Making Inferences about Teacher Observation Scores over Time

Derek C. Briggs

Jessica L. Alzen

University of Colorado Boulder

November 1, 2018

Pre-print of Briggs, D. C. & Alzen, J. L. (2019). Making inferences about teacher observation scores over time. *Educational and Psychological Measurement*.
https://doi.org/10.1177/0013164419826237

Abstract

Observation protocol scores are commonly used as status measures to support inferences about teacher practices. When multiple observations are collected for the same teacher over the course of a year, some portion of a teacher's score on each occasion may be attributable to the rater, lesson and time of year of the observation. All three of these are facets that can threaten the generalizability of teacher scores, but the role of time is easiest to overlook. A generalizability theory framework is used in this study to illustrate the concept of a hidden facet of measurement. When there are many temporally spaced observation occasions, it may be possible to support inferences about the growth in teaching practices over time as an alternative (or complement) to making inferences about status at a single point in time. This study uses longitudinal observation scores from the Measures of Effective Teaching project to estimate the reliability of teacher-level growth parameters for designs that vary in the number and spacing of observation occasions over a two-year span. On the basis of a subsample of teachers scored using the Danielson Framework for Teaching, we show that at least 8 observations over two years are needed before it would be possible to make distinctions in growth with a reliability coefficient of .38.

Introduction

Teaching is a multifaceted practice, and as such it can be very difficult to evaluate. Teachers are expected to establish rapport with their students, well-organized classroom routines, and perhaps most importantly, an environment that is conducive to learning. With respect to instruction, teachers are expected to clearly communicate their learning objectives, actively engage their students with questioning and discussion techniques, use assessment for both formative and summative purposes, and find ways to make topics for instruction culturally relevant. It is usually assumed that teachers who do these sorts of things well are ones who are likely to have the most positive impacts on student learning, and that this should be evident in growth on students' standardized assessment scores. There is, at this point, a well-established literature on methods devised to evaluate teachers on the basis of value-added models (Braun, Chudowsky, & Koenig, 2010; Chetty Friedman, & Rockoff 2014a; Chetty, Friedman, & Rockoff, 2014b; Everson, 2017; Hanushek & Rivkin, 2010; Harris, 2009; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2003; McCaffrey, Han, & Lockwood, 2009; Rothstein, 2010; Stacy, Guarino, & Wooldridge, 2018). Value-added models attempt to quantify the effect of a teacher on her students' academic achievement, but they represent a black box with respect to the specific kinds of practices that produce the effect. In addition, even if one puts aside concerns about the validity and reliability of the inferences that value-added estimates can support (Bacher-Hicks, Chin, Kane, & Staiger, 2017; Ballou & Springer, 2015; Cohen & Goldhaber, 2016; Goldhaber, 2015; Goldhaber & Hansen, 2013; Sanders & Horn, 1998; Sanders, Saxton, & Horn, 1997), even in a best-case scenario such an approach can only be used to estimate these effects for subject domains in which it is feasible to develop valid assessments of

the knowledge, skills and abilities students are expected to develop. For a majority of subjects and grade levels, these assessments are unavailable. Even when they are, at best these assessments only provide information about a sample of what students may or may not have learned, and no information at all about so-called "non-cognitive" attributes that a healthy classroom environment would be expected to foster (Blazer & Kraft, 2015).

For these reasons, any research or evaluation that focuses on differentiating the quality of teachers' classroom practices will invariably require direct observations of these practices, and some method for distinguishing features that make some aspects of these practices better than others. A common tool for this purpose is the *observation protocol* (Bell, Gitomer, McCaffrey, Hamre, Pianta & Qi, 2012; Brabeck, 2014; Goe, Bell & Little, 2008; Pianta & Hamre, 2009). An observation protocol is a set of materials given to an individual who has been asked to observe a unique lesson being facilitated by a teacher for a specific classroom on a specific occasion during the school year. The mode of the observation may be "live" or on video, in which case there will be a lag between the date of the lesson and the date of the observation. The materials in the protocol essentially tell the individual what to pay attention to when observing the interaction between teacher and students. It would be possible to use an observation protocol solely as a guide to recording qualitative observations. When the protocol requires the observer to go beyond this, to transform these qualitative observations into a set of ordered numeric scores for each of some finite set of practices, the protocol becomes the instrument for a measurement procedure, and the observer becomes a "rater." The scores that result from this procedure are typically averaged across multiple practices, lessons, and raters, and then attached to teachers so that they can be compared either to one another, or to some criterion-referenced standard. Finally, these comparisons may be made for purposes that range from those that have generally

low stakes (e.g., research on teacher induction or professional development) to those that have generally high stakes (e.g., evaluation in the context of a system of educational accountability).

Establishing the validity of inferences based on observation protocol scores requires a multifacteted array of evidence, and studies along these lines have become more common in recent years. Bell et al. (2012) provide an excellent framework, premised on Kane's (2006) argument-based approach to test validation, for how such studies can be conceptualized and organized. They do so by sketching out an interpretive argument for what should be in place before the scores from an observation protocol could be validly used to make inferences about teaching quality. The argument has four hierarchical components that start with the *scoring* of the observations, then move to the *generalization* of the scores to some unobserved target score, the *extrapolation* of the score to some behavioral domain outside of the observation, and the *implication* of score use for consequential decision-making. Bell et al. provide examples of analyses that could be done to evaluate each of these components in one or more studies to build a comprehensive validity argument. One of the most important of these involves an analysis of the generalizability of observation scores (Cohen & Goldhaber, 2016), and it is this aspect within the broader context of observation protocol validity that is the focus of the present paper.

