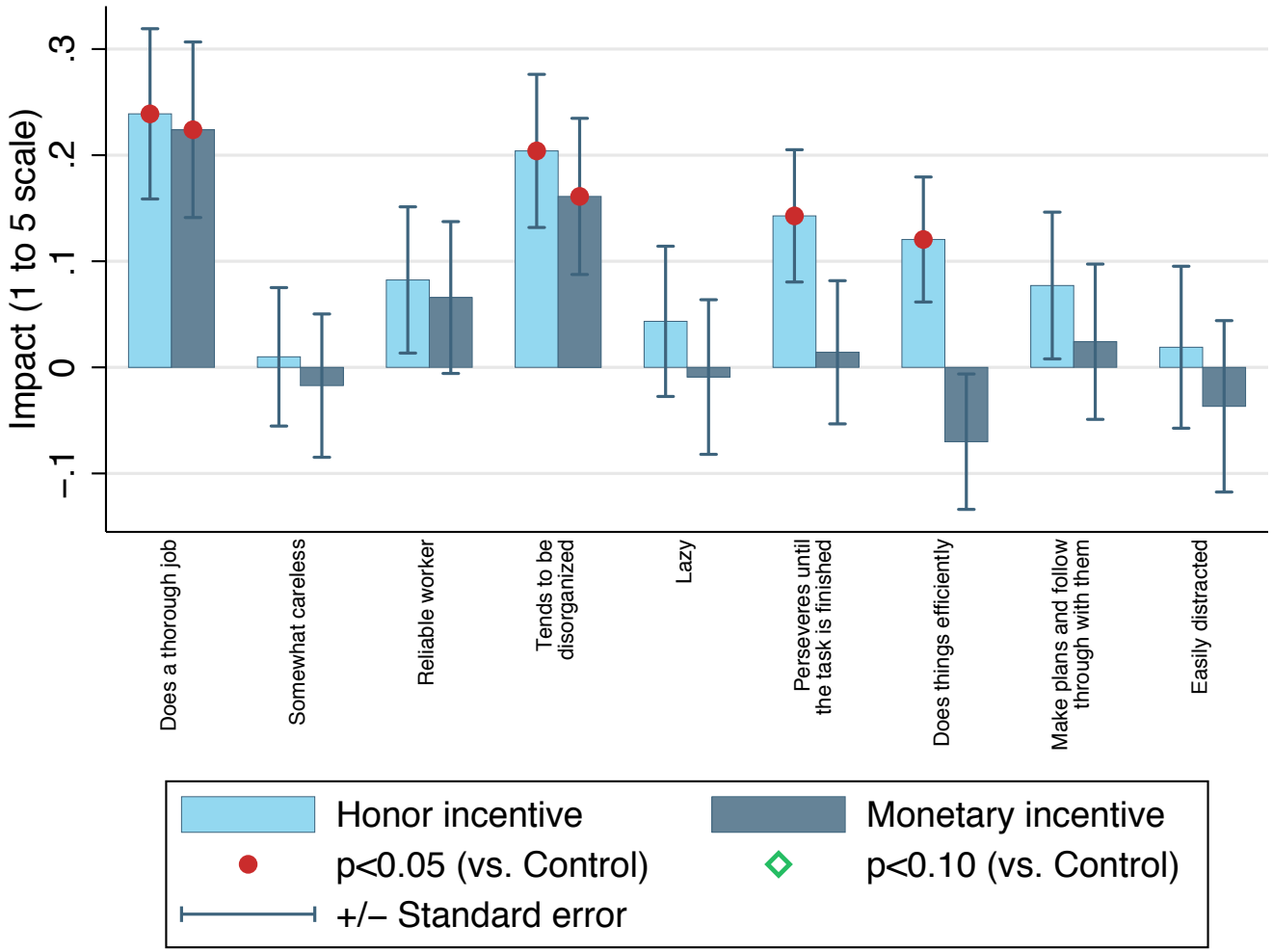
Distribution of differences in math test scores for students below the median on the 2015 test



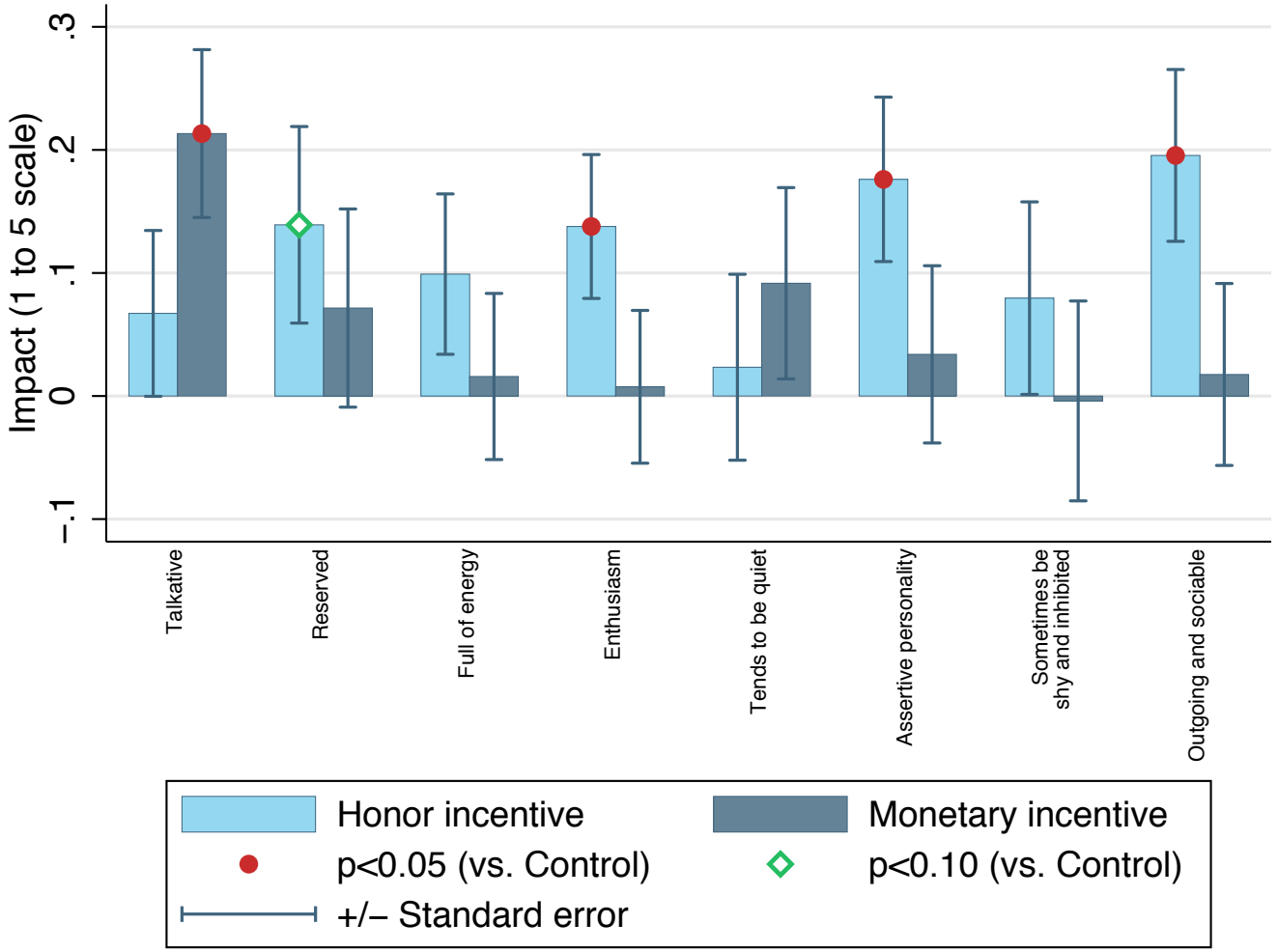
Impact on individual items of Openness to Experience



