4-22-2016

Measuring Teacher Conscientiousness and its Impact on Students: Insight from the Measures of Effective Teaching Longitudinal Database

Albert Cheng
Harvard University, axc070@uark.edu

Gema Zamarro
University of Arkansas, Fayetteville

UNIVERSITY OF ARKANSAS | College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

Measuring Teacher Conscientiousness and its Impact on Students: Insight from the Measures of Effective Teaching Longitudinal Database

Albert Cheng*
Gema Zamarro

April 2016

EDRE Working Paper 2016-05

char·asse·in
THE CHARACTER ASSESSMENT INITIATIVE

Measuring Teacher Conscientiousness and its Impact on Students: Insight from the Measures of Effective Teaching Longitudinal Database

Albert Cheng*
Gema Zamarro

University of Arkansas

April 2016

Abstract
Although research has been unable to find strong links between observable teacher characteristics and a teacher's ability to improve student achievement, it has generally not considered the role that teacher non-cognitive skills play in affecting student outcomes. In this article, we validate several novel performance-task measures of teacher conscientiousness based upon the effort that teachers exert completing a survey and use these measures to examine the role that teacher conscientiousness plays in affecting both student test scores and student non-cognitive skills. We conduct our analysis using the Measure of Effective Teaching Longitudinal Database where teachers were randomly assigned to their classrooms in the second year of the study. We exploit this random assignment to estimate causal impacts of teachers on their students' outcomes during the second year of the MET project. We find that our survey-effort measures of teacher conscientiousness capture important dimensions of teacher quality. More conscientious teachers are more effective at improving their student conscientiousness but not their student test scores. Additional analysis suggests that traditional measures of teacher quality largely fail to capture a teacher's ability to improve student conscientiousness, though measures of teacher quality based upon student ratings and one particular classroom observation protocol are exceptions.

Acknowledgements
We would like to thank Matthew Kraft, Kata Mihaly, and participants at the Association of Education Finance and Policy's 41st Annual Conference as well as Patrick Wolf for their comments on earlier versions of this paper.

*Corresponding Author. Email: axc070@uark.edu

1. Introduction

There is no doubt in the literature that teachers play an important role in improving student performance. Multiple researchers have shown wide variation in teacher effectiveness towards improving student achievement on standardized tests of math and reading (Rivkin, Hanushek, & Kain 2005; Rockoff, 2004). Teachers who are effective at improving student achievement have also been found to have impacts on their students' long-run life outcomes such as educational attainment and employment income (Chetty et al, 2014).

However, recent literature shows that teachers also affect other student outcomes besides test scores. A growing body of research documents the meaningful impacts that teachers have on their students' non-cognitive skills, which refer to personality and character attributes such as diligence, self-control, and a propensity to engage in prosocial behaviors.[1] Indeed, this is what we demonstrate in this analysis. Moreover, this research suggests that a teacher's impacts on student achievement are weakly correlated with the teacher's impacts on student non-cognitive skills. That is, teachers who are most effective at improving student test scores are not necessarily the most effective at improving student non-cognitive skills and vice-versa (Backes & Hansen, 2015; Blazar & Kraft, 2015; Cheng, 2015; Gershenson, forthcoming; Kraft & Grace, 2016; Jackson, 2012; Jennings & Diprete, 2010; Koedel, 2008; Ruzek et al., 2014). It appears that teacher quality is then multidimensional, comprising more than just the ability to improve student achievement.

It is, therefore, possible that some dimensions of teacher quality are left unmeasured if policymakers and practitioners solely rely on traditional value-added scores or other similar

[1] The term *non-cognitive* is intended to differentiate this set of attributes from *cognitive* skills, which consist of intelligence and content knowledge that standardized tests typically capture (West et al, 2016; Duckworth & Yeager, 2015).

measures of teacher quality that are derived from student achievement on standardized tests (Grissom, Loeb, Doss, 2016). Some schools and researchers have responded by engaging in efforts to develop alternative measures of teacher quality in an attempt to capture other aspects of teacher effectiveness and effective teaching practice. Some evaluators have asked students to rate teachers on surveys (Ferguson, 2012) and many others have relied on evaluations of teachers through formal classroom observations (Danielson, 2007; Pianta & Hamre, 2009). Notably, measures of teacher quality based upon these alternative measures are only modestly correlated with value-added scores, suggesting that different measures of teacher quality capture distinct dimensions of teacher effectiveness (Kane, McCaffrey, & Staiger, 2012).

Although it appears that teachers benefit their students in a variety of ways, it is not clear which types of teachers are most effective at improving particular student outcomes. Observable teacher characteristics, such as educational background and credentials, are largely uncorrelated with teacher value-added scores (Buddin & Zamarro, 2009; Goldhaber, 2008; Hanushek & Rivkin, 2006; Jacob, 2007). Similarly, the research identifying teacher impacts on student non-cognitive skills has generally not yet pinpointed the observable teacher characteristics that are associated with such impacts.

In this study, we aim to address this gap in the literature by investigating whether a teacher's level of conscientiousness is correlated with teacher effectiveness, particularly whether teachers with higher levels of conscientiousness produce improvements on levels of that same non-cognitive skill among students. Research has not thoroughly investigated this possibility because measures of conscientiousness and other non-cognitive skills are seldom available. One exception is Rockoff et al, (2011), who found among a sample of novice teachers that non-cognitive skills such as self-efficacy, extraversion and conscientiousness are marginally

correlated with impacts on student achievement and teacher retention (see also Duckworth et al., 2009). We aim to improve upon this observational work in two ways. First, we use data from the Measuring of Effective Teaching (MET) Project where teachers within the same school were randomly assigned to classrooms of students in the second year of the study. This feature allows us to account for the bias that arises from the systematic sorting of students to teachers (Rothstein, 2011) and then to estimate causal impacts of teachers on their students. Second, we rely upon new measures of conscientiousness based upon performance tasks. Following Hitt, Trivitt, and Cheng (in press) and Zamarro et al. (2016), we utilize a type of performance-task measure of conscientiousness that is based upon levels of engagement and effort that respondents exert in completing surveys. We henceforth refer to our measures as *survey-effort measures of conscientiousness*. These survey-effort measures of conscientiousness are readily available in our data and address common limitations of the self-reported measures of non-cognitive skills that Rockoff et al. (2011) used (e.g., social desirability bias, reference group bias, see Duckworth & Yeager [2015]). Furthermore, this same approach enables us to create measures of student conscientiousness by studying the effort that students themselves put in their surveys, expanding the range of student outcomes for study.

We first show that our survey-effort measures of teacher conscientiousness are correlated with some, though not all, traditional measures of teacher quality, including ratings based upon student surveys, formal classroom observation protocols, and subjective principal ratings. We then show that teachers with higher levels of conscientiousness are more effective at improving conscientiousness but not test scores among their students. These are important findings as prior research demonstrates the role that conscientiousness plays in influencing educational attainment, job performance, employment, income, health, criminal behavior, and other

indicators of well-being in life, above and beyond that of cognitive ability (Almlund et al., 2011; Dalal, 2005; Duckworth et al., 2007; Roberts et al., 2007).

Moreover, we find that not all traditional measures of teacher quality are correlated with impacts on student conscientiousness. Specifically, variation in value-added scores and ratings based upon some formal classroom observations does not explain variation in teacher impacts on student conscientiousness, but variation in other formal classroom observation protocols, principal ratings, and student ratings does.[2] We interpret these findings to suggest that teacher quality is multidimensional: Different teachers affect different student outcomes to varying degrees, and an assortment of teacher quality measures are needed to identify distinct effects on a set of relevant student outcomes. If policymakers and school leaders wish to improve the quality of the teacher workforce and student outcomes, they must think more clearly, holistically, and precisely about the multifaceted nature of teacher quality.

The remainder of the article is divided into four sections. We first review the relevant literature associated with teacher quality, emphasizing the ways in which it is currently measured and proposing that teacher non-cognitive skills could be relevant to the matter. We also provide background information related to our novel survey-effort measures of teacher conscientiousness. Next we describe our data, how we construct the variety of measures of teacher quality as well as conscientiousness, and our empirical strategy. We then present the results and, in the final section, discuss their implications.

2. Literature Review

Measuring Teacher Quality

[2]Using the same dataset as we do here, Kraft and Grace (2016) find a similar result based upon self-reports of grit as a student outcome measure.

Teachers play a crucial role in improving student short-run outcomes such as achievement as well as longer-run outcomes such as educational attainment and employment income (Chetty et al., 2014; Koedel, 2008; Rivkin, Hanushek, & Kain 2005; Rockoff, 2004). Despite the importance of identifying, hiring, and retaining effective teachers, policymakers and school leaders are largely unable to consistently do so.

One reason for this difficulty is that many observable characteristics such as educational background and certification are generally uncorrelated with a teacher's ability to raise student achievement on standardized tests (Buddin & Zamarro, 2009; Goldhaber, 2008; Hanushek & Rivkin, 2006; Jacob, 2007). School leaders cannot use such readily-available information to identify and to hire the most effective teachers. Although there is some evidence that teachers improve with more years of experience, school leaders cannot use this fact for hiring novice teachers. Moreover, the returns to experience appear to attenuate after three to five years (Clotfelter, Ladd, & Vigdor, 2006; Hanushek & Rivkin, 2006; but see Wiswall, 2013; Papay & Kraft, 2015).

Due to these limitations, other scholars have alternatively proposed to assess teacher quality based upon observing teacher effectiveness after they have started their career and then making personnel decisions given this new information (Podgursky, 2005; Kane et al., 2008). Typically, this includes estimating how much teachers improve their student test scores, but it is not always possible to calculate these value-added measures with validity, especially if multiple years of data for a teacher are not available (Rothstein, 2009; Koedel & Betts, 2011). Different approaches for isolating causal effects of teachers upon student achievement are available, though how suitable each approach is depends on a variety of contextual details (Gaurino, Reckase, & Wooldridge, 2014; Zamarro et al., 2015). This is not to mention that value-added

scores are sometimes impossible to calculate because students in many grades and subjects are not tested.

But even assuming that value-added scores are valid, approaches that assess teachers solely based upon their ability to improve test scores may fail to capture other important dimensions of teacher quality. Although raising test scores has been found to be important for improving long-run life prospects for students (Chetty et al., 2014), other factors may play a large and independent role in influencing the future well-being of students. A growing body of research in economics and psychology demonstrates the importance of non-cognitive skills in determining outcomes such as educational attainment, employment, income, health, and criminal behavior, even after controlling for performance on tests of cognitive ability (Almlund et al., 2014; Heckman, Stixrud, & Urzua, 2006). Insofar as value-added scores and similar measures based upon student test performance do not capture teacher effects on student non-cognitive skills, they will misstate the benefits that teachers provide to their students.