The issue of score generalization is itself a broader way to conceptualize score reliability. Unlike the standardized achievement test given to a student, where the principal facet of "error" that influences the observed score a student receives comes from the selection of test items, there are multiple facets that can influence the observation protocol score for a teacher that are outside of the control of the teacher being observed. Prominent examples of these facets include the choice of rater, lesson, lesson segment for an observation, as well as the mode in which observations are scored. A number of studies have used Generalizability Theory (G-Theory;

Cronbach, Gleser, Harinder Nanda, & Rajaratnam, 1972; Brennan, 2001) to decompose the variance of observation protocol scores into that which is attributable to the true differences between teachers of interest, and that which is attributable to the facets most central to the protocol scoring design (Bell et al, 2012; Casabianca, McCaffrey, Gitomer, Bell, Hamre, & Pianta, 2013; Charalambous, Kyriakides, Tsangaridou & Kyriakides, 2017; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Mashburn, Meyer, Allen, & Pianta, 2013). However, in all of these studies, it can be argued that there is an important "hidden" facet to the measurement procedure involved in the scoring of a classroom observation. This facet is the occasion, or more specifically, the temporal location of the observation within the academic school year. As a facet of a measurement procedure in a typical G-Theory context, the occasion of observation can influence teacher scores by introducing a random source of variance. A methodological problem comes in the estimation of this source of variance since it is usually confounded with variability due to differences in the lesson of instruction on different occasions. When the problem is ignored, it may introduce bias into the estimate of a generalizability coefficient. In addition, the occasion of measurement may also introduce bias directly into teacher observation scores. This can happen if teaching practices are, in fact, getting better or worse over time. If, for example, teachers improve over the course of a school year, a teacher would benefit from having more observations clustered near the end of the year than at the beginning.

This latter issue was taken up in a study by Casabianca, Lockwood and McCaffrey (2015). The authors use data collected longitudinally from the CLASS-S observation protocol (Pianta, Hamre, Haynes, Mintz, & LaParo, 2007) to develop "augmented" G-Theory models that allow them to separately estimate the variance components for facets such as lesson segments, raters, classrooms and their interactions, as well as variance components for parameters

associated with time trends in these scores. These time trends are specified separately to capture both changes in teaching practices and changes in rater judgments about these practices. Casabianca et al. find that while these main effect trends are responsible for a fairly small proportion of observed score variance, they can still introduce a significant source of bias. Somewhat counterintuitively, teaching practice scores showed evidence of a small linear decline over time. In contrast, rater judgments showed evidence of a "burn-in" trend in which raters generally provide higher scores at the outset, but then rapidly adjust downward and largely stabilize. Though both trends were significant, the impact of the time trend on raters tended to be two to three times as large as the general impact of time on the practices being rated.

The present paper can be regarded as a companion piece to the work by Casabianca et al. with two main purposes. The first purpose is to explain the concept of a hidden facet using the context of a measurement procedure defined within a G-Theory framework. In our experience both researchers and practitioners involved with the design, scoring and interpretation of teacher observations are seldom familiar with the idea of a multifaceted measurement procedure, and even when this has been introduced, it does not include a discussion about the concept of a hidden facet. The second purpose is to use data from the Measures of Effective Teaching (MET) project, which has almost twice as many teacher observations over a two-year span, to specify a simpler version of Casabianca et al.'s augmented G-Theory model: a two-level HLM with observation scores repeated at the first level, and teachers at the second. Under certain conditions, this model could be used to generate teacher-specific estimates of growth in observation scores. We focus attention on the reliability of these estimates, and on the design principles that would need to be in place to maximize reliability.

There are five sections that follow. First, we introduce a conceptual framework premised on G-Theory to show why an occasion of measurement, if it is ignored or viewed as exchangeable, has the potential to introduce bias into estimates of the reliability of an observation protocol score. Next, instead of ignoring the occasion facet, we suggest an approach for modeling it longitudinally as a repeated measure over time. Within this new context there are now two different ways to conceptualize reliability, one that focuses on the reliability of score levels, and another that focuses on the reliability of score growth. In the third section, we describe the empirical data that we use to estimate and compare the reliability of score levels and score growth from the Danielson Framework for Teaching (FFT; Danielson, 2013). These data, which comes from the aforementioned MET project, are not ideal as a means of disentangling the variability in score levels attributable to occasions from the variability attributable to lessons and raters at a single point in time. However, because they include up to 16 observations of unique lessons from individual teachers over two academic school years, they provide for a proof of concept for the estimation and interpretation of score growth. In the fourth section, we present our results from fitting two longitudinal models with data that differ in the way that unique occasions of time are conceptualized. We conclude with a discussion of these results and the implications of our findings for the use of observation protocols to evaluate growth in teacher practices.

The Generalizability of Observation Protocol Scores with Occasion as a Hidden Facet

G-Theory is an extension of Classical Test Theory (Lord & Novick, 1968) in which a person's observed score on a measurement procedure is characterized as a linear model in the

general form $Y = T + \varepsilon$ where the term T is the "true score," or alternatively the "universe score" one would observe if it were possible to compute an expected value for Y taken over one or more design *facets* that can vary over unique hypothetical replications of the measurement procedure. In Classical Test Theory, focus is placed on a single facet of measurement in a testing scenario (the item), and the computation of the reliability coefficient $\frac{\sigma^2(T)}{\sigma^2(T)+\sigma^2(\varepsilon)}$. G-Theory differs from Classical Test Theory in that it allows for the specification of multiple design facets that contribute to measurement error, and this leads to a decomposition of the term $\sigma^2(\varepsilon)$ with respect to these different facets. Hence while G-Theory is still a linear model, it unpacks the undifferentiated term ε in Classical Test Theory into multiple independent random effects.