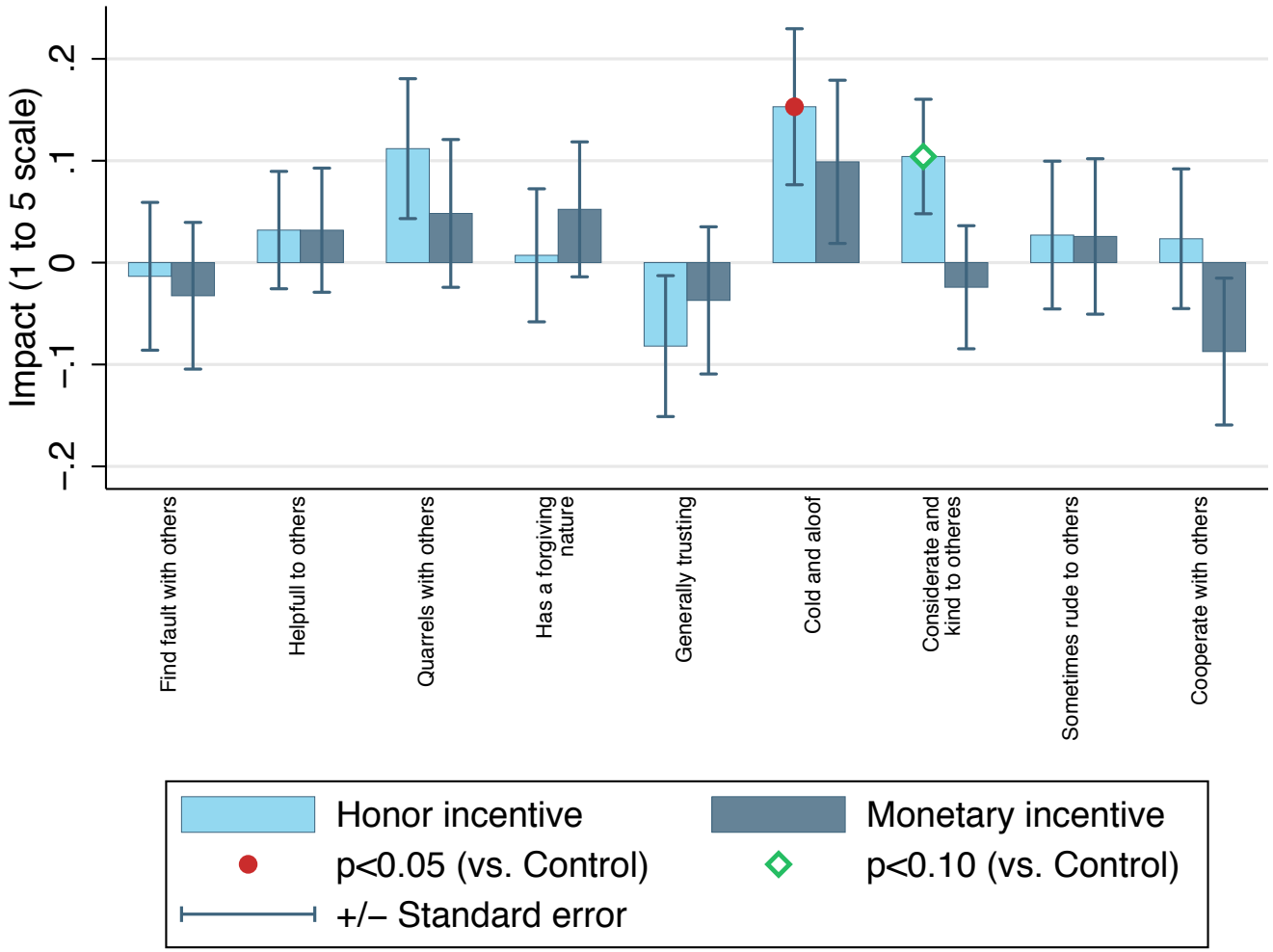
Impact on individual items of Conscientiousness



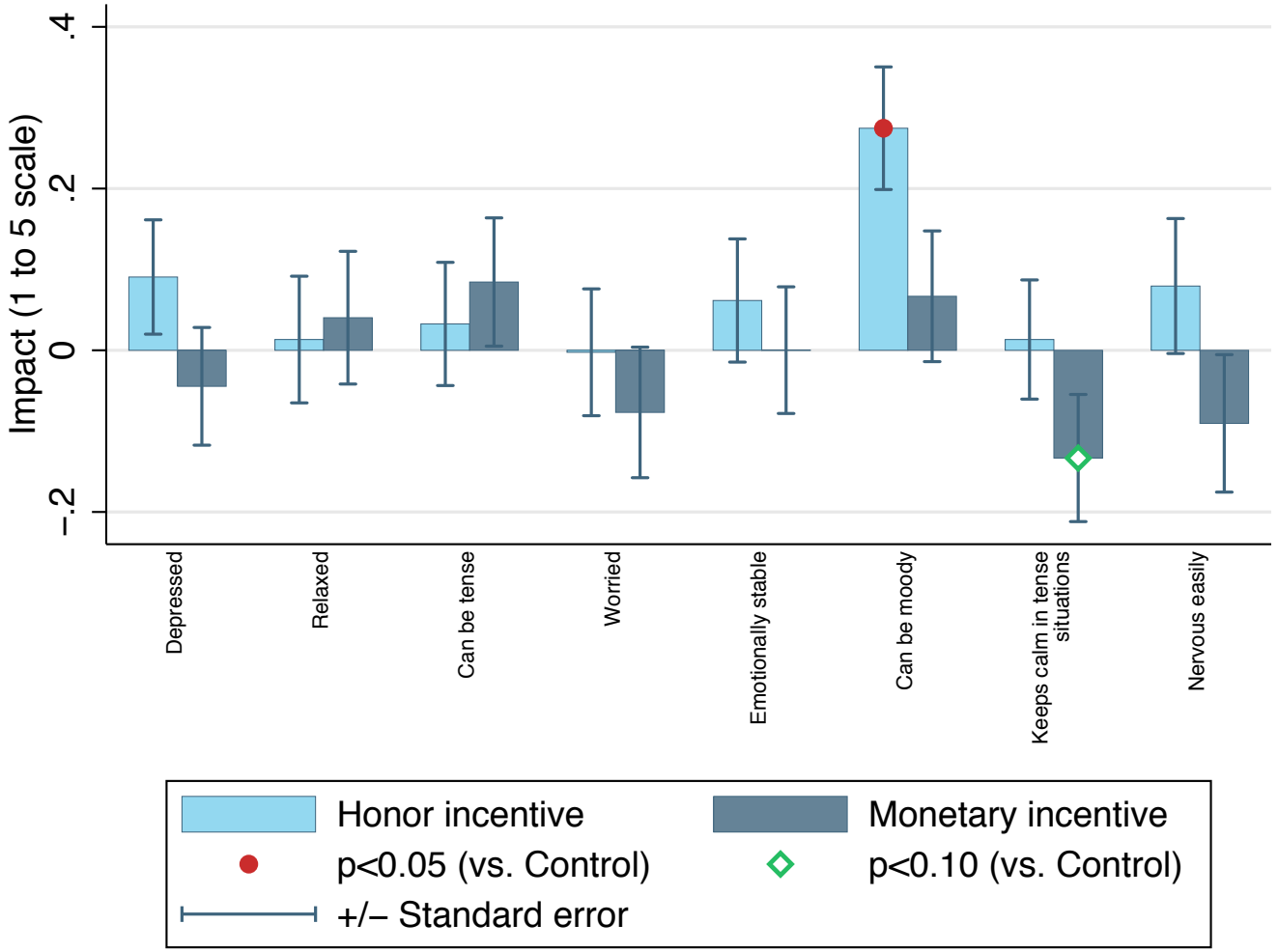
Impact on individual items of Extraversion