Alternatives for measuring teacher quality have been proposed, presumably because they may capture aspects of effective teaching that value-added scores do not. Formal classroom observations and student surveys of teacher performance, for example, potentially provide finer-grained contextual details about a teacher's classroom environment and instructional practices that may bear upon student outcomes (Pianta & Hamre, 2009). Curiously, however, these alternative measures have mainly been validated based upon how strongly they are correlated with student performance on standardized tests (Garret & Steinberg, 2015; Kane et al., 2012). The underlying assumption is that teacher quality is unidimensional and only concerns a teacher's ability to improve student achievement. That is, measures derived from classroom observations and student ratings of teachers are useful insofar as they more fully capture a

teacher's ability to improve test scores than value-added measures alone can. The practical implication, then, is to combine alternative measures with value-added scores to form a more valid and more reliable composite measure of teacher quality since each measure captures independent information about a teacher's ability to raise test scores (Kane et al., 2012; Mihaly et al, 2013).[3]

However, this approach could be misguided if teachers affect their students in meaningful and measurable ways that are not captured by test scores. Indeed, this is what the literature of teacher impacts on student non-cognitive skills suggests. Emerging research shows that teachers who have large effects on test scores do not necessarily have large effects on non-cognitive skills that, in turn, contribute to the future well-being of students. Similarly, teachers who have large effects on student non-cognitive skills do not necessarily have equally sizable effects on student test scores (Blazar & Kraft, 2015; Cheng, 2015; Gershenson, in press; Jackson, 2012; Jennings & Diprete, 2010; Koedel, 2008; Ruzek et al., 2014). Failing to consider a variety of student outcomes may lead to misclassifications of teacher effectiveness as teachers may benefit students on outcomes that are unconsidered or unobserved. Of particular relevance to our work is Kraft and Grace's (2016) analysis of the MET data; they find weak relationships between teacher effects on test scores and teacher effects on self-reports of non-cognitive skills like grit, growth mindset, and effort. In short, there are reasons to doubt the assumption that different measures of teacher quality collectively capture a unidimensional factor of teacher effectiveness. There is not

[3] As it turns out, student and formal classroom observation ratings are only modestly correlated with student achievement on standardized tests (Kane et al., 2012). In a separate study using data from the MET Project, Garrett and Steinberg (2015) find that teacher ratings based on classroom observations that used Danielson's (2007) Framework for Teaching are positively correlated with student test scores but much of this is due to the systematic sorting of higher-achieving students to teachers who have higher classroom observation ratings.

only variation between teachers in the ability to improve a specific student outcome but also variation within a teacher in his or her ability to improve a variety of student outcomes.

In related work, several scholars have found that subjective ratings of teachers given by principals are only moderately correlated with teacher value-added scores. Although these ratings are most strongly correlated with value-added scores among the least and the most effective teachers, they are unable to differentiate teachers within the middle of the distribution of value-added scores (Harris, & Sass, 2014; Jacob & Lefgren, 2008; Rockoff et al., 2012). Notably, Harris, Ingle, and Rutledge (2014) find that principals base their ratings not only upon a teacher's ability to improve test scores but also upon teacher non-cognitive skills, particularly the amount of effort they exert in their everyday work.

The Role of Teacher Non-cognitive Skills

Harris et al.'s (2014) finding that principals form judgments based upon teacher non-cognitive skills suggests that teacher non-cognitive skills may be a key component of teacher effectiveness. Indeed, research from labor economics and psychology demonstrates an association between worker productivity and certain non-cognitive skills. More conscientious workers, for example, exhibit better job productivity (Borghans et al., 2008; Dalal, 2005; Heckman et al., 2006; Roberts et al., 2007). Different combinations of non-cognitive skills may be required to improve worker productivity across different types of occupations (Borghans, ter Weel, & Weinberg, 2008). Although it is possible that more conscientious teachers are more highly-valued by principals, as Harris et al. (2014) suggest, it remains unclear how a teacher's level of conscientiousness relate to other measures of teacher quality and how it directly affects students.

There are only a few instances when scholars collected measures of teacher non-cognitive skills and studied their relationship to educational outcomes. For example, Duckworth et al. (2009), using self-reported measures of teacher non-cognitive skills, have found that teacher self-reports of grit and life satisfaction are predictive of student test scores among Teach for America teachers (Duckworth et al., 2009). In a sample of novice elementary and middle school math teachers, Rockoff et al. (2011) demonstrate that self-reports of conscientiousness, extraversion, and self-efficacy are correlated with subjective ratings that are given by their mentor teachers but are only weakly correlated with student achievement and teacher retention.

While Duckworth et al. (2009) and Rockoff et al. (2011) focus on how teacher non-cognitive skills affect student achievement, other work has focused on how teacher non-cognitive skills affect student non-cognitive skills. For example, Blazar and Kraft (2015) find that fourth- and fifth-grade students exhibit more self-efficacy when they have teachers who are more adept at lending emotional support to students in their interactions and through fostering a safe, positive classroom environment. Presumably, teachers who are more effective at improving student self-efficacy engage in certain classroom practices and processes that are conducive to realizing these outcomes (Pianta & Hamre, 2009). However, some research indicates that the pedagogical practices that teachers utilize and teacher observable characteristics are uncorrelated with impacts on student non-cognitive skills (Jennings & Diprete, 2010, but see Bargagliotti, Gottfried, & Guarino, 2016).

Elsewhere, Cheng (2015) uses longitudinal data to show that students receive increases in conscientiousness in years when they have teachers who exhibit higher levels of that same skill. He draws upon social learning theory to posit that students may learn non-cognitive skills through observing role models such as teachers (Bandura, 1977; Bandura & Walters, 1963). This

theory explains why teachers with a particular set of non-cognitive skills may be more effective at improving the same set of non-cognitive skills among their students. At the very least, students appear sensitive to and influenced by teacher behaviors. In fact, Blazar and Kraft (2015) posit that teachers who provide more social and emotional support improve student self-efficacy and happiness not because of a particular pedagogical approach but because of the behaviors that they model in providing such support.

We extend this line of research by further studying the relationship between teacher conscientiousness and other measures of teacher quality and their role in improving student cognitive and non-cognitive skills. Furthermore, instead of relying upon self-reported measures of conscientiousness, we rely upon newly developed survey-effort measures of conscientiousness, which we describe next.

Measuring of Non-cognitive Skills

The aim of developing and exploiting innovative measures of conscientiousness based upon survey effort is to capture levels of conscientiousness among teachers and students in cases where self-reported measures might not be available or might be affected by reporting biases. As an alternative to self-reported measures, non-cognitive skills data can be collected via performance-task measures. These types of measures begin by asking individuals to complete carefully designed task; researchers then observe variation in the individuals' behaviors as they complete it and interpret differences in behaviors as differences on the level of the skill being measured. Our novel survey-effort measures of conscientiousness are a type of performance-task measure and are constructed by observing how much effort teachers and students exert towards responding to a survey. Completing such clerical tasks requires careful attention to detail and persistence to avoid skipping or providing thoughtless, inaccurate answers (Hitt et al., in press;

Jackson et al., 2010). In other words, we view completing the survey itself as a task that requires conscientiousness and use three approaches to parameterize survey effort and to create measures of teacher conscientiousness: (a) item nonresponse rate, (b) careless answering patterns, and (c) survey omission.

Item nonresponse rate. Respondents sometimes demonstrate low effort in surveys by altogether skipping items or thoughtlessly providing answers of "I don't know." The item nonresponse rate is parameterized as the proportion of questions on a survey that an individual neglects to answer out of the total number of questions he was supposed to answer. Hitt et al. (in press) validate item nonresponse rate as a performance-task measure of non-cognitive skills related to conscientiousness. In six nationally-representative, longitudinal datasets of US secondary school students, item nonresponse rates are found to be predictive of educational attainment, which in turn influence labor market outcomes, in adulthood (see also Cheng, 2015). In our data, we use the item nonresponse rate among teachers as a measure of conscientiousness, provide validation that it captures meaningful teacher attributes, and explore whether it captures a meaningful dimension of teacher quality by investigating whether it is a determinant of student outcomes.

Careless answering patterns. When asked to complete surveys, some individuals begin the survey and do not skip items but still exert low effort by hastily providing thoughtless and random answers. This behavior results in careless answering patterns, which is a behavior that can be detected and parameterized (Hitt, 2015; Meade & Bartholomew, 2012). This measure has been validated as a proxy for conscientiousness in two nationally-representative samples — a sample of US adolescent school-age children and a nationally-representative sample of US adults — and found to be positively correlated with educational attainment, employment income, a

greater likelihood of being employed in a high-skilled job, and self-reported measures of conscientiousness, even after controlling for cognitive ability (Hitt, 2015; Zamarro et al., 2016). We describe the construction of this measure in greater detail in the methods section below.

Survey omission. Rather than skipping items or responding thoughtlessly, some individuals exhibit low survey effort by entirely ignoring a survey even after they are asked to complete it. Although there may be numerous reasons for why teachers, in particular, ignore a survey despite volunteering to participate in the study that requires its completion, there is no reason not to rule out survey omission as a manifestation of low survey effort. In other words, while some teachers exert low effort by beginning the survey but skipping items or providing inaccurate answers, others do not even begin the survey. Cheng (2015) uses a binary variable of whether a teacher completes or does not complete a survey as a measure of teacher conscientiousness, but he was not able to provide empirical validation for it because no data were available to conduct a validation test. Rockoff et al. (2011), however, shows that among new teachers who were invited to fill out a survey, those who did were rated as higher-quality teachers by their mentors than those who did not respond. This result suggests that, on balance, teachers who overlook surveys which they have been invited to complete may also tend to be of lower quality. Our data enable us to provide additional validation of survey omission as a meaningful measure of teacher quality. We further explore whether refraining from completing a survey that one volunteered to do through prior agreement is systematically related to other measures of teacher quality and predictive of student outcomes.

Why use performance-task measures? It is much more commonplace to rely upon self-reported measures of non-cognitive skills. Researchers usually execute this approach by administering surveys and using responses to a series of Likert-type items. Yet there are some

drawbacks; we highlight two. First, self-reported measures are relatively convenient to collect, but they are prone to social desirability and reference group bias (Duckworth & Yeager, 2015; Dobbie & Fryer, 2015; Paulhus 1991; West et al., 2016). In contrast, our survey-effort measures do not face the same threats to validity, especially because respondents typically do not know they are being observed on how much effort they exert to complete a survey. Their behavior while completing the survey then reveals something about their non-cognitive skills without being colored by sources of bias endemic to self-reported measures.[4] The second limitation of self-reported measures is more practical in nature. Self-reported measures of non-cognitive skills are rarely collected for teachers. This is why teacher quality research that uses large scale data sets rarely examines the topic of non-cognitive skills. Our data, which come from the MET study, is no exception. However, our survey-effort measures of teacher conscientiousness can be readily constructed from any dataset that has administered surveys to teachers. Latent information about a respondent's non-cognitive skills can be recovered using our approach within any data set, opening new avenues to research in this understudied area.

Research Questions

In summary, research has found that educational background, years of experience, and other observable teacher characteristics are weak predictors of teacher quality as measured by student achievement. There is likewise little understanding of how other observable teacher characteristics predict teacher impacts on student non-cognitive skills. After all, researchers have only recently begun to study teacher impacts on student non-cognitive skills, finding that the

[4] This critique is not intended to call all survey research into question. Self-reported measures of non-cognitive skills have been validated in a variety of circumstances and scholars have gleaned much knowledge from this research approach. That being said, there are incidences where sources of bias endemic to self-reported measures have possibly distorted results (Dobbie & Fryer, 2015; West et al., 2016). All approaches to measurement have unique strengths and weaknesses.

teachers who are effective at improving them do not necessarily comprise the same group of teachers who are effective at improving student achievement.

This literature, however, has not extensively studied the possibility that teacher non-cognitive skills explain variation in teacher effectiveness. Teacher non-cognitive skills could be related to existing measures of teacher quality and also play an important role in improving both cognitive and non-cognitive student outcomes. We address these issues in our study and contribute to the understanding of teacher quality by answering two specific research questions about teacher conscientiousness.

First, how are our survey-effort measures of teacher conscientiousness correlated with existing measures of teacher quality? Answering this question lends some validity to our measures by showing they are meaningful and systematically correlated to established measures of teacher quality. It also provides insight into what, exactly, current measures of teacher quality are actually capturing. Quite possibly, classroom observation protocols, student ratings, and value-added scores may not only provide an accurate account of the practices that teachers use for teaching and relating to students but also measure personality traits that are also crucial for student outcomes. In other words, we examine the extent, if any, that teacher conscientiousness is captured by value-added scores and ratings based upon classroom observations, subjective principal opinions, or student surveys of teacher performance.