Consider the conventional context of a standardized achievement test taken by a student as an example of a measurement procedure. When the test consists entirely of selected responses to items that are all objectively scored, the key design facet is the choice and number of items that comprise any given testing event. If one can plausibly assume that the items on a test event represent a random sample of *n* items from a defined population of *N* items, then what we wish to estimate is the sampling variance of a student's test score associated with the random selection of *n* items. If data has been collected such that it is defensible to regard students and items as random samples from their respective populations, one can proceed to estimate a *Generalizability Coefficient* ($E\rho^2$), which we write below using notation that is more general than the classical expression for reliability:

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)}\,. \qquad (1)$$

In Equation 1, $\sigma^2(\tau)$ represents variance in scores across the object of measurement (but now described as "universe score variance"), and $\sigma^2(\delta)$ represents variance in observed scores

attributable to measurement error (i.e., variance due to all design facets that interact with the object of measurement). The scenario in which a standardized achievement test is taken by students can be expressed as the simplest possible measurement procedure and data collection design, $p \times i$, where the object of measurement (p) is crossed with a single facet of measurement, items (i).

In contrast to a conventional standardized test, an observation protocol is a *multifaceted* measurement procedure. In addition to the "items" on which ratings are to be provided, other possible facets of an observation protocol could include the choice of rater to score the observation (r), the lesson being observed (l), the segment of the lesson being observed for scoring (s), the occasion on which the observation is being conducted (o), and even the format of the observation (i.e., live or recorded on video) (f). To simplify matters in what follows, we will assume that the lesson being observed and the segment of the lesson being observed are the same thing, that the format facet is fixed for any hypothetical protocol, and that while there are multiple items within a protocol, these are no other items that would be admissible as replacements from some larger population of items. This means that whenever an observation protocol is used to conduct observations for some sample of teachers (p), the same N items are always used, the observation always lasts for the same amount of time at the same point in the lesson, and that all observations are either conducted live or viewed in a recorded format. This leaves us with raters, lessons, and occasions as facets for which it would be possible to conceptualize populations of raters, lessons, and occasions that would all constitute exchangeable conditions for a hypothetical replication of the measurement procedure. In a G-Theory framework, the raters (r), lessons (l), and occasions (o) are all regarded as random samples from populations R, L, and O. It is because of the intent to generalize scores across the

combination of these different facet populations that a G-Theorist would refer to $\sigma^2(\tau)$ as "universe" score variance instead of just "true" score variance, $\sigma^2(T)$.

As an example of an ideal design of a generalizability *study*, administrative staff from a school district might meet over the summer to establish an admissible population of teachers that would be eligible to be observed teaching a class during the coming school year. The district would also need to establish admissible populations of raters, lessons, and occasions eligible for use in any given observation. In the resulting study, teachers might be stratified by school, grade level, and subject area specialization. Random samples would be drawn, first by school, and then by grade level and subject matter specialization within school. This might result in, say, six samples of 20 teachers in elementary, middle, and high school with subject area specializations in math or English Language Arts (for a total of 120 unique teachers). Teachers in each of these six samples would be videotaped teaching two different lessons on the same day on three different occasions during the school year. These observations would each be scored by three different raters. In this design, unique lessons are nested within occasions, unique occasions are nested within teacher, and both facets of measurement (lessons, occasions) are crossed with raters and the fixed items on the protocol for a $l{:}o{:}p$ x r x I design. This idealized design is given a visual representation with the Venn Diagram in Figure 1. The circle for the item facet is denoted with a dashed line to indicate that it is a fixed facet. Analysis of variance techniques could be used to partition the total observed variance across teachers into true differences among teachers and different sources of measurement error. This could be used to estimate the Generalizability Coefficient shown in Equation 2.

Insert Figure 1 about here

$$E\rho^2 =$$

$$\frac{\sigma^2(p)+\sigma^2(pi)/N_i}{\left(\sigma^2(p)+\frac{\sigma^2(pi)}{N_i}\right)+\left(\frac{\sigma^2(pr)}{n_r}+\frac{\sigma^2(o:p)}{n_o}+\frac{\sigma^2(io:p)}{N_i n_o}+\frac{\sigma^2(l:o:p)}{n_l n_o}+\frac{\sigma^2(pri)}{N_i n_r}+\frac{\sigma^2(il:o:p)}{N_i n_l n_o}+\frac{\sigma^2(irl:o:p,e)}{N_i n_r n_l n_o}+\frac{\sigma^2(ri:op)}{N_i n_r n_o}+\frac{\sigma^2(r:op)}{n_r n_o}+\frac{\sigma^2(rlo:p)}{n_r n_l n_o}\right)}$$

(2)

Comparing the expression in Equation 2 to the more general expression in Equation 1, note that $\sigma^2(\tau) = \sigma^2(p) + \sigma^2(pi)/N_i$, which indicates that universe score variance is driven not just by true variance among teachers, but by the interaction of teachers with the particular set of items that raters use to score teachers according to guidelines of the observation protocol. The latter is not a source of measurement error because it is fixed, which is why $\sigma^2(pi)$ is divided by N_i (representing the population size) rather than n_i (representing the sample size). Continuing with a comparison to Equation 1, note that the term $\sigma^2(\delta)$ has been decomposed into 10 different source of measurement error through the interactions of lessons, occasions, items and raters with the object of measurement, teachers. Each of these can be seen visually as a specific intersection between the four circles shown in Figure 1. With this equation and estimated variance components from the study in hand, one could settle upon an optimal design for the observation protocol with respect to the number of raters, lessons and occasions one would need to observe in order to minimize these different sources of measurement error and thereby maximize the generalizability of the observation protocol scores.