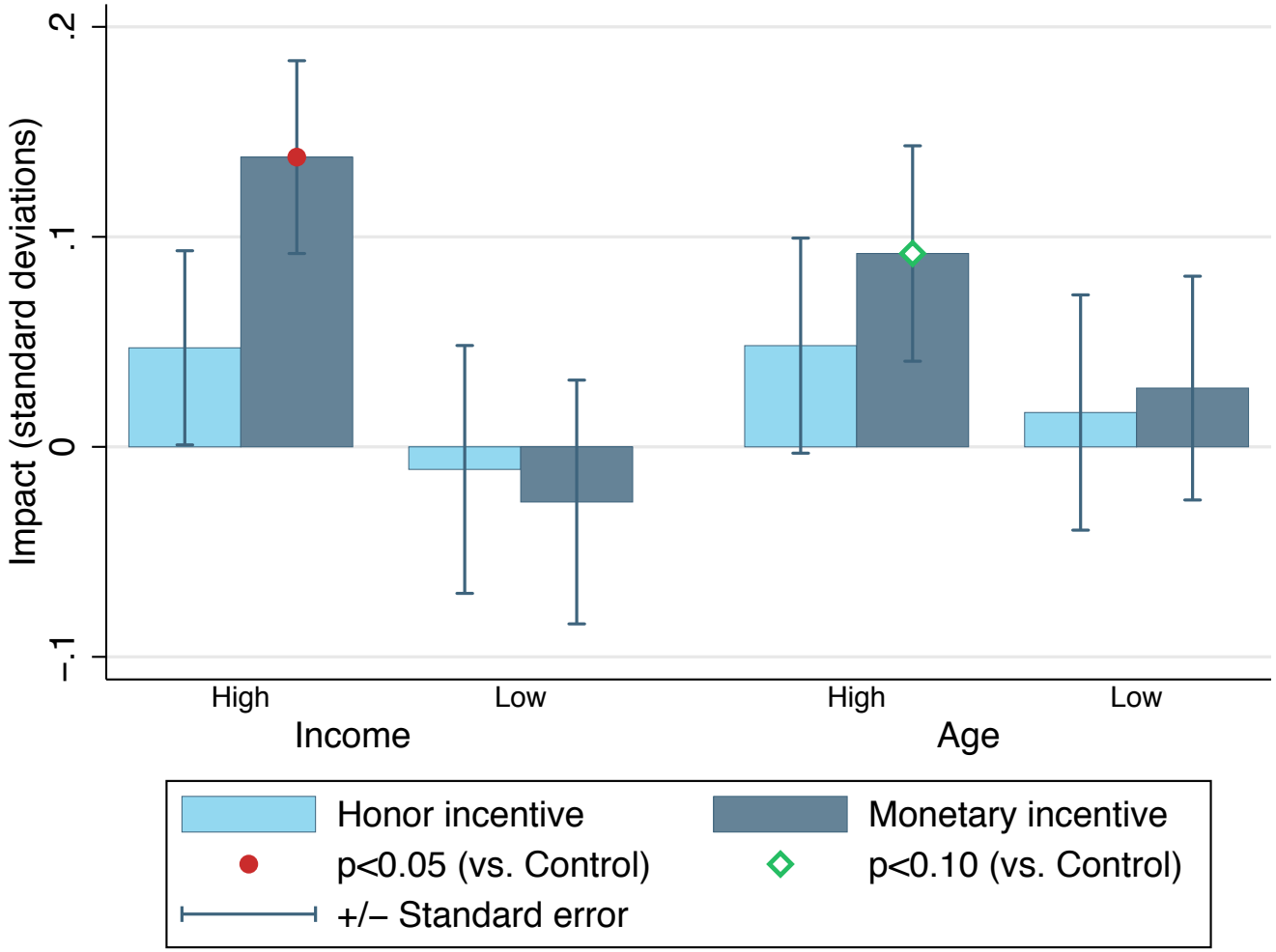
Impact on individual items of Agreeableness



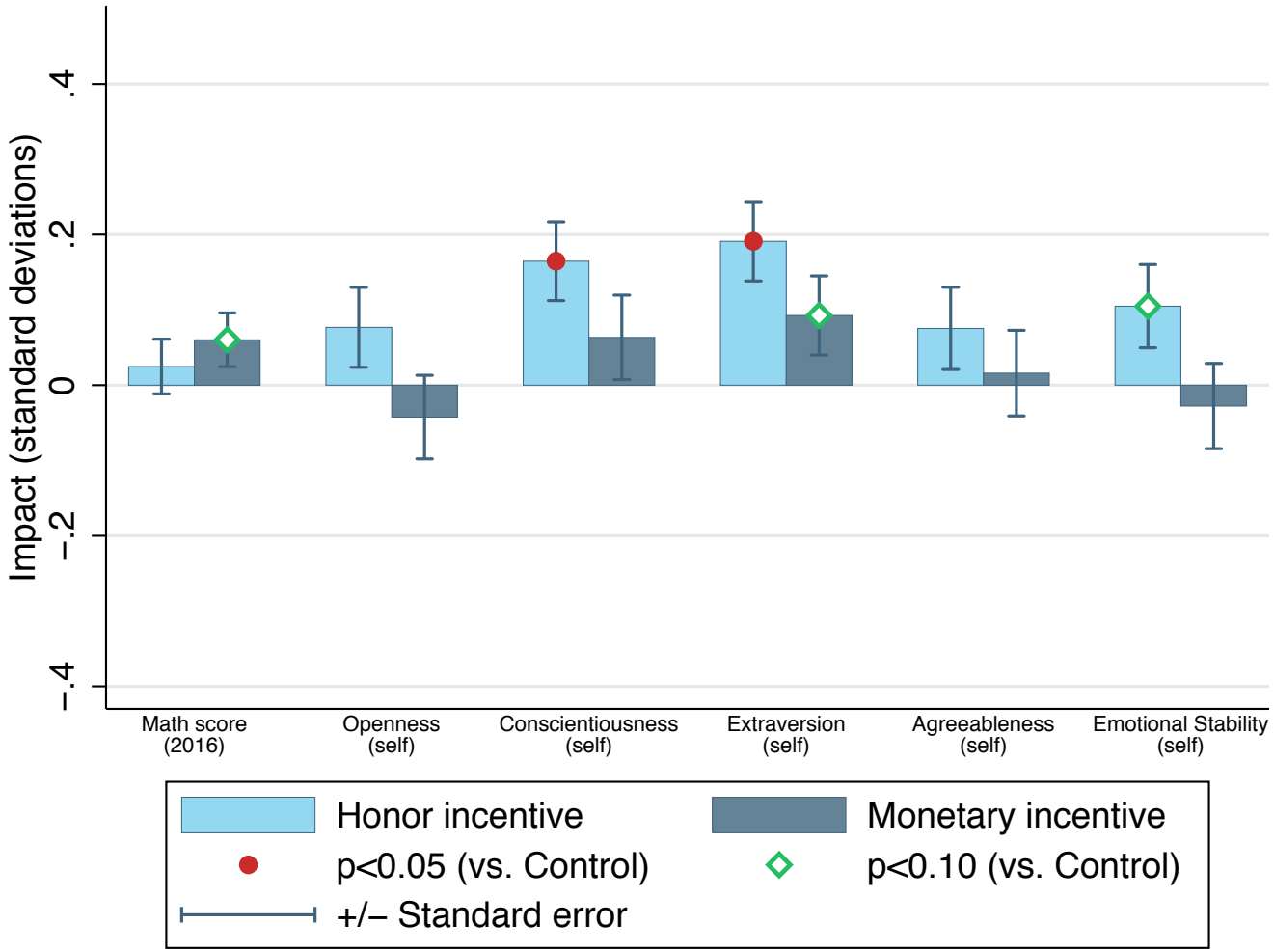
Impact on individual items of Neuroticism