Second, how are our survey-effort measures of teacher conscientiousness correlated with student cognitive and non-cognitive outcomes? We pay particular attention to whether more conscientious teachers are more effective at increasing student achievement and student conscientiousness based upon self-reported and survey-effort measures. If so, we may have uncovered an observable teacher characteristic that is linked to teacher effectiveness — a result

that has largely eluded researchers in teacher quality. As a point of comparison, we also examine the extent to which higher-performing teachers, as judged by traditional measures of teacher quality (e.g., student ratings, principal subjective ratings, formal classroom observations, and value-added scores) affect student conscientiousness as captured by our survey-effort measures. Kraft and Grace (2016) have already shown that student ratings of their teachers are positively correlated with teacher impacts on student grit, while traditional value-added scores and ratings based upon formal classroom observations are uncorrelated with them. Other prior research using MET data has already established the predictive power of teacher value-added scores and, to a much lesser degree, student ratings and formal classroom observations to forecast student achievement outcomes (Kane et al., 2012; Mihaly et al., 2013). Whether traditional measures of teacher quality also forecast survey-effort measures of conscientiousness is an unanswered question which we address.

Answers to these two questions provide a more refined picture of teacher quality by describing how teacher conscientiousness is linked to a variety of student outcomes. We describe our analytical methods next.

3. Methods

Data

Data for this study come from the MET Project. Six large, urban public school districts participated in the MET project, which lasted two school years from 2009-2011. The districts involved are New York City Department of Education, Charlotte-Mecklenburg Schools, Denver Public Schools, Memphis City Schools, Dallas Independent School District, and Hillsborough County Public Schools. In the second year of the study over 1,500 teachers from nearly 300 schools were randomly assigned within schools and grades to classrooms of students ranging

from fourth to ninth grade (White & Rowan, 2012). We leverage this random assignment to estimate the causal effect of teachers on a variety of student outcomes observed at the end of the second year.

Our measures of teacher quality are constructed based on data from the first year of the MET study.[5] In other words, we assume that teacher quality is captured with validity in the first year of the MET study, and then we estimate how teachers of different quality affect a variety of student outcomes measured in the second year. This strategy of predicting how teachers will affect student performance in a given year based upon performance in other years with a different group of students has been used in prior research (Chetty et al., 2014; Jackson, 2012). Note too that we assume that teacher conscientiousness is a stable trait – a claim that possesses some empirical evidence (Cobb-Clark & Schurer, 2012).

Survey-Effort Measures of Teacher Conscientiousness

We begin by describing how we construct our survey-effort measures of teacher conscientiousness: item nonresponse rate, careless answering patterns, and survey omission.

Item nonresponse rate. To compute item nonresponse rates for teachers, we use the Teacher Working Conditions Survey that teachers participating in the MET study completed during the first year of the study. Teachers were asked to answer 144 items on this survey, and

[5] Whenever possible, we prefer to build measures of teacher quality based upon data from the first year of the MET study. We refrain from using data from the second year of the MET study as it may confound the causal direction of the relationships between measures of teacher quality and student outcomes – the latter which are measured in the second year. That is, a teacher's randomly assigned classroom may influence the teacher's behavior and, ultimately, measures of teacher quality. In such a case it is unclear to what extent the teacher is affecting student outcomes or vice-versa. The only exception is that we use data from both years of the MET study to create teacher value-added scores since research has documented that value-added measures based on one year of data are unstable (McCaffrey et al., 2009). In contrast, other work suggests that other measures based upon classroom observations or student ratings possess greater within-year stability (Polikoff, 2015; Pianta et al., 2008). It is also worth mentioning that MET teachers did not receive their teacher quality ratings from the first year of the study so that their behaviors in the second year of the study are uninfluenced by any such feedback (Polikoff, 2015).

new teachers were asked to answer an additional 39 items. All surveys were administered through a confidential online system (Rowan & White 2012). Dividing the total number of questions that teachers did not answer by the total number of questions that they were supposed to answer yields our first survey-effort measure of teacher conscientiousness, item nonresponse rate. On average, teachers skipped or responded "Don't know" to 9 percent of the items.[6]

Careless answering patterns. Our second proxy of teacher conscientiousness identifies careless answering patterns to create a measure of survey effort. Methods for building this measure can be found in Hitt (2015) and Zamarro et al., (2016), which is a generalization of Meade & Bartholomew (2012). We provide a sketch of the method below. In the present study, teacher careless answering is derived from the Teacher Working Conditions Survey, which contains several scales that are designed to capture unique constructs such as time use, school leadership, the management of student conduct, and other aspects of the teacher's school and professional life.

We use response data from eight scales on the survey to run a series of bivariate regressions where the dependent variable is the response to an individual item in a given scale and the independent variable is an average of the responses to the remaining items in the scale. The residual in this regression captures the deviations between a teacher's actual response to an item and his expected response to the item based upon his as well as the sample's responses to the other items on the same scale. We obtain residuals from regressions for all items from each of the eight scales that we use. Then, we average the absolute values of the residuals within each scale and standardize each average to have a mean equal to 0 and standard deviation equal to 1.

[6]Not counting responses of "Don't know" as an instance of nonresponse does not substantively change the results, a pattern consistent with other research validating the use of item nonresponse (see Hitt et al., in press).

These values capture levels of careless answering within each of the eight scales. Higher values indicate more careless answering as larger residuals, in terms of absolute value, indicate greater deviations from expected responses. In other words, higher values on this measure indicate lower levels of conscientiousness. Finally, we compute an overall level of careless answering by averaging levels of careless answering from each of the eight scales and, once again, standardizing these values to have a mean equal to 0 and standard deviation equal to 1. Importantly, Cronbach's alpha for the scales we use range from 0.81 to 0.94. These figures suggest a high level of internal consistency within each scale and lend credence to our assumption that deviations in expected responses to an item are attributable to a lack of survey effort and not random measurement error in the scale.

As it turns out, the measures based upon careless answering patterns and those based upon item nonresponse are uncorrelated (ρ = -0.03). This low correlation does not necessarily imply that our measures are capturing distinct latent traits. It could simply indicate that different individuals exhibit low survey effort in distinct ways – either by skipping questions or by hastily providing thoughtless answers (Zamarro et al, 2016). We return to a discussion about this possibility in the final section of this article.

Survey omission. We also use the Teacher Working Conditions Survey to build our final survey-effort measure of conscientiousness: an indicator of survey omission. This variable takes on a value equal to 1 if a teacher who volunteered to be in the MET study never submitted nor even started the Teacher Working Conditions Survey as asked. This variable equals 0 for teachers who responded to the survey as asked. About 27 percent of teachers in our sample failed to respond to the Teacher Working Conditions Survey. Because teachers who did not begin the survey do not provide responses from which we could construct item nonresponse and careless

answering measures, we cannot report correlations between our measure of survey omission and our other two measures of teacher conscientiousness.

Traditional Measures of Teacher Quality

Having described the derivations of our survey-effort measures of teacher conscientiousness, we now describe the derivations of traditional measures of teacher quality, which include value-added scores, scores based upon formal classroom observation rubrics, student ratings of their teachers, and subjective ratings made by principals.

Value-Added Scores. We use two years of student test scores based upon state assessments to compute value-added scores for each teacher. Test scores are standardized by district, grade, subject, and year. Our value-added scores are computed by estimating models that include teacher fixed effects and then implementing an empirical Bayes adjustment to mitigate measurement error due to variation in the number of student observations that are available for computing each teacher's effect.

In particular, we estimate

$$Y_{ijt} = \alpha Y_{i(t-1)} + \boldsymbol{X_{it}\beta} + \delta \bar{Y}_{j(t-1)} + \bar{\boldsymbol{X}}_{\boldsymbol{jt}}\boldsymbol{\gamma} + \boldsymbol{\theta_i} + \epsilon_{it}. \quad (1)$$

In equation (1) Y_{it} is the test score for student *i* in classroom *j* during year *t*, while $Y_{i(t-1)}$ is the test score for student *i* in the prior year. $\boldsymbol{X_{it}}$ is a vector of demographic characteristics for student *i* including age and indicators for gender, race, free and reduced-priced lunch status, English language learner status, gifted status, and special education status. $\bar{Y}_{j(t-1)}$ and $\bar{\boldsymbol{X}}_{\boldsymbol{ijt}}$ represent measures of prior-year test scores and student demographic characteristics, respectively, averaged across all students in classroom *j*. $\boldsymbol{\theta_j}$ is a vector of teacher fixed effects and ϵ_{ijt} is the error term. Value-added scores for each teacher are computed by taking estimates of $\boldsymbol{\theta_j}$ and, following Tate (2004) applying an empirical Bayes adjustment. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the mean and

variance, respectively, of the estimated distribution of teacher value-added scores across the sample and let $\hat{\sigma}_j^2$ be the estimated variance of the estimated value-added score for teacher *j*. Bayes-adjusted teacher value-added scores, θ_j^{EB} are thus given by:

$$\theta_j^{EB} = \theta_j \lambda_j + \hat{\mu}(1 - \lambda_j), \text{ where } \lambda = \hat{\sigma}^2/(\hat{\sigma}^2 + \hat{\sigma}_j^2).$$[7]

Formal classroom observations. Researchers in the MET Project video-recorded multiple lessons for each classroom section that a participating teacher taught during both years of the study. These videos were then shown to evaluators trained in the use of one of two classroom observation rubrics. Some evaluators were trained to rate lessons based on the Classroom Assessment Scoring System (CLASS) developed by Pianta, La Paro, and Hamre (2008). The CLASS instrument is designed to capture the extent to which teachers support student learning and emotional growth through fostering a safe and positive classroom climate, managing classroom time and student behavior, engaging students, and using effective pedagogy. Other evaluators were trained to rate lessons based upon Danielson's (1996) Framework for Teaching rubric, hereafter FFT. This rubric is similar to CLASS as it is also designed to capture the extent to which teachers cultivate a classroom environment that is conducive to learning and whether they use effective instructional techniques that promote student learning. All videos were rated by these evaluators, and composite CLASS and FFT scores were created by averaging scores on the various components of each respective rubric. A teacher's overall classroom observation

[7]The MET study also contained pre-constructed value-added scores that could be used in the analysis. However, we preferred to construct our own scores to be sure of the empirical specification used. Correlations between our value-added scores and the pre-constructed value-added scores are 0.90 for math and 0.88. Replicating our analysis using the pre-constructed value-added scores rather than our data does not substantively change the results.

score, whether it is based upon the CLASS or the FFT instrument, is constructed by averaging his composite scores across all of his raters.[8]

Student perceptions. Students of teachers participating in the MET Project were annually administered the Tripod survey developed by Ferguson (2012). Based upon the student's responses, the Tripod survey captures seven dimensions of effective teaching. For example, the dimension named *Care* captures the extent to which teachers foster a sense of safety, belonging, and support in the classroom for their students. The extent to which teachers push students to work hard, exert greater effort to learn, and to think critically or deeply about a topic is captured by the dimension named *Challenge*. Other dimensions capture other instructional practices that support student engagement and learning. A teacher's overall Tripod score is created by averaging responses at the individual-student level and then averaging these scores again at the classroom-teacher level.

Principal Subjective Ratings. As part of the original MET study, principals in participating schools were asked to rate up to twelve teachers who were also a part of the MET study. Principals were asked to rate these teachers on a six-point ordinal scale, which included the following categories: Exceptional (top 5%), Very Good (top 25%), Good (top 50%), Fair (top 75%), Poor (bottom 25%), and Very Poor bottom 5%). In our model specifications, we dichotomize principal subjective ratings. The variable takes on a value equal to 1 for teachers receiving any of the highest three ratings (i.e., teachers whom principals deemed in the upper half of the effectiveness distribution); the variable takes on a value equal to 0 for teachers

[8]Based on the CLASS instrument a teacher received, on average, 19 ratings with a standard deviation of about 8 ratings. Based on the CLASS instrument and 5 ratings with a standard deviation of 2 ratings.

receiving any of the lowest three ratings as another.[9] We elected to dichotomize this variable to facilitate the interpretation of our results as well as our estimation techniques, which we now describe.