In practice, outside of the work by Casabianca et al. (2015), there has been no study of which we are aware that has attempted to disentangle the variance in observation protocol scores attributable to lessons from the variance uniquely attributable to occasions. Instead, it is more

common to see a *l:p* x *r* x *I* design with lessons nested in teachers crossed by raters and fixed items. For example, using unique data from a supplementary study conducted for the MET project, Ho & Kane (2013) illustrate three scenarios consisting of different combinations of lessons and raters (with the time spent doing the observation held constant). In a design scenario with two lessons, each observed and scored by a different rater, the predicted generalizability coefficient would .59. In a design with four lessons with three unique raters, the coefficient would increase to .69. In these designs, occasions are fully confounded with lessons and vice-versa. As a result, even the relatively low generalizability coefficients estimated by Ho & Kane may well be overestimated as a result of bias caused by "hidden" occasion facets (Brennan, 2001). These facets exist conceptually as a potential source of measurement error, even if they have not been estimated empirically. That is, in reality both a lesson facet *and* an occasion facet are nested within teachers, but only the variance associated with their interaction can be estimated.

This concept is illustrated by the Venn Diagram in Figure 2. The four distinct, shaded areas within Figure 2 represent specific sources of observed score variance that cannot be distinguished from universe score variance. Among these shaded areas, going clockwise from left to right, $o{:}p$ represents variance in a teacher's score due to choice of occasion (e.g., some teachers will look better or worse on a Monday relative to a Wednesday, or on cold day relative to a warm day, etc.); $r\ x\ o{:}p$ represents variance due to the specific rater or raters observing a given teacher on a given occasion (e.g., some raters may be harsher or more generous at certain times of the school year); $r\ x\ I\ x\ o{:}p$ represents variance due to the specific rater or raters observing a given teacher on a given occasion with this specific set of protocol items (e.g., the severity of raters can depend on both the mixture of occasion and the specific things they are

asked to observe about teachers on that occasion); and $I\ x\ o{:}p$ represents variance due to an occasion that is specific to the fixed set of items on the protocol (e.g., student-teacher interactions, a common element on observation protocol scores, might be different during times such as the week before a major holiday break).

Insert Figure 2 about here

Under an *l:p* x *r* x *I* design, the Generalizability Coefficient takes on the specific form in Equation 3

$$E\rho^2 = \frac{\sigma^2(p)+\frac{\sigma^2(pi)}{N_i}+\left\{\frac{\sigma^2(o{:}p)}{n_o}+\frac{\sigma^2(io{:}p)}{N_i n_o}\right\}}{\left(\sigma^2(p)+\frac{\sigma^2(pi)}{N_i}+\left\{\frac{\sigma^2(o{:}p)}{n_o}+\frac{\sigma^2(io{:}p)}{N_i n_o}\right\}\right)+\left(\left\{\frac{\sigma^2(pr)+\sigma^2(r{:}op)}{n_r}\right\}+\left\{\frac{\sigma^2(pri)+\sigma^2(ri{:}op)}{N_i n_r}\right\}+\frac{\sigma^2(il{:}o{:}p)}{N_i n_l}+\frac{\sigma^2(l{:}o{:}p)}{n_l}+\frac{\sigma^2(rlo{:}p)}{n_r n_l}+\frac{\sigma^2(irl{:}o{:}p,e)}{N_i n_r n_l}\right)}$$

(3)

A first key distinction between Equations 2 and 3 is the shift in the location of the variance components $\frac{\sigma^2(o{:}p)}{n_o}$ and $\frac{\sigma^2(io{:}p)}{N_i n_o}$ (shown in brackets in both the numerator and within the first term in parentheses in the denominator) from representing components that contribute to specific sources of measurement error (i.e., $\sigma^2(\delta)$ in Equation 1) to representing sources that now contribute to universe score variance (i.e., $\sigma^2(\tau)$ in Equation 1). The intuition here is that there is variability in the different scores given to teachers each time a new lesson is observed that is not due to the lesson, but due to the occasion on which the lesson is being observed. If some teachers happen, by chance, to be more "on" or "off" on some days, weeks or months of

the year irrespective of the lesson they happen to be teaching, this will result in variability in scores that will be mistakenly attributed to true differences among teachers.

A second key distinction between Equations 2 and 3 is that the terms $\frac{\sigma^2(r{:}op)}{n_r n_o}$ and $\frac{\sigma^2(ri{:}op)}{N_i n_r n_o}$ are now confounded within the respective sources of error variance, $\frac{\sigma^2(pr)}{n_r}$ and $\frac{\sigma^2(pri)}{N_i n_r}$. In Equation 2, the combined sources of error variance were $\left\{\frac{\sigma^2(pr)}{n_r} + \frac{\sigma^2(r{:}op)}{n_r n_o}\right\}$ and $\left\{\frac{\sigma^2(pri)}{N_i n_r} + \frac{\sigma^2(ri{:}op)}{N_i n_r n_o}\right\}$. Compare these two terms from Equation 2 to the compatible terms in brackets within the set of parentheses of the Equation 3 denominator: $\left\{\frac{\sigma^2(pr)+\sigma^2(r{:}op)}{n_r}\right\}$ and $\left\{\frac{\sigma^2(pri)+\sigma^2(ri{:}op)}{N_i n_r}\right\}$. In both Equations 2 and 3, $\sigma^2(r{:}\,op)$ and $\sigma^2(ri{:}\,op)$ represent sources contributing to error variance that interacts with the rater facet (some raters tend to be more severe or lenient with the teachers they observe on certain occasions, irrespective of the lesson they are observing). But because they are hidden facets in Equation 3, error variance can only be reduced by dividing, respectively, by the n_r (the total number of raters), and $N_i n_r$ (the product of the fixed number of items and total number of raters) as opposed to dividing a portion of the variance by $n_o n_r$ and $N_i n_r n_o$ respectively.