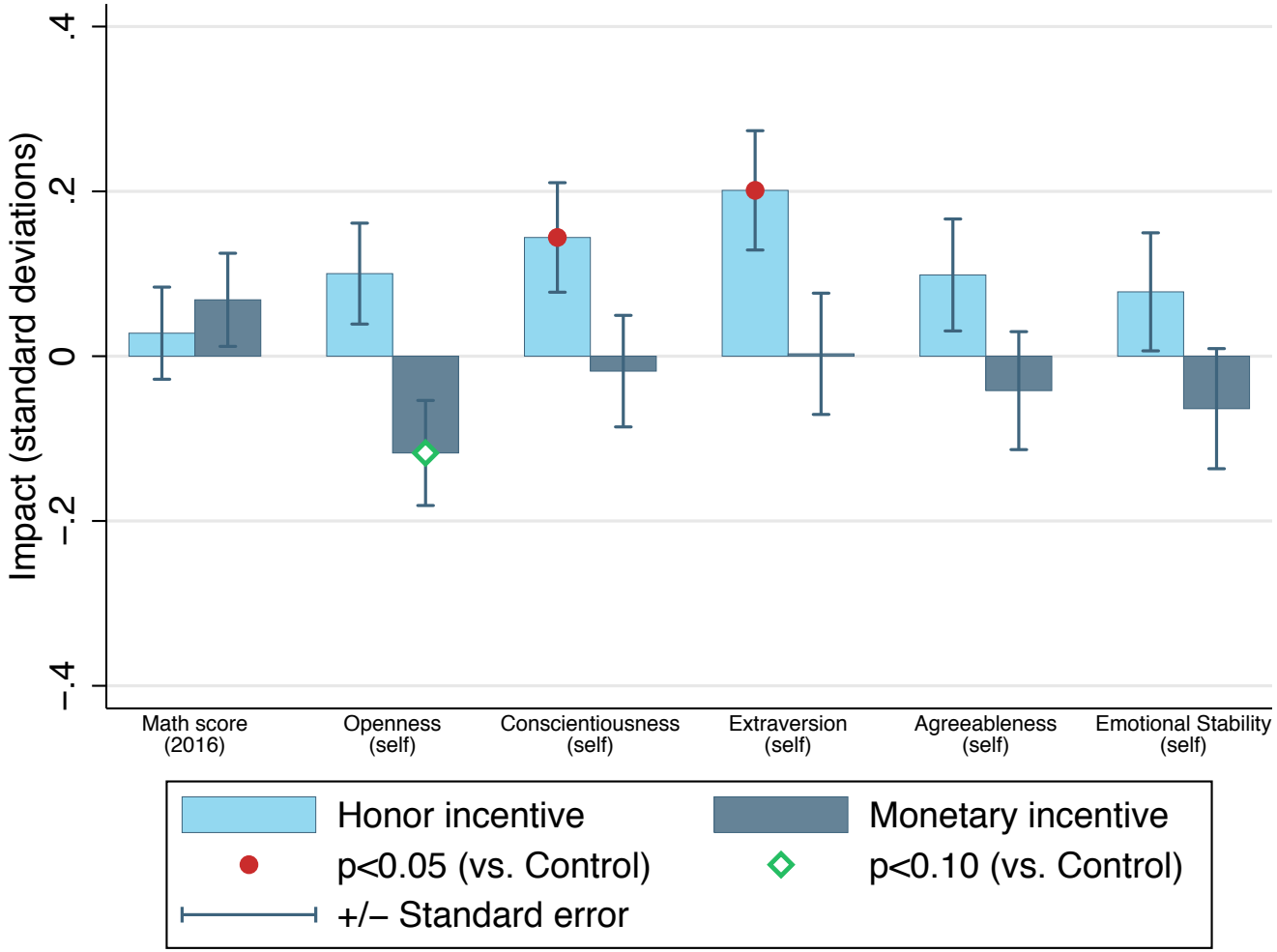
Impact on math test scores by subgroup based on demographics



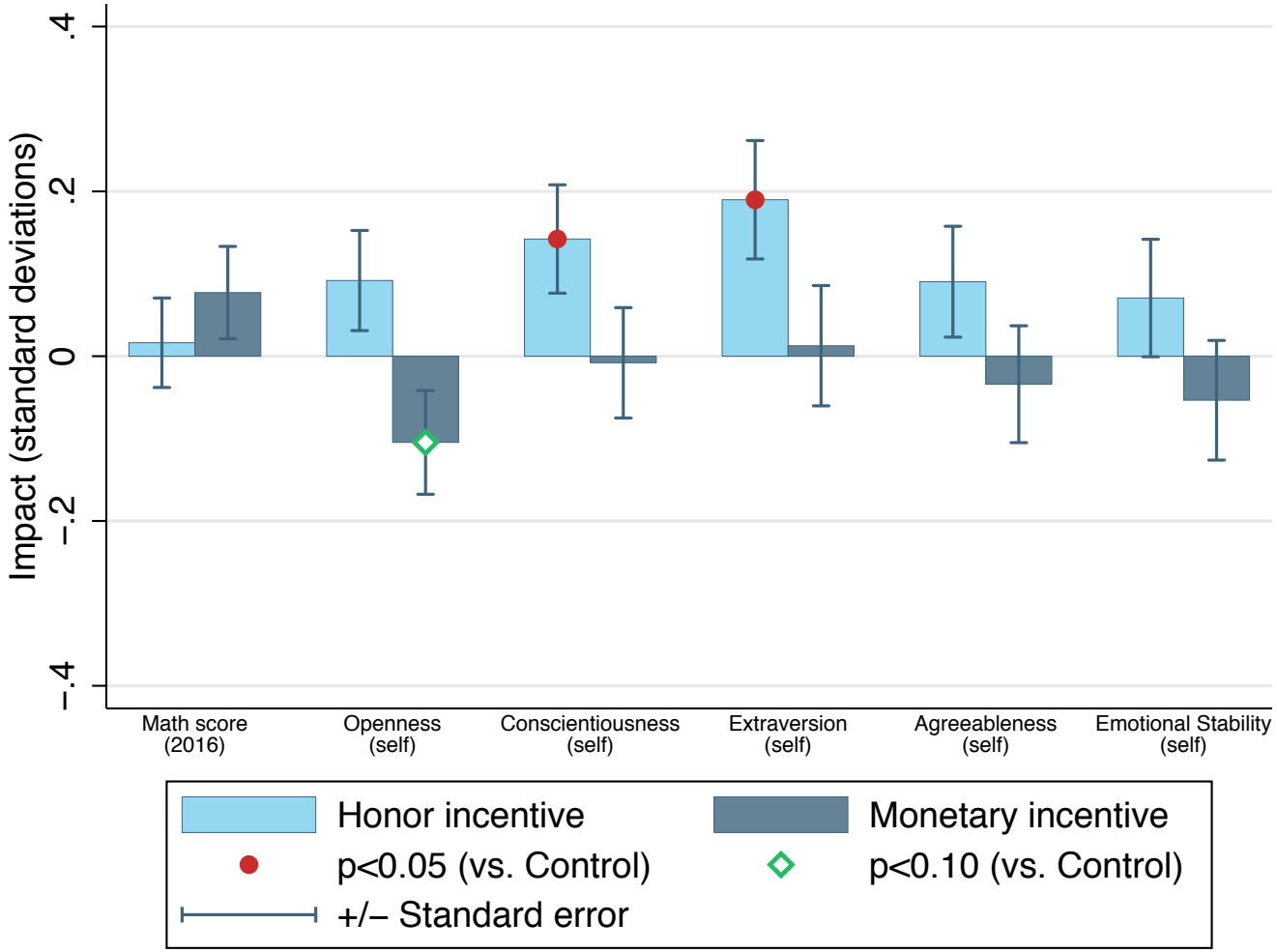
Impact on math performance and self-reported Big Five (Full Sample), mean scores