Empirical Strategy

Using the aforementioned survey-effort measures of teacher conscientiousness and traditional measures of teacher quality, we conduct a series of analyses to answer our two research questions. Recall that we first ask how our survey-effort measures of teacher conscientiousness are correlated with traditional measures of teacher quality. We then, pertaining to our second research question, examine whether our survey-effort measures are correlated with student cognitive and non-cognitive outcomes. That is, do more conscientious teachers affect students in different ways than less conscientious teachers?

Relationships between measures of teacher quality. To address our first research question, we investigate whether our three survey-effort measures of teacher conscientiousness are correlated with other measures of teacher quality. Although prior work has validated our survey effort measures as proxies for conscientiousness among adolescents and adults, including teachers (see Hitt, 2015; Hitt et al, in press; Rockoff et al., 2011; Zamarro et al., 2016), this analysis provides additional evidence of whether our behavioral measures are merely random noise or actually meaningful signals of teacher quality. If our survey-effort measures of conscientiousness are indeed meaningful signals, this analysis provides a sense of what, exactly, our measures capture with respect to other widely-used measures of teacher quality.

[9] Interestingly, this categorization approximately divided the teachers into equal halves along the distribution of principal ratings. There does not appear to be evidence of a "Lake Wobegon effect" where virtually all teachers receive higher ratings (Donaldson, 2009).

In this analysis, we run a series of bivariate regressions where the dependent variable is a traditional measure of teacher quality (e.g., teacher value-added scores, classroom observation scores, student ratings, principal ratings) and the independent variable is one of our survey-effort measures of teacher conscientiousness (e.g., item nonresponse rate, survey omission, careless answering patterns). We express all variables, except our dichotomous indicators of survey omission and principal ratings, in terms of standard deviations for ease of interpretation.

Teacher impacts on student outcomes. For our second analysis, we examine whether each survey-effort measure of teacher conscientiousness is predictive of student outcomes. Specifically, we consider teacher impacts on a variety of student cognitive and noncognitive outcomes: (a) test scores in math and reading, (b) a student self-reported measure of grit, (c) a student self-reported measure of effort, (d) student item nonresponse, and (e) student careless answering patterns on a survey. We discuss each of these measures in turn.

Test scores in math and reading are based upon state-mandated assessments administered during the second year of the MET study. All test scores are standardized by district, year, and grade to have a mean and standard deviation equal to 0 and 1, respectively.

During each school year, students also completed the Student Perceptions Survey administered by MET researchers. This survey included the Tripod instrument as well as items soliciting basic demographic information. In the second year of the MET study, the Student Perceptions Survey included the Duckworth and Quinn (2009) Grit Scale and several items designed to measure the amount of effort that the student exerts in class.

Duckworth and Quinn define grit as “perseverance and passion for long-term goals” and have found it to be positively correlated with academic outcomes such as retention and grade point average in postsecondary students (p. 172). For our data, we create scale scores by

averaging each student's responses to the eight Likert-type items on the Grit Scale, reverse coding items when necessary. Scale scores range from 1 to 5, with higher values indicating higher levels of grit, and have a mean of 3.56 with a standard deviation of 0.67 in our student sample.

Items for the self-reported measure of student effort are shown in Appendix A. The scale consists of five and four Likert-type statements for secondary- and primary-school students, respectively, to answer. We reverse-coded items as necessary and then averaged the responses. Higher values indicate that the student exerts more effort in class. Scale scores range from 1 to 5 with an average of 4.07 and a standard deviation of 0.67.

The remaining measures of student non-cognitive skills are two survey-effort measures of student conscientiousness – item nonresponse and careless answering patterns.[10] These measures are based upon student effort on the Student Perceptions Survey administered in the second year of the MET study. Elementary school students were asked to complete 75 items, while secondary school students were asked to complete 83 items. The average item nonresponse rate was 2.7 percent with a standard deviation of 9.0 percent. Values for the careless answering measure are, by construction, standardized to have a mean equal to 0 and standard deviation equal to 1. We build the measure of student careless answering based upon six of the seven subscales in the Tripod Survey administered in the second year of the MET study. We use the Care, Control, Clarify, Challenge, Captivate, and Confer subscales.[11] These scales have Cronbach's alphas that

[10]There is no measure of survey omission for students as they were compelled to complete the survey once consenting to participate. Teachers, in contrast, consented to the study but could freely decide whether or not to comply with the study by completing their respective survey.

[11] Ferguson's (2012) Tripod instrument, which is a measure of teacher quality, is also included on the Student Perceptions Survey from which we derive several student outcome variables such as self-reported grit, self-reported effort, item nonresponse, and careless answering. However, it is important to reemphasize that measures of teacher quality are all based upon surveys administered in the first year of the study, while student outcome measures come from surveys administered in the second year of the study. Thus, student item nonresponse or careless answering on

range from 0.68 to 0.85 for secondary school students and 0.63 to 0.84 for elementary school students. The Tripod survey has a seventh subscale named *Consolidate* but we opted to omit it due to its apparent lack of reliability as indicated by a low Cronbach's alpha value ($\alpha = 0.52$).[12]

Table 1 displays the correlations between each student outcome measure. One can easily observe that measures of student non-cognitive skills are, if anything, modestly correlated with student test scores. The magnitudes of the correlation coefficients never surpass 0.25. Interestingly, the two self-reported measures of student non-cognitive skills (i.e., grit and effort) are more strongly correlated with each other than they are to test scores or survey-effort measures of conscientiousness. Meanwhile, item nonresponse appears uncorrelated with all other measures. As it was the case for teachers, there is essentially no association between item nonresponse and careless answering among students. This is expected if leaving answers blank in the survey or carelessly answering are substitutive strategies for low effort. On the other hand, careless answering seems to be equally correlated with test scores and self-reported measures of non-cognitive skills, albeit modestly.[13]

---

the Tripod survey during the second year of the MET study does not distort measures of teacher quality, which are based upon the Tripod survey administered in the first year of the study. Moreover, students only enter our data if they have a teacher participating in MET, and because students in the data typically do not have the same teacher for two consecutive years, our measures are not affected by same-source bias. That is, responses on the Tripod instrument used to construct measures of teacher quality are not provided by the same students who provide responses from which we build our student outcome measures.

[12] We are unable to examine whether our survey-effort measures of student conscientiousness are predictive of later-life outcomes in the MET data as such data are unavailable. Instead, we rely on other research that has documented external validity of these measures. Lower item nonresponse rates and lower levels of careless answering as measured in adolescence have been found to be associated with greater levels of educational attainment and a greater likelihood of employment when measured in adulthood, even after controlling for cognitive ability as measured by standardized test scores (Cheng, 2015; Hitt et al., forthcoming; Hitt, 2015). Moreover, other studies of schoolchildren demonstrate that conscientiousness is associated with academic and labor-market success (Duckworth et al., 2007; MacCann, Duckworth & Roberts 2009; Poropat, 2009; Trautwein et al. 2006; Tsukayama, Duckworth & Kim 2013; Kraft & Grace, 2016; Roberts et al., 2007).

[13] It is interesting that correlations between self-reported measures of noncognitive skills are stronger than correlations between self-reported and performance-task measures. This is a pattern which other work has also found (Zamarro et al., 2016). The higher correlations among self-reported measures could be driven by the common

≪Table 1 Here≫

As introduced above, in this paper, we leverage the random assignment of teachers to classrooms in the second year of the MET Project to estimate causal impacts of teachers, who vary in our survey-effort measures of conscientiousness, on each of the student outcomes. We compute both intent-to-treat (ITT) estimates and local average treatment effect (LATE) estimates using an instrumental variables (IV) strategy.

For the ITT analysis, we use ordinary least squares to estimate models of the form

$$Y_i = \beta_0 + \beta_1 T_i + \boldsymbol{\beta_2 X_i} + \boldsymbol{\phi_i} + \epsilon_i, \quad (2)$$

where Y_i is one of the outcomes of interest for student *i* measured in the second year of the MET study, $\mathbf{X_i}$ is a vector of student demographic characteristics as in equation (1) but now also includes prior-year test scores in math and English, and ϵ_i is the usual error term but clustered at the classroom level. T_i is the independent variable of interest and represents one of our survey-effort measures of conscientiousness for the teacher to whom student *i* was randomly assigned. Again, these measures of teacher characteristics are obtained based upon data in the prior (first) year of the MET study. The associated coefficient, β_1, captures the magnitude of the effect that teachers with varying levels of conscientiousness have on a particular student outcome. To better capture the experimental design and randomization process, we include randomization block fixed effects, $\boldsymbol{\phi_i}$. In the MET study, teachers teaching in the same school, grade, and subject were first placed into blocks and then randomized to classrooms (Rowan & White, 2012). Note, then,

mode in which the measures are collected, namely, through self-reports. Whether this is evidence of bias that is common across self-reported measures is unclear. Moreover, magnitudes partial correlations between survey-effort and self-reported measures of conscientiousness in the MET data are similar to those found in other work (Zamarro et al., 2016; Galla et al., 2014).

that the randomization block fixed effect also controls for unobserved characteristics at the school by grade by subject level.

We employ a two-stage least squares approach to compute our IV estimates. In this framework, we use survey-effort measures of teacher conscientiousness for each student's randomly assigned teacher as an instrument for the same measure of teacher conscientiousness for the student's actual teacher. Again, teacher characteristics are measured based on data from the first year of the MET study. In particular, we estimate the following two-stage model:

$$T_i^A = \gamma_0 + \gamma_1 T_i^R + \boldsymbol{\gamma_2 X_i} + \boldsymbol{\phi_i} + \mu_i \quad (3)$$

$$Y_i = \beta_0 + \beta_1 \hat{T}_i^A + \boldsymbol{\beta_2 X_i} + \boldsymbol{\phi_i} + \epsilon_i, \quad (4)$$

where T_i^R is a measure of teacher conscientiousness for the teacher to whom student *i* was randomly assigned and T_i^A is the corresponding measure of teacher conscientiousness for the teacher whom student *i* actually had during the school year[14], and $\hat{T}_i^A$ represents fitted values of T_i^A based upon estimations of Equation 3. The other variables correspond to those presented in equation (2).[15] These models, together with the random assignment of teachers to students, provide causal estimates of how student outcomes are altered when they have teachers with varying levels of conscientiousness. Again, standard errors are clustered at the classroom level.

[14] MET data that was made available to us contained multiple teacher identification numbers for each student: (a) a randomly assigned teacher identifier, (b) an actual teacher in October identifier, (c) an actual teacher in May identifier, and (d) a global teacher identifier. However, the global teacher identifier and the identifiers for actual teachers in October and May were not always consistent, so it was sometimes unclear which variable represented the most accurate information of the teacher a student actually had. In order to test for the effect of these data inconsistencies on our results, we ran our analysis using each of these identifiers and found that our results are not sensitive to the choice of identifier. Thus, we decided to present results for the specification that uses the identifier for the student's actual teacher in October.

[15] Strictly speaking, it is not absolutely necessary to include variables to control for prior-year measures of our student outcome variables because we rely on the random assignment of teachers to students. Including them, however, could be useful to improve the precision of our estimated coefficients. However, doing so is not possible in the data since we do not observe all students in both years of the MET study. Students are only part of the MET study during years where they have MET teachers. The only exception to this rule is the inclusion of prior-year test scores for all students because they are provided in the MET data.

As an additional analysis, we estimate the same models substituting our survey-based measures of teacher conscientiousness with traditional measures of teacher quality (e.g., value-added scores, student ratings, scores on formal classroom observations, principal subjective ratings[16]). These models provide a point of comparison for our survey-effort measures. In other words, they reveal how well the survey-effort measures of teacher conscientiousness predict and explain variation in student outcomes relative to traditional measures of teacher quality, further honing similarities and distinctions between each type of measure of teacher quality. At the very least, we would like to examine the ability of traditional measures of teacher quality based upon data from the first year of the MET study to predict survey-effort measures of student conscientiousness in the subsequent year.