Unless the increase to true teacher variance by the addition of $\frac{\sigma^2(o{:}p)}{n_o}$ and $\frac{\sigma^2(io{:}p)}{N_i n_o}$ to the numerator when going from Equation 2 to 3 is perfectly offset by the inflation in error variance caused by replacing $\left\{\frac{\sigma^2(pr)}{n_r} + \frac{\sigma^2(r{:}op)}{n_r n_o}\right\}$ and $\left\{\frac{\sigma^2(pri)}{N_i n_r} + \frac{\sigma^2(ri{:}op)}{N_i n_r n_o}\right\}$ with $\left\{\frac{\sigma^2(pr)+\sigma^2(r{:}op)}{n_r}\right\}$ and $\left\{\frac{\sigma^2(pri)+\sigma^2(ri{:}op)}{N_i n_r}\right\}$ in the denominator, the presence of the hidden occasion facet is likely to bias estimates of reliability based on Equation 3. If the relative increase to the numerator is larger than the relative increase to the denominator, then estimates of reliability will

be biased upwards. Now, from a theoretical point of view, it may seem hard to imagine that the terms $\left\{\frac{\sigma^2(o:p)}{n_o} + \frac{\sigma^2(io:p)}{N_i n_o}\right\}$ will be especially large, in which case the bias to $E\rho^2$ may be negligible, but this is also an empirical issue that remains mostly unexplored.

The ideal way to gain empirical insights into this issue would be to design a study in which the variance attributable to occasions and lessons can be disentangled. As in the hypothetical example at the outset of this section, such a study would require, at a minimum, samples of teachers able to teach multiple distinct lessons within and across multiple occasions. However, it is also possible that at least some of the variability in scores on a given occasion is something that is not random but instead caused by true differences in the growth in teaching practices over time. If this is the case, then the application of a conventional G-Theory model for variance decomposition would fail to properly parameterize this trend. The approach taken by Casabianca et al (2015) represents something of a compromise. On the one hand, their data collection design does not allow for the estimation of all unique variance components associated with an occasion facet. On the other hand, through the specification of an augmented G-Theory model, they are able to at least distinguish, at any given point in time, between true variability among teachers and the variability induced by a main effect for growth in teaching practices.

An interesting idea that the augmented G-Theory model approach raises is the possibility of focusing attention not just on the status of teacher scores at a single point in time (or averaged across multiple points in time), but instead making inferences about teacher-specific growth trends over time. This was not attempted in the Casabianca et al. study in part because no teacher had more than a total of four observations available. In the present study, we take advantage of the availability of data for teachers who typically have seven to eight observations over a two year span. Though our data does not include a design with multiple raters per

observation, (making it impossible for us to apply the same augmented G-Theory model as in the Casabianca et al. study) we show how issues of observation protocol score generalizability can still be examined in the context of a simple instantiation of an augmented linear model, one that will be both familiar to many and easy to implement: a two-level HLM.

Methods

Let Y_{tp} represent the observation protocol score of teacher p on time occasion t after scores have been averaged across the unique items (or dimensions) of the protocol. The scores could be based on judgments from a single rater on that occasion, or from multiple raters[1]. If they are based on multiple ratings, these would also be averaged. In other words, the score Y_{tp} is an average over two facets of the measurement procedure, items and raters. One could choose to model and decompose Y_{tp} as a function of the time variable *X* using the simple HLM below.

$$Y_{tp} = \pi_{0p} + \pi_{1p}(X_{tp}) + \varepsilon_{tp} \quad (4a)$$

$$\pi_{0p} = \beta_{00} + r_{0p} \quad (4b)$$

$$\pi_{1p} = \beta_{10} + r_{1p} \quad (4c)$$

where $\varepsilon_{tp} \sim N(0, \sigma_\varepsilon^2)$, $r_{0p} \sim N(0, \sigma_0^2)$, and $r_{1p} \sim N(0, \sigma_1^2)$. It is assumed here that ε_{tp} is independent of r_{0p} and r_{1p}, but that r_{0p} and r_{1p} come from a bivariate normal distribution with an unknown covariance, σ_{01}. When expressed as a single equation by substituting Equations 4b

[1] A complication arises when the temporal occasion of the rating is distinct from the temporal occasion of the lesson being observed. In the data that we ultimately use from the MET project this was in fact the case, but only the time of the actual lesson was available to us, so we are essentially ignoring this complication. See Casabianca et al. (2015) for details of an approach when both time of lesson and time of rating are available.

and 4c into 4a, this model can be viewed as an instance of the augmented G-Theory model described by Casabianca et al. (2015, p. 321, Eqn 2). Here, the total variance in observation scores across teachers and occasions, $\sigma^2(Y_{tp})$, is being decomposed into the error variance across teachers within any given occasion, σ_ε^2, the universe score variance within an occasion, σ_0^2, and the true variance in the growth of scores across occasions, σ_1^2.