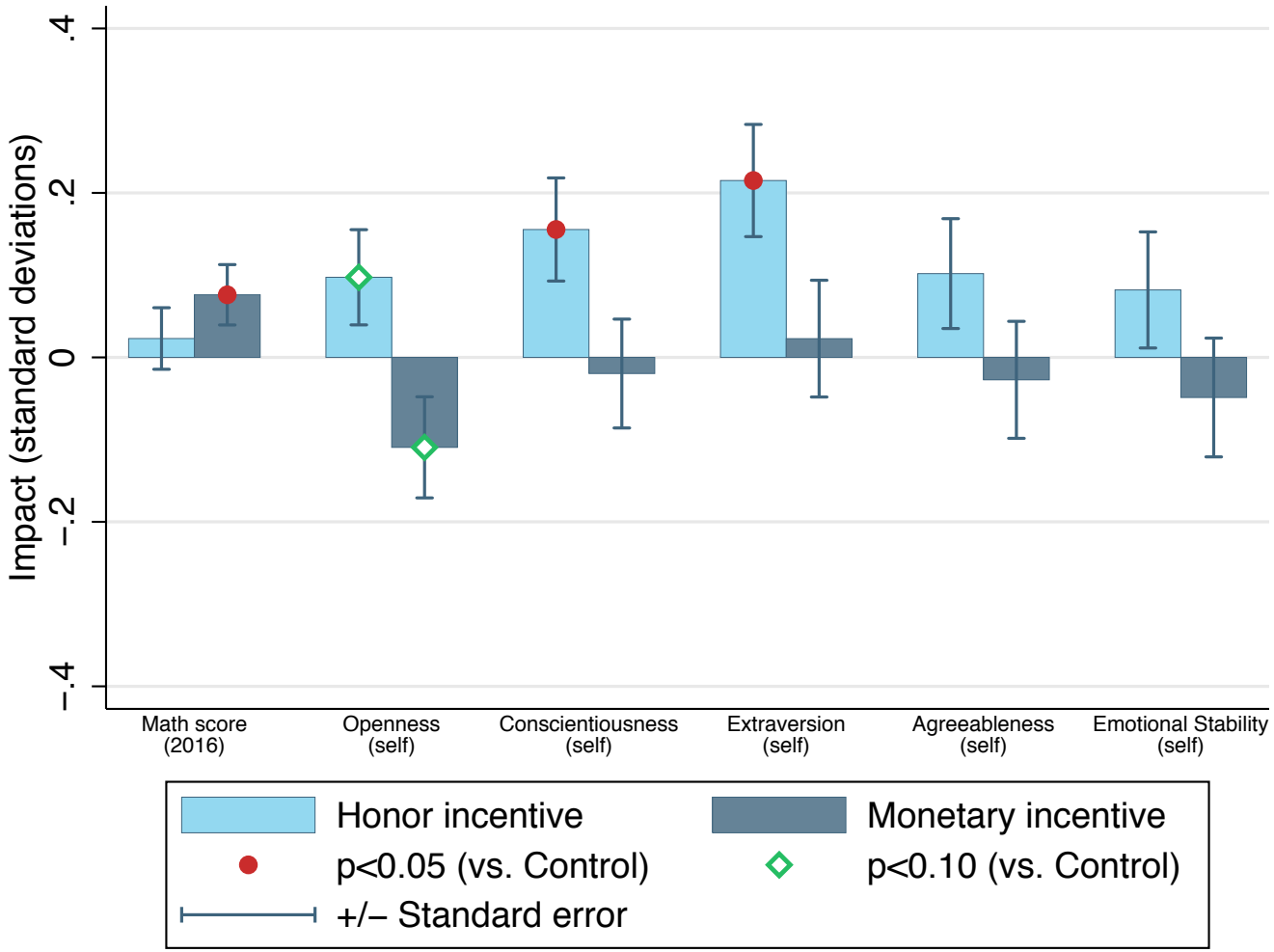
Impact on math performance and self-reported Big Five (Full Sample), no controls



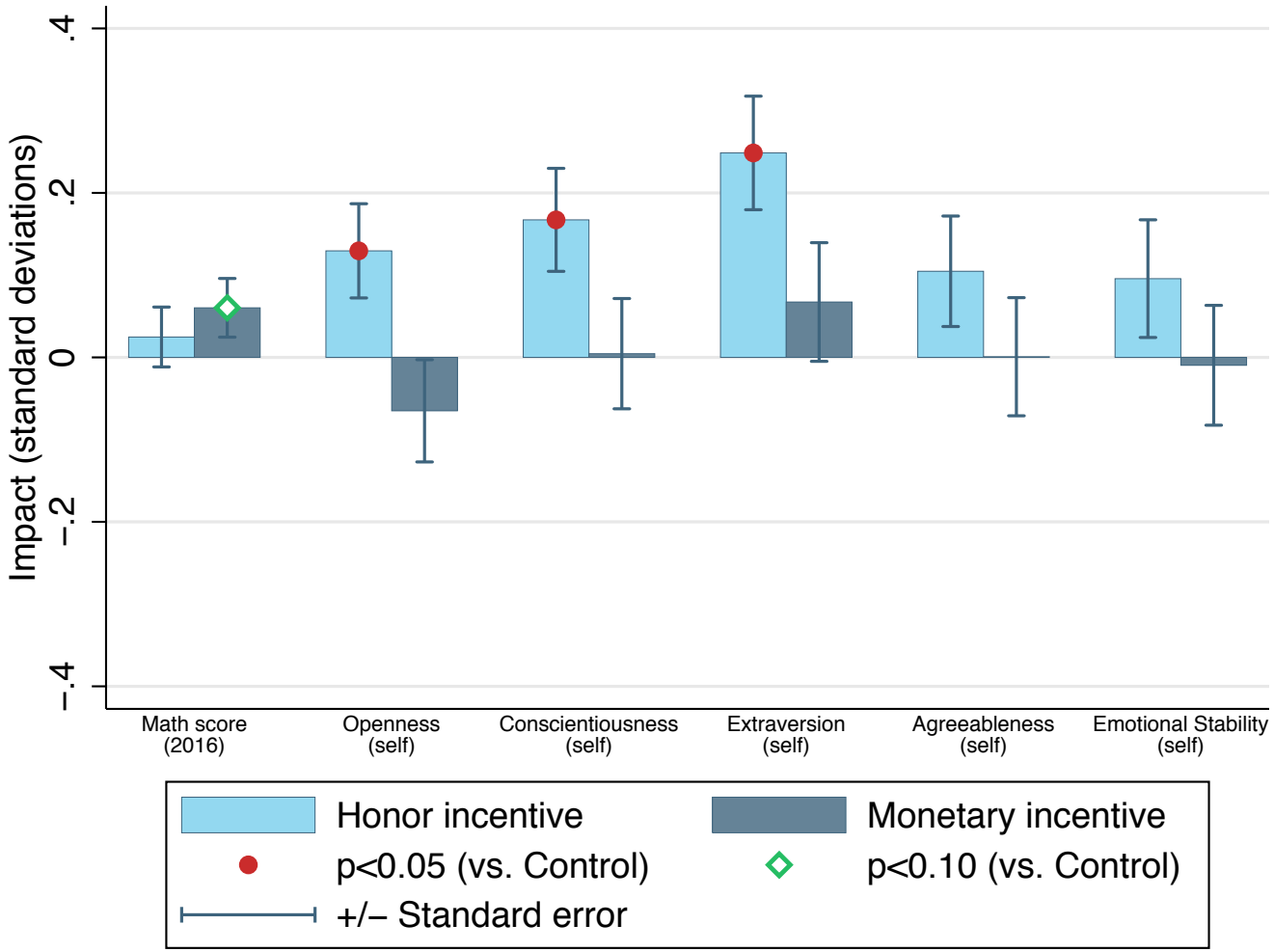
Impact on math performance and self-reported Big Five (Full Sample), basic demographics