4. Results

Relationships between Measures of Teacher Quality

We begin by presenting relationships between our survey-effort measures of teacher conscientiousness with traditional measures of teacher quality. Table 2 lists coefficient estimates and standard errors from bivariate regressions where our behavioral measures of teacher quality are independent variables and traditional measures of teacher quality are dependent variables. In general, our behavioral measures appear to be uncorrelated with value-added scores but are correlated to other indicators of teacher quality such as those based upon formal classroom

[16]Although the principal subjective ratings variable is binary, we use linear probability models in the first stage of the IV models to predict ratings that principals gave to students' actual teachers. Instead of a linear probability model, we also ran specifications where we estimate probit models in the first stage. Results whether we use linear probability models or probit models do not substantively differ. Moreover, an ITT analysis using a nonlinear specification that include dummies for each of the six principal ratings yielded findings similar to those presently shown in the results section.

observations, student ratings, and principal ratings. Teachers who are less conscientious as captured by our survey-effort measures have worse scores on other measures of teacher quality.

≪Table 2≫

For instance, a one standard deviation increase in the item nonresponse rate (i.e., lower conscientiousness) is associated with between a 0.05 to 0.10 standard-deviation decrease in FFT scores, CLASS scores, and Tripod ratings. Similarly, we observe that teachers who more extensively engage in careless answering appear to have lower FFT scores. A one standard deviation increase in careless answering is associated with a decrease in FFT scores by 0.05 standard deviations, though the result is only significant at the 90 percent confidence level ($p = 0.06$).

We finally consider relationships between survey-effort measures of conscientiousness and our indicator for whether a teacher overlooks the Teacher Working Conditions Survey. As shown in the last row of Table 2, teachers who fail to begin the surveys, on average, have FFT and CLASS scores that are lower than teachers who complete the survey. The differences in scores based upon these observation protocols are approximately 0.14 and 0.19 standard deviations, respectively. Moreover, teachers who fail to complete the survey are about 5 percent more likely to receive one of the three lower subjective ratings from their principals rather than one of the three higher ratings. However, there are no discernable differences in value-added scores between teachers who do or do not begin the survey.

Teacher Impacts on Student Non-cognitive Skills

Turning attention to our second research question, we find that our survey-effort measures of teacher conscientiousness are important determinants of certain student outcomes but not others. Again, we leverage the randomized assignment of teachers to students to estimate

the causal effects that teachers with varying levels of conscientiousness have on student non-cognitive skills and test scores.

Post-Randomization and Post-Attrition Covariate Balance. Before presenting results, we check if the randomization process yielded covariate balance. If so, this would lend credence to the claim that the randomization of teachers to students was properly executed so as to remove the bias in our estimates that are attributable to systematic sorting of students to teachers. As can be expected, a substantial amount of noncompliance with random assignment occurred in the execution of the original MET study (Kane et al., 2013; Rowan & White, 2015). The ITT and LATE estimates, described above, would address bias issues due to noncompliance if all no-compliers continue to stay in our sample. However, students attrite from the sample if they transferred districts, switched to non-participating schools in a participating district, or remained at the same participating school but transferred to a classroom taught by a teacher who was not participating in the MET study. Outcomes for non-complying students are not observed in our data. For this reason, our checks for post-randomization covariate balance are based upon data that has been collected post-attrition. Detecting covariate balance in this data would lend more credence that, despite attrition, our estimates can legitimately be interpreted as causal.

We check for post-attrition, post-randomization covariate balance in two ways. First, we test to see if classrooms within randomizations blocks differed along student demographic characteristics and student prior-year test scores. To do this, we first demean each student characteristic or prior-year test score by randomization blocks. We then regress demeaned values of a particular student characteristic or a prior-year test score on a vector of dummy variables that indicate a student's randomly assigned teacher. Finally, we conduct an F-test to see if we can reject the null hypothesis that the estimated coefficients on the set of teacher dummy variables

are jointly equal to each other. Failing to reject the null hypothesis would provide evidence that classrooms were balanced across student covariates within randomization blocks and suggest that the randomization process occurred appropriately so as to eliminate the systematic sorting of students to teachers. Results for these estimates are shown in Appendix Table B1. Although we do not find evidence of imbalance along a majority of student characteristics across classrooms within the same randomization blocks, there is some evidence of disparity. For instance, English language learners, gifted students, and students with higher test scores are more likely to be assigned to particular teachers within the same randomization blocks. To address this issue in our empirical strategy, we include controls for a wide range of classroom-level characteristics, including but not limited to the average prior-year test scores and proportions of students who are classified as gifted or an English-language learners.

As a second test for covariate balance, we run a series of bivariate regressions where we use each observable student demographic characteristic to predict each measure of teacher quality from a student's randomly-assigned teacher. If the randomization process was implemented appropriately, we should not observe any explanatory variables attaining statistical significance in these regressions. Results are shown in Table B1 in Appendix B. Though we find statistically significant differences in student characteristics in a few of our regressions, we cannot clearly rule out the possibility that these have occurred by chance given the number of statistical tests that were conducted. More precisely, we find 8 statistically significant covariates out of 72 tests, when one would expect about 7 to happen just by chance (assuming a significance level of $\alpha = 0.1$). Again, we control for all available observable individual-level and classroom-level demographic characteristics and prior-year test scores in our analyses to control for any source of bias due to these observable characteristics.

Ultimately, given the inclusions of this wide range of covariates together with evidence of post-randomization and post-attrition balance across a majority of student covariates, we maintain that we have sufficiently eliminated systematic sorting of students to teachers. In other words, we have strong reason to believe that our estimates are both causal and valid.

Teacher Noncognitive Skills and Student Noncognitive Skills. We first present ITT estimates in Table 3. Each panel in Table 3 displays coefficients and standard errors that are estimated from Equation 2 for one of our three survey-effort measures of teacher conscientiousness and each student outcome. For example, the first panel under column 5 suggests student item nonresponse rate increases by about 2.4 percent of a standard deviation when a student is randomly assigned to a teacher who, all else equal, is one standard deviation higher on the distribution of teacher item nonresponse.

In fact, we find that all three survey-effort measures of teacher conscientiousness are predictive of student item nonresponse rates (Column 5) but are not predictive of student achievement in math (Column 1) or English (Column 2), student self-reported effort (Column 4), and careless answering (Column 6). As shown in column 5, effect sizes range from 2 to 6 percent of a standard deviation in item nonresponse. Focusing on results for grit in Column 3, we find that students self-report lower levels of grit when they are randomly assigned to teachers who fail to start the Teacher Working Conditions survey. In other words, students randomly assigned to a less conscientious teacher appear to become less conscientious, primarily according to the item-nonresponse proxy and self-reported grit measures. In contrast, student test scores do not seem to be affected when they are randomly assigned to teachers of varying levels of conscientiousness.

≪Table 3≫

Results based upon the IV estimates are shown in Table 4 and generally comport with results based upon the ITT analysis. As shown in the second panel of Column 5, students experience increases in item nonresponse rates when they have teachers who exhibit more carelessness while answering surveys. In the third panel under Column 5, one can also see that relative to students who have teachers that complete the Teacher Working Conditions Survey, students with teachers who do not complete the Teacher Working Conditions Survey have item nonresponse rates that are approximately 0.14 standard deviations higher, all else being equal. In contrast, the associations between (a) student item nonresponse with teacher item nonresponse and (b) student self-reported grit and teacher survey omission, which were significant under the ITT models, are no longer statistically significant under the IV models. The coefficient estimates are positive and larger than the ITT estimates but imprecisely estimated. Furthermore, the IV estimates indicate that students self-report less effort in their classes if they are taught by teachers who refrain from responding to the Teacher Working Conditions Survey. Relative to students with teachers who do indeed respond to the survey, these students' self-reported effort ratings are about 0.15 standard deviations lower.

≪Table 4≫

Traditional Measures of Teacher Quality and Student Non-cognitive Skills. To shed additional light onto our survey-effort measures of teacher conscientiousness, we now present estimates of the relationships between student non-cognitive skills and traditional measures of teacher quality. ITT results are shown in Table 5 and are displayed in a fashion analogous to Table 3 but with traditional measures of teacher quality rather than our survey-effort measures of teacher conscientiousness. Findings shown in the first panel demonstrate that teachers who have been rated more highly by their students on the Tripod survey during the first year of the MET

study are more effective at improving conscientiousness in their subsequent set of students during the second year of the MET study. Students who are randomly assigned to a teacher whose Tripod rating is one standard deviation higher have self-reported grit scores that are about 0.03 standard deviations higher, self-reported effort scores that are about 0.05 standard deviations higher, and careless answering scores that are 0.09 standard deviations lower. Higher quality teachers as judged by student ratings do not appear to have an effect on student item response rates.

In the second panel, one can also observe that students who are randomly assigned to teachers who receive higher principal ratings exhibit less careless answering. That is, these students appear more conscientiousness. Relative to students assigned to teachers who received one of the three lower categories of principal ratings, students assigned to teachers who received one of the three higher categories of principal ratings have careless answering scores that are almost 10 percent of a standard deviation lower.

Turning to the third and fourth panels, variation in teacher quality as measured by classroom observations protocols sometimes explains variation in teacher effectiveness at improving student non-cognitive skills. While FFT ratings are uncorrelated with such impacts, CLASS ratings are modestly predictive of student grit and item nonresponse but slightly more strongly predictive of careless answering. A one standard deviation increase in a teacher's rating based upon the CLASS protocol is associated with an increase in grit and item nonresponse by about 0.03 standard deviations. The corresponding effect size for careless answering is 0.04 standard deviations.

Finally, we observe in the last two panels that that teacher value-added scores show no association with impacts on student non-cognitive skills.

≪Table 5≫

Corresponding estimates based upon IV techniques are displayed in Table 6. These results are generally consistent with ITT results in Table 5, where teachers who are rated higher by their students are more effective at improving student non-cognitive skills as measured by student self-reported grit, self-reported effort, and careless answering. Likewise, students exhibit less careless answering when they have teachers who receive higher ratings from their principals or have higher ratings on the CLASS protocol. There are, however, a few differences between the ITT and IV results that are worth highlighting. Teachers rated more highly on the CLASS protocol now do not appear more effective at improving student self-reported grit and item nonresponse as the ITT results demonstrated. Notably, these relationships were only significant at the 90 percent confidence level in the ITT models. Lastly, students taught by higher-performing English teachers as measured by value-added scores also appear to experience decreases in item nonresponse rate, though the result is only significant at the 90 percent confidence level.

≪Table 6≫

5. Discussion and Conclusion

Summary

In this article, we have aimed to measure teachers' levels of conscientiousness and to assess how they affect similar non-cognitive skills in students. Little research has investigated the role that a teacher's conscientiousness plays in educational outcomes (Duckworth et al., 2009; Rockoff et al., 2011). The paucity of research is attributable to the fact that teacher non-cognitive skills data are rarely collected and available. We overcome this data limitation by utilizing three innovative survey-effort-based measures of conscientiousness that can be

constructed using data collected from teacher surveys: (a) item nonresponse, (b) careless answering, and (c) survey omission. Based upon conceptual reasons from survey methods research and empirical evidence, we view these behavioral measures as proxies for non-cognitive skills related to conscientiousness (Hitt et al., in press; Hitt, 2015; Krosnick, 1991; Smith, 1982; Zamarro et al., 2016). Furthermore, we build survey-effort measures of conscientiousness for students, which is typically unavailable in many datasets, to expand the range of student outcomes available for analysis, although in this case we did have some self-reported measures of non-cognitive skills for students.