There are two teacher-level parameters in this model. The first parameter is π_{0p}, which represents the observation score for a teacher when the time variable is 0. This will typically be designated as either the first observation score in a sequence or some other designated point in the time interval under consideration. The second parameter is π_{1p}, which represents the slope of a teacher-specific score trajectory. For both of these parameters it is possible to estimate "reliability" coefficients[2]

$$\rho(\pi_{0p}) = \frac{\sigma_0^2}{\sigma_0^2+\sigma_\varepsilon^2}, \text{ and} \qquad (5)$$

$$\rho(\pi_{1p}) = \frac{\sigma_1^2}{\sigma_1^2+\sigma_\varepsilon^2/SST_p} \ , \qquad (6)$$

where SST_p stands for "sum of squared time" and is computed as $\sum_{t=1}^{T}(X_{tp}-\bar{X}_p)^2$.

The reliability term $\rho(\pi_{0p})$ in Equation 5 is identical to the basic generalizability coefficient introduced in Equation 1, with $\sigma_0^2 = \sigma^2(\tau)$ and $\sigma_\varepsilon^2 = \sigma^2(\delta)$. It provides an indication

[2] We put reliability in quotes because, on the one hand, they are referred to in this way by Raudenbush & Bryk (2002) and these terms can be found as output when using HLM software. However, the term in Equation 5, $\rho(\pi_{0p})$, is better understood a generalizability coefficient in keeping with the G-Theory framework previously introduced. The sense in which $\rho(\pi_{1p})$ is a reliability coefficient in a classical sense of the term can also be called into question in the sense that is does not derive from the assumptions underlying Equation 1. The terminology is tricky and there is no one solution. In what follows we refer to estimates of both $\rho(\pi_{0p})$ and $\rho(\pi_{1p})$ as reliability coefficients without the quotes, but in each case they simply represent the proportion of the variance in observed score levels and score growth that can be attributed to real differences between teachers.

of the degree to which the scores on an observation protocol are generalizable over the particular sample of raters the single lesson that was used in computing the average score *Y* for subject *p* on occasion *t*. It is instructive to contrast the estimate for $\rho(\pi_{0p})$ that would result from a combined model based on equations 4a-4c

$$Y_{tp} = \beta_{00} + \beta_{10}X_{tp} + \beta_{10}r_{1p} + r_{0p} + \varepsilon_{tp} \quad (7)$$

to an unconditional model of repeated measurements model of the form

$$Y_{tp} = \beta_{00} + r_{0p} + \varepsilon_{tp}. \quad (8)$$

The difference between the combined model shown in equations 7 and 8 is the inclusion or exclusion of a linear time trend ($\beta_{10}X_{tp}$) and error component (r_{1p}). If both β_{10} and σ_1^2 (the variance component associated with r_{1p}) are significant, their exclusion will have an effect on $\rho(\pi_{0p})$ similar to that of the hidden occasion facet described in the previous section. So, although the HLM shown in equation 4a-4c and equation 7 is a fairly simple example of an augmented G-Theory model, and does not provide for the same design insights one could glean from a fully augmented G-Theory model as in Casabianca et al (2015), it does represent a potential improvement over the estimate of reliability one would generate if the only observable facets were lessons and occasions, and the two were confounded.

In contrast to the reliability term $\rho(\pi_{0p})$, although the term $\rho(\pi_{1p})$ in Equation 6 is also similar in structure to the Generalizability Coefficient in Equation 1, the underlying parameter in question is now growth in scores over time rather than level of scores at one point in time. The reliability of teacher growth estimates will be a function of three terms, the true variability in teacher growth, σ_1^2, the extent of measurement error in the scoring of teachers on each occasion $\sigma^2(\delta)$, and both the number and the spacing of distinct occasions when teachers are observed, summarized by SST_p. If all teachers are observed for the same number of occasions, with the

same spacing between occasions, then SST_p is a constant. When the number of occasions and spacing can both vary there is no single reliability coefficient for growth; rather, there is a unique estimate of reliability for each unique observation pattern, where each observation pattern can be conceptualized as a unique measurement design.

Gain scores based on the computation of score differences across two equally spaced time occasions, Y_1 and Y_2, represent a special case of the HLM characterized by equation 7. When only two time points are observed for each teacher, then $SST = 0.5$, and $\rho(\pi_{1p}) = \frac{\sigma_1^2}{\sigma_1^2+2\sigma_\varepsilon^2}$. Further intuition behind the factors that influence the resulting reliability of the slope parameter (note that with two time points the slope is just the difference between time points) can be appreciated by comparing Equation 6 to the equivalent[3] classical expression for the reliability of a difference score (Lord, 1956; Rogosa, Brandt, & Zimowski, 1982; Willett, 1989)

$$\rho(\pi_{1p}) = \rho(\mathrm{D}) = \frac{\sigma_{Y_1}^2\rho(Y_1)+\sigma_{Y_2}^2\rho(Y_2)-2\sigma_{Y_1}\sigma_{Y_2}\rho(Y_1Y_2)}{\sigma_{Y_1}^2+\sigma_{Y_2}^2-2\sigma_{Y_1}\sigma_{Y_2}\rho(Y_1Y_2)} \qquad (9)$$

where

- $\rho(D)$ is the reliability of a difference or gain score,
- $\sigma_{Y_1}^2, \sigma_{Y_2}^2$ are the variances of the observed scores at time 1 and time 2 respectively,
- $\rho(Y_1Y_2)$ is the correlation between the observed time 1 and time 2 scores, and
- $\rho(Y_1), \rho(Y_2)$ are the reliabilities of the time 1 and time 2 scores.