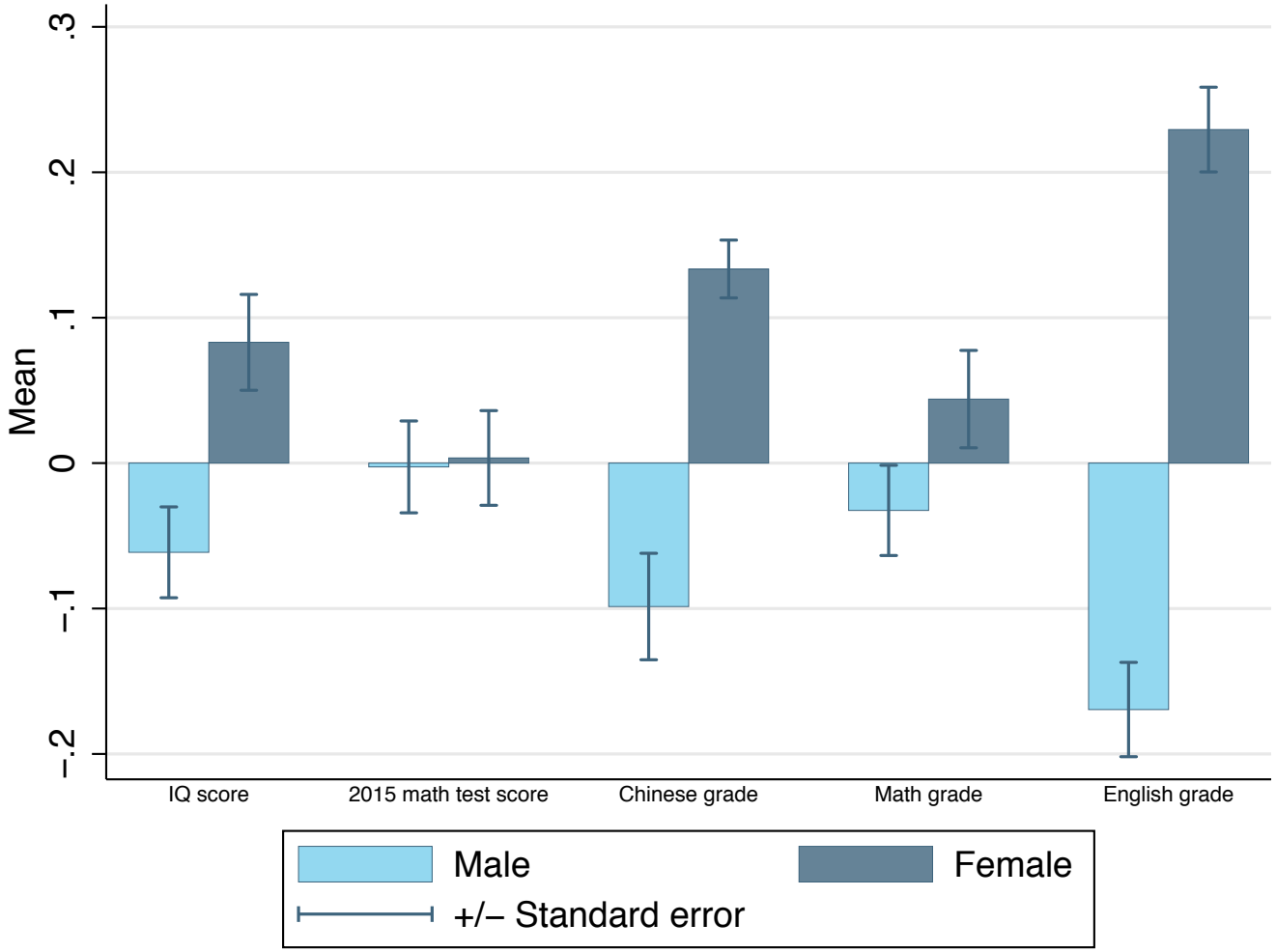
Impact on math performance and self-reported Big Five (Full Sample), basic demographics and ability