We summarize our three main findings. First, our survey-effort measures of teacher conscientiousness are correlated with some but not all existing measures of teacher quality, namely those based upon formal classroom observations, student ratings, and principal ratings but not value-added scores. This result provides additional validation for the use of item nonresponse, careless answering, and survey omission as proxies for meaningful teacher characteristics. These survey-effort measures may represent important observable teacher characteristics that can be easily obtained from existing data and used for further research on teacher quality.

Specifically, we maintain that we have captured levels of teacher conscientiousness with our survey-effort measures, as other work suggests (Hitt, 2015; Hitt et al, in press, Zamarro et al., 2016). In fact, findings from prior research suggest that ratings based upon principals, students, or other classroom observations may capture teacher conscientiousness. For instance, Rockoff et al. (2011) find that novice teachers who have higher self-reported levels of conscientiousness also received higher subjective ratings from their mentor teachers. Interestingly, Rockoff et al. additionally found that novice teachers who did not complete their survey were rated lower by

the mentor teacher than those who completed the survey. The intention of this comparison was not a test of teacher conscientiousness but a robustness check to address the issue of missing data. It is possible that these researchers, by happenstance, actually uncovered more evidence that novice teachers with lower conscientiousness receive lower subjective ratings. The proposition that ratings based upon observations of teachers capture personality traits has also been raised by Harris et al. (2014) and is worth further investigation. Certainly, it would be valuable for scholars and practitioners to reflect upon what, exactly, ratings by principals, students, and other observers of teachers in their classrooms actually capture.

Second, we find that teachers who exhibit more conscientiousness as measured by our survey-effort measures are more effective at improving student conscientiousness. Yet we do not find evidence that teacher conscientiousness is tied to student achievement as measured by test scores. Leaning upon the random assignment of teachers to students in our data, we interpret our results as causal. Students exhibit higher item nonresponse rates but not higher test scores when they have teachers who exhibit higher item nonresponse rates[17], greater levels of careless answering, or fail to complete a survey when asked to do so. There is also some suggestive evidence that students self-report lower levels of grit when they are assigned to teachers who do not complete surveys and self-report lower levels of effort when they are actually taught by those teachers.

Third, we do not find much evidence that teachers with high formal classroom observation or value-added scores are effective at improving student conscientiousness. The exception to this result is that higher-quality teachers as measured by the CLASS observation

[17] Again this result was only significant at the 90 percent confidence level in the ITT analysis and imprecisely estimated in the IV analysis.

protocol appear to be more effective at improving student conscientiousness as measured by careless answering. However, the traditional measures of teacher quality that is most strongly predictive of student conscientiousness are student ratings of teachers — similar to what Kraft and Grace (2016) have found — and principal subjective ratings.[18] Taken together, these results indicate that existing measures of teacher quality do not fully capture all the relevant ways in which teachers influence their students. In particular, our survey-effort measures of conscientiousness capture impacts that teachers have upon student item nonresponse — something traditional measures of teacher quality are, according to our analysis, less able to do.

Overall, we interpret our estimates as lower bounds for the impacts that teachers have upon student conscientiousness. Arguably, each of our survey-effort measures capture latent traits that are only proxies for conscientiousness. For instance, both conscientiousness and problems with survey administration could explain variation in our measure of survey omission. A variety of logistical reasons may influence why some teachers did not begin the survey. Likewise, variation in careless answering could partially reflect genuine variation in teacher response patterns or natural measurement error in the scales (even though we already selected scales with higher levels of Cronbach's alpha) in addition to a lack of conscientiousness. All of these issues generate random measurement error and, at worst, cause our results to attenuate. Furthermore, while student test scores are arguably mostly influenced by one teacher in a particular content area, student non-cognitive skills can be influenced by multiple teachers. For secondary students that have more than one teacher, this possibility introduces additional

[18] The FFT, CLASS, and Tripod instruments also have a variety of subscales designed to capture different dimensions of teacher quality. We estimate models using scores on these subscales in addition to models that use scores based on the entire scale. Results when using separate scores from the seven subscales of the Tripod instrument all reflect overall results. Results when using separate scores from subscales of the FFT and CLASS were mixed with no obvious patterns.

attenuation bias in our results. For these reasons, we interpret our results to be conservative estimates. That we find systematic relationships between survey-effort measures of teacher conscientiousness and student outcomes, therefore, deserves special attention..

Although we cannot provide evidence for the mechanisms for why these relationships exist, these findings are consistent with social learning theory where students learn to be more conscientious by observing their more conscientious role models (Bandura, 1977; Bandura & Walters, 1963). These patterns are similar to results in Cheng (2015) where students, over the course of their secondary schooling, become more conscientious (as measured by item nonresponse) during school years where they have more conscientious teachers (as measured by survey omission). It is also possible that more conscientious teachers tend to be more effective at implementing certain classroom management or instructional techniques that could be conducive to fostering student conscientiousness (Blazar & Kraft, 2015). Indeed, our survey-effort measures of teacher conscientiousness are correlated with teacher quality measures based upon formal classroom observations by the CLASS instrument. Still, we do not find that measures of teacher quality based upon the FFT or value-added scores are predictive of student outcomes in conscientiousness. Again, more work investigating what teacher traits, exactly, are captured by classroom observation protocols or how classroom interactions between teachers and students affect student non-cognitive outcomes will be useful.

Implications for understanding teacher quality

Our findings suggest that teacher quality is multidimensional. Some teachers are effective at improving student test scores while others are more effective at improving student conscientiousness. Moreover, it appears that teachers who are themselves more conscientious are more effective at improving student conscientiousness but not necessarily student test scores.

This finding tracks with Kraft and Grace (2016) who use MET data to find that teacher effects on student non-cognitive skills and achievement are weakly correlated. Indeed, these findings align with those from a growing body of teacher quality research (Blazar & Kraft, 2015; Cheng, 2015; Gershenson, in press; Jackson, 2012; Jennings & Diprete, 2010; Koedel, 2008).

This paper joins this body of work studying teacher impacts on student non-cognitive outcomes, all of which suggest evaluations of teachers, schools, and other educational interventions to consider both student achievement and non-cognitive skills as outcomes. The former have been the focus and standard by which educational programs are evaluated, but research overlooks impacts on non-cognitive skills by solely relying on cognitive measures as outcome variables. Such oversight is not inconsequential because non-cognitive skills have been found to be important determinants for later-life outcomes, even after accounting for cognitive ability (Heckman, Stixrud, & Urzua, 2006). Without considering impacts on non-cognitive skills, the benefits that teachers impart to their students will likely be misstated (Heckman, Pinto, & Savelyev, 2013). Teachers that realize large gains in student non-cognitive skills will likely be categorized as ineffective if evaluation systems only rely on gains in student achievement (Grissom et al., 2016). In fact, it is worthwhile to reiterate that even our survey-effort measures of non-cognitive skills are predictive of student educational attainment and labor-market outcomes; this finding cannot be replicated in the MET data but has been documented in several longitudinal analyses (Hitt et al., in press; Hitt, 2015; Cheng, 2015).

Our findings also speak to prior research from the original MET study. Kane et al. (2012) find that classroom observations, student ratings, and value-added scores are only weakly correlated with student test scores and recommend creating a composite score that aggregates all these measures to gauge teacher impacts on student achievement. The composite score, as they

argue and demonstrate, is more reliable than the each measure alone and is more predictive of teacher value-added scores (see also Mihaly et al, 2013). However, the assumption behind Kane et al.'s (2012) recommendation is that teacher quality is a unidimensional construct (e.g., the ability to improve student achievement) and that different measures of teacher quality capture mutually exclusive parts of that construct. Our results provide reason to dispute that assumption. Not only is teacher quality multidimensional but different measures of teacher quality capture different aspects of teacher quality. For instance, our survey-effort measures of teacher conscientiousness clearly predict student conscientiousness but not student achievement, and student ratings of teachers are predictive of student grit, effort, and careless answering as well as test scores. Aggregating different measures of teacher quality to ascertain their predictive power to forecast student achievement may mask teacher effects on other student outcomes and obscure the multifaceted ways in which different teachers benefit their students.[19]

Implications for Future Research and Practice

Identifying effective teachers has been elusive. Observable characteristics such as years of experience, educational background, and licensure are at best only weakly correlated with student achievement outcomes (Buddin & Zamarro, 2009; Goldhaber, 2008; Jacob, 2007). In this study, we follow Rockoff et al.'s (2011) approach by first hypothesizing that teacher non-cognitive skills play an important role in shaping student outcomes. Our innovation, however, is

[19] Certainly, one could undergo the empirical exercise of recalculating ideal weights as in Kane et al. (2012) and Mihaly et al. (2013) to create composite measures of teacher quality by including our survey-effort measures of teacher and student non-cognitive skills. One of the intents behind Kane et al. (2012) and Mihaly et al. (2013) was to provide guidance for practitioners who desire to utilize traditional measures of teacher quality in a systematic way. However, we caution that our survey-effort measures of teacher quality are currently suitable for research purposes only, not for use by schools and policymakers in their everyday operations and practice.

to present evidence supporting this hypothesis by capturing a teacher's level of conscientiousness using survey effort measures.

We view this as our main contribution to the research literature. An increasing number of studies find that teachers vary in their effectiveness at improving a variety of student outcomes, but it still remains unclear what types of teachers are effective at improving particular outcomes (Hanushek & Rivkin, 2006). We are only aware of four other studies that have examined this issue. First, Rockoff et al. (2011) provides some evidence that teacher self-efficacy is related to student test scores in math, but the correlation is modest. Second, Blazar and Kraft (2015) have descriptively shown that teachers who are rated highly on the CLASS rubric appear to be more effective at improving self-efficacy among students. Third, Cheng (2015) demonstrates that students experience gains in conscientiousness in years when they have more conscientious teachers. Fourth, Duckworth et al. (2009) show that teacher grit and life satisfaction are predictive of student achievement. More work in the same vein as these studies and this present study needs to be undertaken to better understand what kinds of teachers are effective at improving both student cognitive and non-cognitive outcomes. Testing for associations between teacher personality traits, teaching practices, and student outcomes, as this work and Rockoff et al. (2011) have done, could be a more promising avenue to uncovering the elusive observable teacher characteristics that are predictive of student outcomes. One additional advantage is that these traits could be measured at the moment of hire in contrast with traditional measures of teacher quality that only available once teachers are in the classroom. More generally,

identifying the kinds of teachers that effectively develop student cognitive and non-cognitive skills is a task that warrants additional scholarly attention.[20]

For now, this work as well as other studies that find heterogeneity in the ways that teachers affect their students, suggests that scholars, policymakers, and practitioners need to think more critically about the ways in which they conceptualize teacher quality. It is not wholly unreasonable to construe teacher quality as a teacher's ability to improve student achievement. After all, student achievement and cognitive ability are crucial components of human capital development and play an important role in determining a student's future educational attainment, employment, earnings, and other long-run life outcomes (Becker, 1964; Chetty et al., 2014). However, non-cognitive skills also play their own role in determining the same outcomes, above and beyond the role that student achievement plays (Heckman et al., 2006; Almlund et al., 2011). It is also unclear to what extent scores on tests of cognitive ability are driven by student effort or student content knowledge (Borghans & Schils, 2013; Mendez et al., 2015; Hitt, Zamarro, & Mendez, 2016). A conception of teacher quality that only focus on impacts on student achievement are therefore incomplete, overlooking the nontrivial ways in which teachers benefit students.

We also view our use of survey-effort measures of teacher and student conscientiousness as a key contribution to research. These measures can be readily constructed in most data sets that rely on self-reports and provide a viable research strategy to answer questions about student and teacher non-cognitive skills – a topic that is receiving increasing attention in several

[20] Bargagliott et al., (2016) show that kindergarten teachers who utilize certain pedagogical practices to teach math appear to be more effective at improving particular non-cognitive skills in their students. Research investigating teaching practices associated with improving student non-cognitive skills is equally important, though some studies fail to find a relationship between pedagogical approaches and student outcomes (Jennings & Diprete, 2010).

academic fields. More importantly, we have shown in this study that these survey-effort measures are not simply random noise. They are related to other traditional measures of teacher quality and predictive of student outcomes. Even if one disputes their validity as measures of conscientiousness, such relationships demand explanation.