Inspection of equation 9 indicates that the reliability of a gain score depends primarily upon two things, the reliability of scores at each time point, and the correlation of these scores across the

[3] To verify this equivalence, note the assumptions under a linear error model that $Y_1 = t_1 + \epsilon_1$, $Y_2 = t_2 + \epsilon_2$, ϵ_1 and ϵ_2 are independent, $\sigma^2(Y) = \sigma^2(t) + \sigma^2(\epsilon)$ and $cov(Y_1, Y_2) = cov(t_1, t_2)$. It follows that in equation 6, $\sigma_1^2 = cov(t_2 - t_1) = \sigma_{t_1}^2 + \sigma_{t_2}^2 - 2cov(t_1, t_2)$.

two time points. The reliability of the gain score is inversely related to the correlation of the scores across occasions; as the correlation increases from -1 to 1, the reliability decreases. On the other hand, all else equal, the reliability of the gain score is always enhanced when the two scores used to compute a gain score are themselves reliable.

Choosing an Admissible Unit of Time

From the discussion in the preceding pages, it follows that an important decision point in any empirical investigation into the reliability of both the status and growth of teacher scores from an observation protocol is to come to some decision about the smallest admissible unit for temporal occasion (*t*). In the present study, we consider two different cases, one in which the smallest unit for *t* represents a week (i.e., small grain size), and another where the smallest unit for *t* represents a year (i.e., large grain size). When the smallest admissible unit of time is a year, any observations within that year will be averaged. This means that only two observations are possible, so only a gain score can be modeled. This choice brings to light some interesting tradeoffs when considered relative to Equations 9.

- If unique occasions are defined with respect to a small grain size, one has the opportunity to maximize the size of SST_p, but this will typically come at the expense of greater measurement error on each occasion because the score on each occasion will typically come from a single lesson (i.e., scores are not being averaged over multiple lessons).
- If unique occasions are defined with respect to a large grain size for the time unit, one is treating different lessons observed within that unit of time as exchangeable observations that can be averaged. This averaging will decrease measurement error due to the choice

of lesson, and this results in scores per aggregated occasion that will be more generalizable. However, some portion of this increased generalizability may be an illusion because now occasions within each year represent a hidden facet. Also, there will be fewer occasions available to model growth over time.

It is therefore somewhat of an open empirical question whether and when it is better to estimate teacher growth across many occasions with less precise measures, or across few occasions with more precise (though possibly biased) measures .

Data

The data used in this study comes from the MET project (Bill and Melinda Gates Foundation, 2013). The full study includes information on 2,741 fourth- through ninth-grade volunteer teachers at over 300 schools in six US school districts: Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City, New York; and Memphis, Tennessee (Cantrell & Kane, 2013). Data collection occurred during the 2009-2010 and 2010-2011 school years. The MET project design required teachers to submit four classroom videos per subject between February and June of 2010 for Year 1, and then another four more videos per subject between October 2010 and June 2011 for Year 2. In total then, the duration for which classroom observations were available spanned about 16 months over two calendar years, or a little less than one and half academic school years. Excluding the three summer months of 2010, total duration over which observations were collected spanned 50 weeks. Project researchers encouraged teachers to spread the video recordings over time within each year to ensure that the recordings were more representative of instruction than a series of

closely timed lessons. However, there were no constraints set on the amount of time required between each video occasion (Bill and Melinda Gates Foundation, 2013). Each of these videos received scores from multiple observation protocols, but we focus our attention on just one of these, the Danielson Framework for Teaching.

The Danielson Framework for Teaching

Observation protocols can be broadly divided into two categories: those that are subject-specific and those that are subject-neutral. Some examples of subject-specific protocols include the Mathematical Quality of Instruction (MQI; Hill et al., 2008), the Reformed Teaching Observation Protocol (RTOP; Piburn & Sawada, 2000) for mathematics and science, the UTeach Observation Protocol for math and science (UTOP; Marder et al., 2012), the Protocol for Language Arts Teaching Observation (PLATO; Grossman, Loeb, Cohen, & Wyckoff, 2013), and the TEX-IN3 for literacy (Hoffman, Sailors, Duffy, & Beretvas, 2004). These protocols were designed to capture information about content-specific elements of classroom practices such as the richness of the content as present in the lesson or the teacher's pedagogical knowledge directly related to the content area. Other elements include the extent to which the teacher values and prioritizes a broad spectrum of material within the content area. For example, the extent to which the teacher includes multiple types of text in a language arts class as opposed to favoring only narrative text or the way a math teacher presents a concept in varied contexts as opposed to only one presentation of rote procedure.

Conversely, protocols which are subject-neutral focus on more general elements of teaching. Examples of these include the Charlotte Danielson Framework for Teaching (FFT;

Danielson, 2013), the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2006), the Sheltered Instruction Observation Protocol (SIOP, Echevarria, Vogt, & Short, 2008), and the Marzano Observation Protocol (Marzano, 2007). Each of these protocols are designed to be used across subjects and include scoring elements related to teacher practices such as clear communication between students and teachers, positive classroom environment, and meaningful assessment and feedback practices.