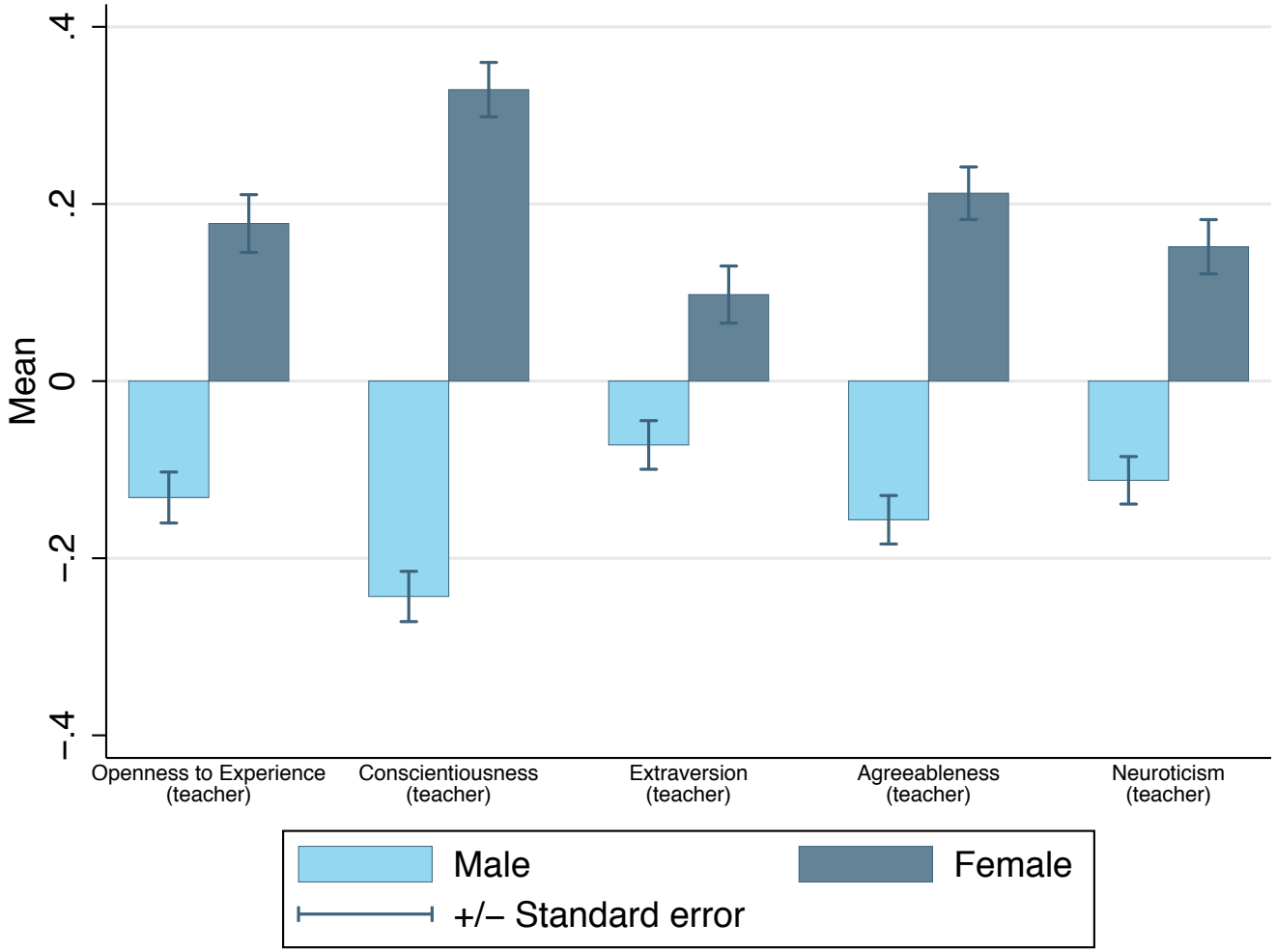
Impact on math performance and self-reported Big Five (Full Sample), basic demographics, ability, and school fixed effects



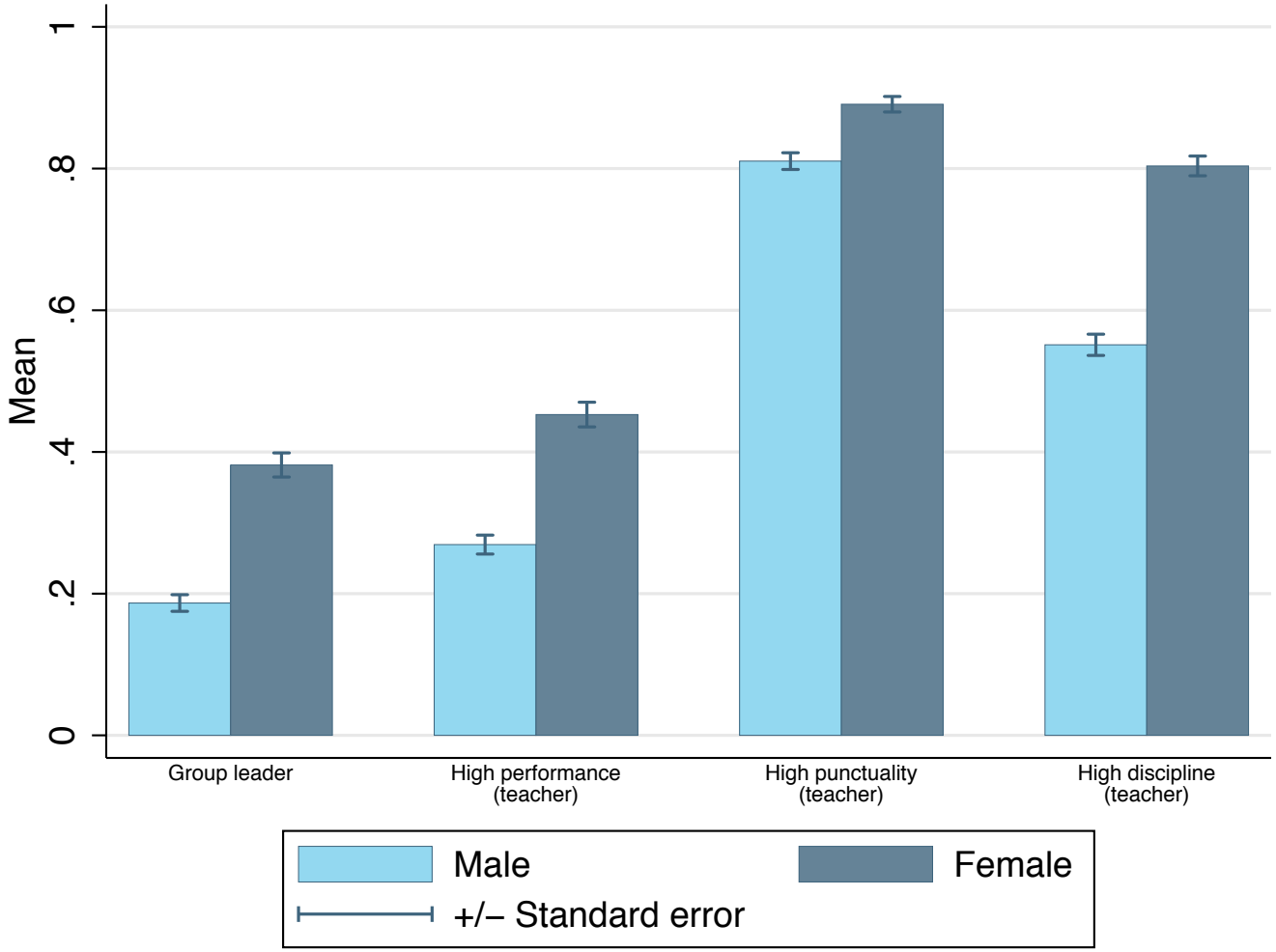
Female students had higher levels of cognitive ability