More study of these survey-effort measures needs to be completed. Little is known, for example, about the stability of such measures and how they will behave in different survey contexts. Indeed, in other work, we have found that the ways respondents shirk on surveys varies depending upon whether the survey is compulsory or voluntary and whether the survey is administered via pencil-and-paper, with an interviewer present, or via a computer (Zamarro et al., 2016; Hitt et al., in press). In fact, the differences in survey mode may explain why we found that our measures of teacher conscientiousness were predictive of student item nonresponse but not of student careless answering. Careless answering may have been a more expedient way to shirk than skipping items on the particular survey that students were tasked to complete — a pattern found elsewhere (Zamarro et al., 2016; Hitt et al., in press). This need for additional research also explains why we do not support the use of our survey-effort measures for high-stakes teacher evaluation. For now, we maintain that these measures provide useful information that can be used for research purposes that will enhance the understanding of teacher quality and how to improve student outcomes. Different teachers shape their students in many ways that affect their long-run life prospects and future well-being. It behooves researchers and policymakers to better understand the underlying mechanisms behind these developmental and formative processes.

References

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

Backes, B., & Hansen, M. (2015). Teach for America Impact Estimates on Nontested Student Outcomes. (CALDER Working Paper No. 146). Washington, DC: National Center for Analysis of Longitudinal Data in Education.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A., & Walters, R.H. (1963). *Social learning and personality development*. New York: Holt, Rinehart and Winston.

Bargagliotti, A., Gottfried, M.A., Guarino, C. (2016). The effects of kindergarten mathematical instructional oractices on young children's noncognitive development. Paper presented at the Association for Education Finance and Policy 41st Annual Conference. Denver, CO.

Blazar, D., & Kraft, M. A. (2015). Teacher and teaching effects on students' academic behaviors and mindsets (Working Paper No. 41). Cambridge, MA: Mathematica Policy Research.

Borghans, L., & Schils, T. (2012). The leaning tower of PISA. *Unpublished Manuscript.* Available at http://www.sole-jole.org/13260.pdf (Accessed 30 March 2016).

Borghans, L., ter Weel, B., & Weinberg, B.A. (2008). Interpersonal styles and labor market outcomes. *Journal of Human Resources, 43*(4), 815-858.

Buddin, R., & Zamarro, G. (2009). Teacher Qualifications and Student Achievement in Urban Elementary Schools. *Journal of Urban Economics, 66*(2), 103–115.

Cheng, A. (2015). Like teacher, like student: Teachers and the development of student noncognitive skills (EDRE Working Paper No. 2015-02). Fayetteville, AR: Department of Education Reform, University of Arkansas.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Cobb-Clark, D.A., & Schurer, S. (2012). The stability of big-five personality traits. *Economic Letters*, *115*(1), 11-15.

Dalal, R.S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology*, *90*(6): 1241-1255.

Danielson, C. (2007). *Enhancing professional practice: A framework (2nd ed)*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dobbie, W., & Fryer, R. G. (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, *123*(5), 985–1037.

Donaldson, M. (2009). So long, Lake Wobegon? Using teacher evalauation to raise teacher quality. Washington, DC: Center for American Progress.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and Passion for Long-Term Goals. Journal of Personality and Social Psychology, 92(6), 1087–1101

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (Grit-S). *Journal of Personality Assessment*, *91*(2), 166–174.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, *44*(4), 237–251.

Duckworth, A. L., Quinn, P. D., & Seligman, M. E. P. (2009). Positive predictors of teacher effectiveness. *The Journal of Positive Psychology*, *4*(6), 540–547.

Ferguson, R.F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24-28.

Galla, B., M., Plummer, B. D., White, R. E., Meketon, D., D'Mello, S. K., & Duckworth, A. L. (2014). The Academic Diligence Task (ADT): Assessing individual differences in effort on tedious but important schoolwork. *Contemporary Educational Psychology*, *39*(4), 314–325.

Garrett, R., & Steinberg, M. P. (2015). Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence from the Randomization of Teachers to Students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1), 117–156.

Gershenson, S. (in press). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*.

Goldhaber, D. D. (2008). Teachers Matter, But Effective Teacher Quality Policies are Elusive. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of Research in Education Finance and Policy* (pp. 146–165). New York, NY: Routledge.

Grissom, J. A., Loeb, S., & Doss, C. (2016). The multiple dimensions of teacher quality: Does value-added capture teacher's nonachievement contributions to their schools? In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation Systems: Making the most of multiple measures* (pp. 37–50). New York, NY: Teachers College Press.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher Quality. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 1052–1078). Amsterdam: Elsevier.

Harris, D. & Sass, T. (2014). Skills, productivity, and the evaluation of teacher performance. *Economics of Education Review, 40,* 183-204

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability. *American Educational Research Journal, 51*(1), 73–112.

Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411–482.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review, 103*(6), 2052–2086.

Hitt, C.E. (2015). Just filling in the bubbles: Using careless answer patterns on surveys as a proxy measure of noncognitive skills (EDRE Working Paper 2015-06). Fayetteville, AR: Department of Education Reform, University of Arkansas.

Hitt, C.E., Trivitt, J.R., Cheng, A. (in press). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review.*

Hitt, C.E., Zamarro, G., & Mendez, I. (2016) What if students don't care? Reexamining international differences in achievement and non-cognitive skills. Paper presented at the Association for Education Finance and Policy 41st Annual Conference. Denver, CO.

Jacob, Brian A. 2007. The challenges of staffing urban schools with effective teachers. *Future of Children* 17(1): 129–53.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–136.

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511.

Jackson, K. (2012). *Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina* (NBER Working Paper No. 18624). Cambridge, MA: National Bureau of Economic Research.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher Effects on Social and Behavioral Skills in Early Elementary School. *Sociology of Education*, *83*(2), 135–159.

Kane, T., Rockoff, J. E., & Staiger, D. (2008). What Does Certification Tell Us about Teacher Effectiveness. *Economics of Education Review*, *27*(6), 615–631.

Kane, T., McCaffrey, D.F., & Staiger, D.O. (2012). *Gathering feedback for teaching: Combining High-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*(3), 560–572.

Koedel, C., and Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18–42.

Kraft, M.A., & Grace, S. (2016). Teaching for tomorrow's economy? Teacher effects on complex cognitive skills and social-emotional competencies (Working Paper). Brown University: Providence, RI.

MacCann, C., Duckworth, A.L., & Roberts, R.D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, 19(4), 451–458.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy, 4*(4), 572–606.

Mendez, I., Zamarro, G., Clavel, J.G., & Hitt, C.E. (2015). Non-cognitive abilities and Spanish regional differences in student performance in PISA 2009 (EDRE Working Paper No. 2015-05). University of Arkansas: Fayetteville, AR.

Mihaly K., McCaffrey D. F., Staiger D., & Lockwood J. R. (2013). A composite estimator of effective teaching. Seattle, WA: Bill & Melinda Gates Foundation.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics, 130*, 105–119.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Pianta, R.C., & Hamre, B.K., (2009) Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational*

Researcher, *38*(2), 109–119.

Pianta, R.C., LaParo, K.M., & Hamre, B.K. (2008). Classroom Assessment Scoring System Manual, Pre-K. Baltimore, MA: Brookes Publishing Co.

Podgursky, M. J. (2005). Teacher licensing in U.S. public schools: The case for simplicity and flexibility. *Peabody Journal of Education*, *80*(3), 15–43.

Polikoff, M.S. (2015). The stability of observation and student survey measures of teaching effectiveness. *American Journal of Education*, *121*, 183-212.

Poropat, A.E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322-338.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458.

Roberts, B.W., Harms, P.D., Caspi, A., & Moffitt, T.E. (2007). Predicting the counterproductive employee in a child-to-adult prospective study. *Journal of Applied Psychology*, *92*(5), 1427–1436.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, *94*(2), 247–252.

Rockoff, J. E., Jacob, B., Kane, T., & Staiger, D. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, *6*(1), 43–74.

Rockoff, J. E., Staiger, D., Kane, T., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*(7), 3184–2313.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Ruzek, E.A., T. Domina, A.M. Conley, G.J. Duncan, & Karabenick, S.A. (2014). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence*, *35*(5-6), 852–882.

Tate, R. (2004). A cautionary note on shrinkage estimates of school and teacher effects. Florida *Journal of Educational Research*, *42*, 1–21.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology*, *98*(2), 438-456

Tsukayama, E., Duckworth, A.L. & Kim, B. (2013). Domain-specific impulsivity in school-age children. *Developmental Science, 16*(6), 879–893.

West, M., Kraft, M. A., Finn, A. S., Martin, R., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and Paradox: Measuring Students' Non-cognitive Skills and the Impact of Schooling. *Educational Evaluation and Policy Analysis*, *38*(1), 148–170.

White M., & Rowan B. (2012). A user guide to the "core study" data files available to MET early career grantees. Ann Arbor: Inter-University Consortium for Political and Social Research, The University of Michigan

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics, 100*, 61-78.

Zamarro, G., Engberg, J., Saavedra, J. E., & Steele, J. (2015). Disentangling Disadvantage: Can We Distinguish Good Teaching from Classroom Composition? *Journal of Research on Educational Effectiveness*, *8*(1), 84–111.

Zamarro, G., Cheng, A., Shakeel, M., & Hitt, C. (2016). Comparing and validating measures of character skills: Findings from a nationally representative sample. Paper presented at the Association for Education Finance and Policy 41st Annual Conference. Denver, CO.

Table 1: Student Outcomes Correlation Matrix

	Math Test Scores	English Test Scores	Self-Reported Grit	Self-Reported Effort	Item Nonresponse
Math Test Scores	-				
English Test Scores	0.72	-			
Self-Reported Grit	0.14	0.09	-		
Self-Reported Effort	0.22	0.23	0.44	-	
Item Nonresponse	0.08	0.09	0.04	0.04	-
Careless Answering	-0.25	-0.25	-0.21	-0.23	-0.07

Table 2: Relationships between Different Measures of Teacher Quality

	(1) Value-added in Math	(2) Value-added in English	(3) FFT Score	(4) CLASS Score	(5) Student Tripod Ratings	(6) Principal Subjective Ratings
Item Nonresponse	-0.011 (0.034) N = 813	0.014 (0.033) N = 904	-0.059** (0.029) N = 1,217	-0.068** (0.028) N = 1,238	-0.085*** (0.022) N = 1,960	-0.006 (0.013) N = 1,385
Careless Answering	-0.015 (0.033) N = 813	0.043 (0.030) N = 904	-0.054* (0.029) N = 1,217	-0.037 (0.028) N = 1,238	-0.001 (0.022) N = 1,960	-0.022 (0.013) N = 1,385
Survey Omission	0.036 (0.067) N= 1,109	0.056 (0.064) N= 1,240	-0.141** (0.061) N = 1,555	-0.185*** (0.061) N = 1,580	0.017 (0.046) N = 2,597	-0.046* (0.027) N = 1,852

Note: Value-added scores are based upon two years of data. FFT, CLASS, and Student Tripod ratings are based lessons evaluated in the first year of the MET study. All measures, except the binary Survey Omission and Principal Subjective Ratings variables, are standardized to have a mean equal to 0 and standard deviation equal to 1. Columns 1 through 5 report coefficients from bivariate regressions where survey-effort measures of non-cognitive skills are the independent variable. Column 6 reports marginal effects after estimating an analogous probit model. Standard errors are in parenthesis and sample sizes are written below. ***p<0.01; **p<0.05; *p<0.1.

Table 3: ITT Estimates of Teacher Effects (Based on Survey-effort Measures of Conscientiousness) on Student Outcomes

	Student Outcome					
	(1) Math Test Scores	(2) English Test Scores	(3) Self-Reported Grit	(4) Self-Reported Effort	(5) Item Nonresponse	(6) Careless Answering
Teacher Item Nonresponse	-0.008	-0.008	0.021	-0.015	0.024*	-0.010
	(0.013)	(0.010)	(0.013)	(0.015)	(0.014)	(0.017)
R^2	0.70	0.67	0.21	0.14	0.21	0.18
Observations	5,712	7,247	10,518	10,682	10,706	10,648
Teacher Careless Answering	-0.013	-0.006	0.017	0.021	0.035***	0.017
	(0.010)	(0.007)	(0.012)	(0.014)	(0.012)	(0.014)
R^2	0.70	0.67	0.21	0.14	0.21	0.18
Observations	5,712	7,247	10,518	10,682	10,706	10,648
Survey Omission	-0.051	-0.009	-0.058*	0.084	0.064**	0.059
	(0.033)	(0.029)	(0.035)	(0.037)	(0.030)	(0.043)
R^2	0.68	0.67	0.20	0.15	0.21	0.18
Observations	7,003	8,814	12,809	13,022	13,051	12,991

Note: All regressions control for student gender, race, age, special education status, free or reduce-priced lunch status, gifted status, English learner status, prior-year test scores, and randomization blocks, as well as classroom composition (i.e., average prior year test scores, proportion students in the class of a particular gender, race, special education status, free or reduce-priced lunch status, gifted status, and English learner status). Standard errors clustered at the classroom level. ***p<0.01; **p<0.05; *p<0.1

Table 4: IV Estimates of Teacher Effects (Based on Survey-effort Measures of Conscientiousness) on Student Outcomes

	Student Outcomes					
	(1) Math Test Scores	(2) English Test Scores	(3) Self-Reported Grit	(4) Self-Reported Effort	(5) Item Nonresponse	6) Careless Answering
Teacher Item Nonresponse	-0.011	-0.012	0.031	0.021	0.033	-0.019
	(0.020)	(0.015)	(0.023)	(0.026)	(0.023)	(0.028)
R^2	0.70	0.67	0.21	0.14	0.21	0.18
Observations	5,712	7,247	10,102	10,253	10,277	10,229
Teacher Careless Answering	-0.023	-0.011	0.030	0.040	0.059***	0.026
	(0.017)	(0.012)	(0.024)	(0.026)	(0.022)	(0.027)
R^2	0.70	0.67	0.21	0.14	0.21	0.18
Observations	5,712	7,247	10,102	10,253	10,277	10,229
Survey Omission	-0.086	-0.015	-0.086	-0.146**	0.137***	0.134*
	(0.055)	(0.050)	(0.064)	(0.068)	(0.052)	(0.074)
R^2	0.68	0.67	0.20	0.13	0.21	0.18
Observations	7,003	8,814	12,228	12,425	12,454	12,404

Note: All regressions control for student gender, race, age, special education status, free or reduce-priced lunch status, gifted status, English learner status, prior-year test scores, and randomization blocks, as well as classroom composition (i.e., average prior year test scores, proportion students in the class of a particular gender, race, special education status, free or reduce-priced lunch status, gifted status, and English learner status). Standard errors clustered at the classroom level. ***p<0.01; **p<0.05; *p<0.1

Table 5: ITT Estimates of Teacher Effects (Based on Traditional Measures of Teacher Quality)

	Student Outcomes			
	(1) Self-Reported Grit	(2) Self-Reported Effort	(3) Item Nonresponse	(4) Careless Answering
Student Tripod Ratings	0.034***	0.051***	-0.003	-0.091***
	(0.011)	(0.014)	(0.017)	(0.017)
R^2	0.20	0.13	0.19	0.18
Observations	12,364	12,563	12,587	12,527
Received Higher Principal Ratings	-0.030	0.024	-0.006	-0.098***
	(0.032)	(0.034)	(0.026)	(0.037)
R^2	0.20	0.13	0.25	0.18
Observations	9,873	10,042	10,064	10,010
FFT Score	0.002	-0.000	0.018	-0.009
	(0.016)	(0.018)	(0.020)	(0.021)
R^2	0.21	0.13	0.21	0.18
Observations	11,786	11,988	12,016	11,956
CLASS Score	0.027*	-0.010	0.027*	-0.039**
	(0.016)	(0.017)	(0.014)	(0.020)
R^2	0.21	0.13	0.21	0.18
Observations	11,786	11,988	12,016	11,956
Teacher Value Added (English)	0.013	0.016	-0.020	-0.027
	(0.018)	(0.021)	(0.019)	(0.024)
R^2	0.21	0.12	0.24	0.18
Observations	7,190	7,302	7,324	7,310
Teacher Value Added (Math)	-0.008	0.028	0.051	-0.001
	(0.024)	(0.025)	(0.033)	(0.034)
R^2	0.22	0.13	0.12	0.18
Observations	6,762	6,886	6,894	6,844

Note: Principal ratings is a dichotomous variable where the omitted category represents teachers in approximately the lower half of the distribution of principal ratings. All regressions control for student gender, race, age, special education status, free or reduce-priced lunch status, gifted status, English learner status, prior-year test scores, and randomization blocks, as well as classroom composition (i.e., average prior year test scores, proportion students in the class of a particular gender, race, special education status, free or reduce-priced lunch status, gifted status, and English learner status). Standard errors clustered at the classroom level. ***p<0.01; **p<0.05; *p<0.1

Table 6: IV Estimates of Teacher Effects (Based on Traditional Measures of Teacher Quality)

	Student Outcomes			
	(1) Self-Reported Grit	(2) Self-Reported Effort	(3) Item Nonresponse	(4) Careless Answering
Student Tripod Ratings	0.054**	0.075***	-0.009	-0.146***
	(0.021)	(0.025)	(0.029)	(0.028)
R^2	0.20	0.13	0.19	0.18
Observations	11,840	12,024	12,048	11,998
Received Higher Principal Ratings	-0.052	0.041	-0.011	-0.171***
	(0.056)	(0.059)	(0.044)	(0.064)
R^2	0.21	0.13	0.24	0.18
Observations	9,873	10,042	10,064	10,010
FFT Score	0.008	0.019	0.013	-0.045
	(0.027)	(0.030)	(0.033)	(0.033)
R^2	0.21	0.13	0.21	0.18
Observations	11,263	11,450	11,478	11,428
CLASS Score	0.045	0.021	0.036	-0.077**
	(0.028)	(0.029)	(0.024)	(0.034)
R^2	0.21	0.13	0.21	0.18
Observations	11,263	11,450	11,478	11,428
Teacher Value Added (English)	0.024	0.033	-0.067*	-0.081
	(0.042)	(0.050)	(0.039)	(0.049)
R^2	0.21	0.12	0.24	0.18
Observations	6,895	6,995	7,017	7,011
Teacher Value Added (Math)	-0.012	0.059	0.113	-0.003
	(0.058)	(0.060)	(0.078)	(0.079)
R^2	0.22	0.13	0.12	0.19
Observations	6,385	6,504	6,512	6,464

Note: Principal ratings is a dichotomous variable where the omitted category represents teachers in approximately the lower half of the distribution of principal ratings. All regressions control for student gender, race, age, special education status, free or reduce-priced lunch status, gifted status, English learner status, prior-year test scores, and randomization blocks, as well as classroom composition (i.e., average prior year test scores, proportion students in the class of a particular gender, race, special education status, free or reduce-priced lunch status, gifted status, and English learner status). Standard errors clustered at the classroom level. ***p<0.01; **p<0.05; *p<0.1

Appendix A
Items on Student Effort Scale

Items for Elementary School Students

1. In this class, I take it easy and do not try very hard to do my best. *(Reverse coded)*

2. In this class, I stop trying when the work gets hard. *(Reverse coded)*

3. When doing schoolwork for this class, I try to learn as much as I can and don't worry about how long it takes.

4. I have pushed myself hard to completely understand lessons in this class.

Items for Secondary School Students

1. In this class, I take it easy and do not try very hard to do my best. *(Reverse coded)*

2. In this class, I stop trying when the work gets hard. *(Reverse coded)*

3. When doing schoolwork for this class, I try to learn as much as I can and don't worry about how long it takes.

4. I have pushed myself hard to completely understand lessons in this class.

5. Overall, between homework, reading, and other class assignments, I work hard.

Appendix B: Post-Attrition Baseline Covariate Balance after Randomization

Table B1. Post-Attrition Baseline Covariate Balance across Randomly Assigned Teachers within Randomization Block

	F-Statistic	P-Value
Student Characteristic		
Age	0.82	1.000
Male	0.57	1.000
English-language Learner	1.28	0.000
Special Education	0.93	0.953
Gifted	1.55	0.000
Free or reduced-priced lunch	0.67	1.000
Black	0.68	1.000
Hispanic	0.72	1.000
Asian	0.69	1.000
White	0.71	1.000
Prior Year Math Test Scores	1.60	0.000
Prior Year English Test Scores	1.50	0.000

Notes: We estimated a model that used teacher fixed effects to predict each student characteristic, while also controlling for randomization blocks. This table displays F-statistics and p-values from tests that coefficients estimates for teacher fixed effects are jointly equal to zero. **p<0.05; *p<0.1.

Table B2. Post-Attrition Baseline Covariate Balance between Student Characteristics and Randomly Assigned Teacher Characteristics

	Teacher Quality Measure of Randomly Assigned Teacher					
Student Characteristic	Tripod Score	FFT Score	CLASS Score	Survey Omission	Item Nonresponse	Careless Answering
Age	0.022	-0.011	-0.002	-0.018	-0.013	0.008
	(0.020)	(0.019)	(0.009)	(0.014)	(0.065)	(0.005)
Male	0.006	0.030	0.013	-0.003	-0.088	0.011**
	(0.020)	(0.019)	(0.011)	(0.012)	(0.062)	(0.005)
English-language Learner	-0.010	-0.025*	-0.005	0.012	0.033	0.003
	(0.011)	(0.014)	(0.007)	(0.013)	(0.050)	(0.003)
Special Education	0.003	-0.006	-0.001	0.007	0.014	0.002
	(0.010)	(0.010)	(0.006)	(0.008)	(0.034)	(0.003)
Gifted	0.004	0.003	0.002	0.021	0.046	-0.008*
	(0.011)	(0.015)	(0.006)	(0.009)	(0.055)	(0.004)
Free or reduced-priced lunch (FRL)	0.006	-0.014	-0.016*	0.012	0.053	-0.004
	(0.016)	(0.020)	(0.009)	(0.012)	(0.059)	(0.005)
Black	-0.002	0.002	-0.001	-0.018*	-0.023	0.003
	(0.012)	(0.014)	(0.006)	(0.010)	(0.042)	(0.003)
Hispanic	0.010	-0.011	0.001	-0.002	0.039	0.001
	(0.014)	(0.019)	(0.001)	(0.012)	(0.052)	(0.004)
Asian	-0.001	0.014	0.004	0.009	0.029	-0.001
	(0.009)	(0.014)	(0.006)	(0.007)	(0.023)	(0.002)
White	-0.006	-0.007	-0.006	-0.015**	0.034	-0.004
	(0.013)	(0.014)	(0.006)	(0.008)	(0.044)	(0.004)
Prior Year Math Test Scores	-0.037	0.067	0.017	0.001	-0.037	-0.030**
	(0.043)	(0.047)	(0.023)	(0.032)	(0.168)	(0.013)
Prior Year English Test Scores	-0.059	0.021	-0.006	-0.023	0.111	-0.028**
	(0.044)	(0.048)	(0.023)	(0.033)	(0.156)	(0.133)

Notes: This table displays coefficient estimates from separate bivariate regressions where the dependent variable is a student characteristic and the independent variable is a teacher quality measure. Standard errors are in parenthesis. All regressions control for randomization block. **p<0.05; *p<0.1.