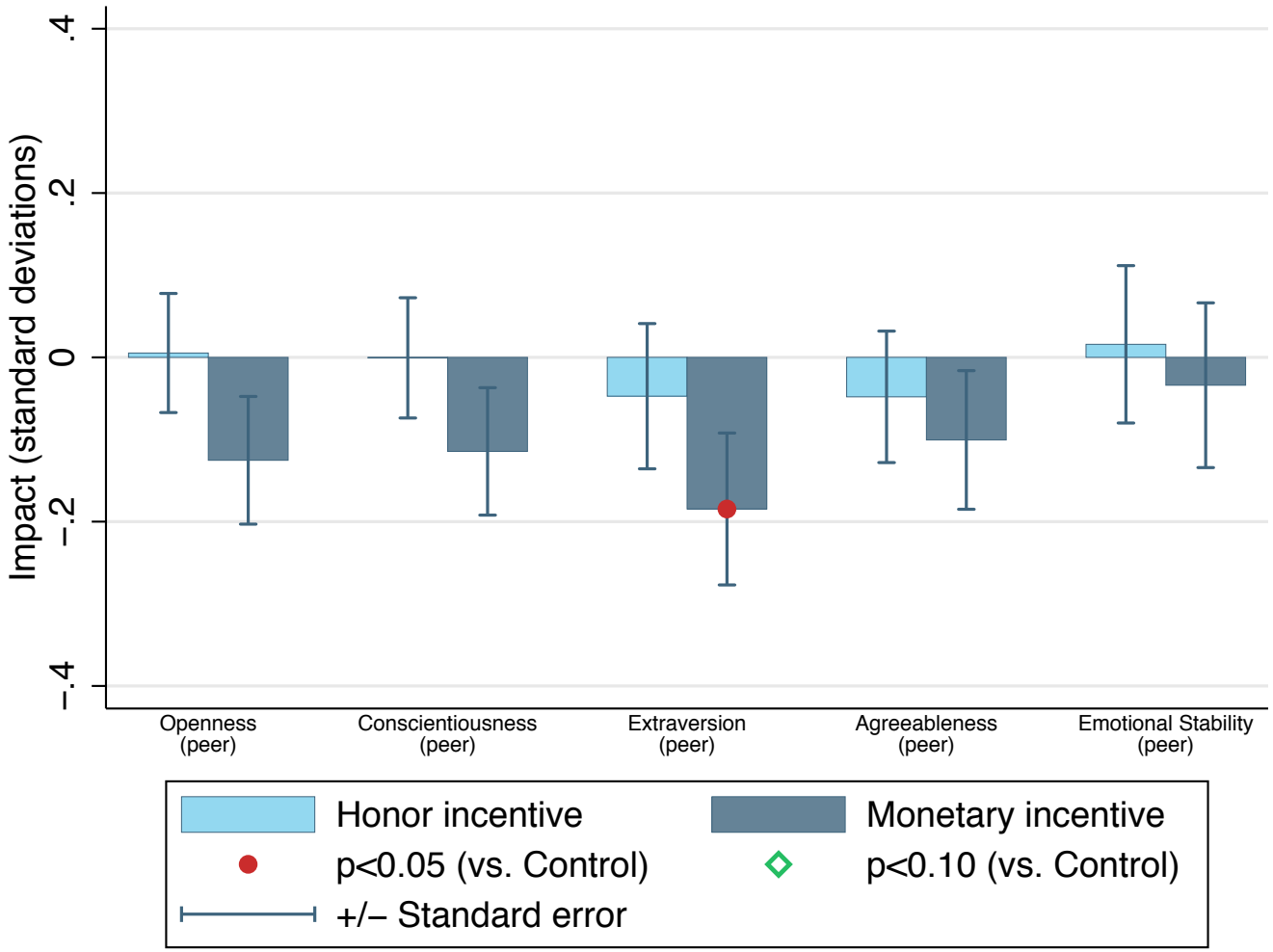
Female students had higher levels of Big Five personality



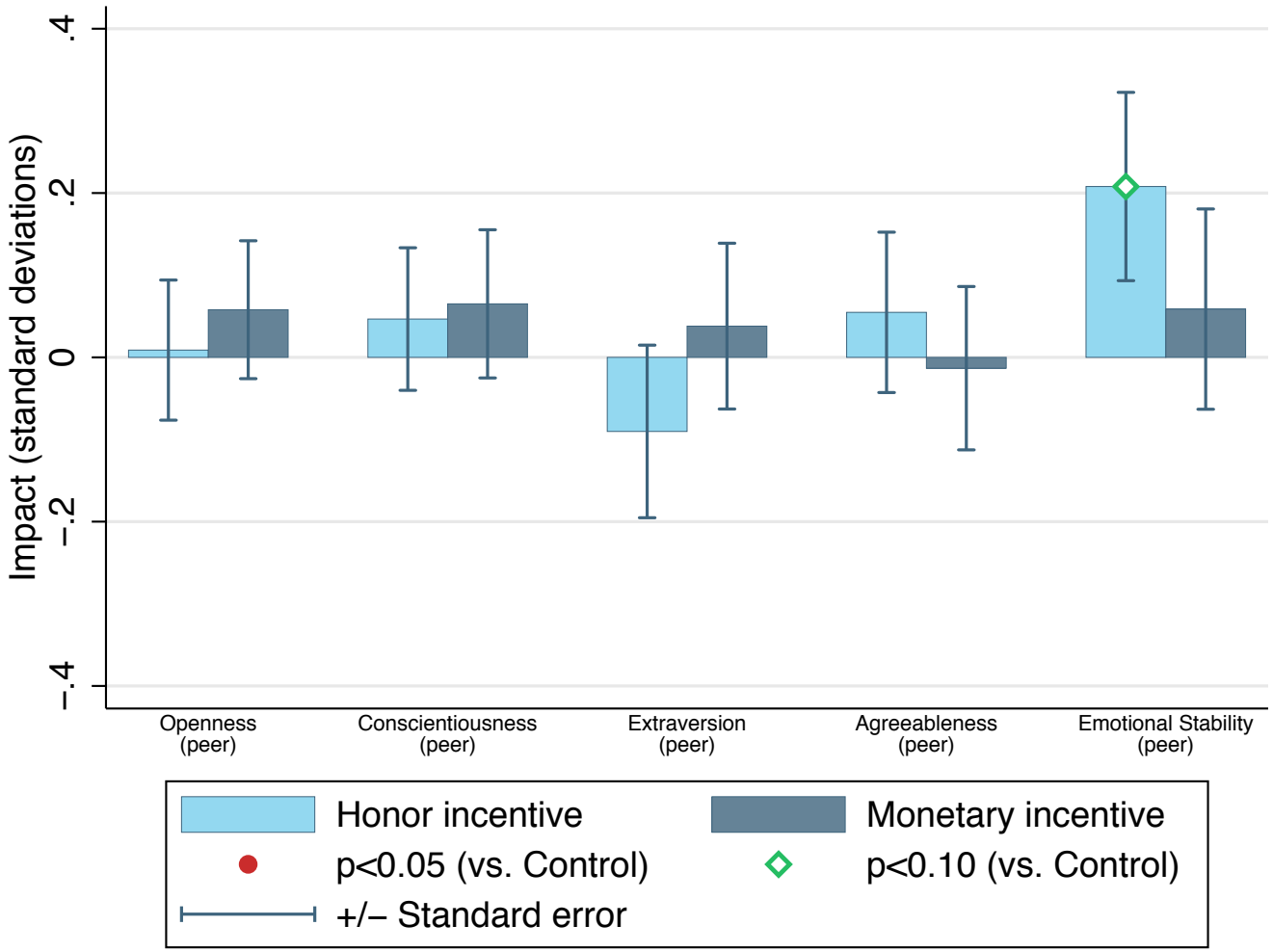
Female students had higher levels of other non-cognitive measures



Impact on how treatment groups are assessed by peers (Males)



Impact on how treatment groups are assessed by peers (Females)



Standard error inflation factor for treatment T (Cameron and Miller, 2015):

$$\tau_T \approx 1 + \rho_T \rho_\varepsilon \left(\bar{N}_s - 1 \right).$$

ρ_T: within-cluster correlation of T_{is}
ρ_ε: within-cluster correlation of ε
$\bar{N}_s$: average cluster size

Almlund, M., A. Duckworth, J. J. Heckman, and T. Kautz (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, and L. Wößmann (Eds.), *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Amsterdam: Elsevier.

Borghans, L., H. Meijers, and B. ter Weel (2008, January). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry 46*(1), 2–12.

Cameron, A. C. and D. L. Miller (2015). A practitioners guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Durlak, J. A., R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, and K. B. Schellinger (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development 82*(1), 405–432.

Kautz, T., J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. *OECD*.

Segal, C. (2012, August). Working when no one is watching: Motivation, test scores, and economic success. *Management Science 58*(8), 1438–1457.