We use the FFT in the current study for two reasons. First, a subject-neutral protocol maximizes the number of potential data points for any given teacher in the MET data. Over the course of the MET project, teachers provided up to eight videos for each subject taught. Elementary teachers, who instructed the same group of students in multiple subjects, had the opportunity to provide up to 16 videos over the course of the two-year study: a maximum of eight in English Language Arts (ELA) and a maximum of eight in math. Secondary teachers provided a maximum of eight videos in one subject or the other. Since scores on the FFT are agnostic to course content, all scores, regardless of content, are available for building longitudinal growth trajectories. Second, there is one version of the FFT used across grades 4-9 while the CLASS has two forms, one for grades 4-5 and another for grades 6-9. Using the FFT allows for inclusion of all grades in the study. Although we use the FFT to explore the use of observation scores over time, the methodological issues we explore here apply to the use of scores from any observation protocol.

The FFT as used in the MET project included eight dimensions divided into two domains. Each video in the project received a score on every dimension, and each dimension has its own scoring rubric with exemplars of the teacher-student interactions that would lead to a

score in one of four ordered categories[4]. The eight dimensions that are scored by raters appear in Table 1. Figure 3 illustrates the distribution of scores for each of the eight dimensions on the FFT across all occasions for the data used in this study.

Insert Table 1 and Figure 3 here

As evident in Figure 3, most teachers received scores of two or three for all of the dimensions of the FFT. Another way of considering these scores is as the average dimension-level score per occasion. In other words, for each teacher by occasion, we calculated the average of all eight dimension-specific scores. The mean average dimension-level scores across all teachers and occasions is 2.5 and the standard deviation (SD) is 0.47. In the analysis that follows this section, our we use the average FFT score across all eight dimensions as our outcome of interest[5].

When scoring the FFT, raters viewed the first fifteen minutes of each video and then skipped to viewing minutes 25 through 35. After viewing both segments, raters gave one score per item for the video overall. Raters were randomly assigned to teachers within randomized blocks of classrooms within schools and only scored a single video for a given teacher in each year. That is, the most a single rater scored a single teacher was twice over both years of the study, so in each year, unique teacher observations are scored by multiple raters. However, no procedures ensured that a rater scored a specific teacher twice in the study over the two year span. Furthermore, only about 5% of videos received scores from two raters. For each video that

[4] For details on the scoring criteria used in these rubrics, see Danielson, 2011.
[5] Previous studies (e.g., Hill et al., 2012; Praetorius et al, 2014) have found differences in estimates of reliability by different domains and dimensions of observation protocols, but this is not an issue we examine here.

was double-scored, one randomly selected set of scores appears in the final data set for estimating growth trajectories in this project. All the rest of the videos only received scores from one rater per protocol. Because nearly every score assigned to a given teacher came from a unique rater and only a very small number of the videos received scores from two raters, we could not decompose and estimate rater facets of variance. This represents a significant limitation on the empirical portion of this study since we know from past research that raters can differ considerably in their scoring practices. On the other hand, each teacher in our study is in fact cumulatively scored by multiple raters, and these raters were allocated to teachers at random. As a result, any time FFT scores are averaged across occasions (i.e., when we use year as the smallest admissible unit for time), we are implicitly averaging across multiple raters as well.

Teacher Demographics

Of the original 2,741 teachers involved in the project at large, 1,569 received FFT scores, but only about 953 teachers had dates associated with their videos. In addition, approximately half of these 953 teachers did not have values for a variable indicating their years of experience. This is because two of the districts participating in the MET study did not provide this information (Kane & Staiger, 2012). We drop cases in which there is no information regarding a teacher's years of experience[6]. As a result, the final dataset used to conduct the analysis includes 458 teachers who provide 3372 unique videos.

[6] The data used in this study come from a prior project in which this sample restriction was imposed.

The information in Table 2 provides a summary of the demographic information available for the teachers used in the study compared to those without data about their years of experience as well as those without dates associated with their videos. These numbers suggest that the group of teachers in this study are more likely to be male and white and less likely to have a master's degree or higher than the other teachers eliminated from the sample due to missing information. Further, about half of the teachers without information about experience and 40% of those without dates attached to their videos taught elementary school, while only about one quarter of the teachers in the current study taught elementary school. This means that most teachers in the study would have been required to submit only four videos per year. Thus, although it was possible to submit 16 videos per teacher, the average number of videos available per teacher is 8 due to the grade-level make-up of our teacher sample.

Insert Table 2 about here

Timing of Observation Protocols

Since teachers in the MET project had autonomy in choosing the number and spacing of their video recordings, each teacher represents a different potential observation design, which has implications for the reliability of growth estimates. The variability in the number and spacing of occasions, represented by differences in the variable SST_p, distinguishes each unique design. We use X_{tp} as our time variable, defined according to two different admissible units, week and year. When year is used, X_{tp} takes on one of two values, 0 for year 1, and 1 for year 2. The FFT score Y_{tp} associated with a year is the average of all FFT scores collected on unique occasions within

the year. When week is used as the smallest admissible time unit, a value of 0 indicates the first week of the MET project, and values can range from 0 to 50. Weeks are numbered across both years of the project for which data was collected, excluding the summer in between.

The frequency distribution of occasions across our teacher sample is displayed in Figure 4. For most teachers, 7 or 8 observations were available (70% of the sample), and at least 4 or more observations were available for 91% of the sample. The small number of teachers for whom only a single observation is available contribute no information toward the estimation of a growth parameter and associated variance component (i.e., see equation 4c), but they do contribute to the estimation of the variance component associated with score levels (i.e., see equation 4b).

Insert Figure 4 about here

Results

HLM Parameter Estimates

Table 3 presents the results from two different approaches to defining an admissible unit of time. When the smallest admissible unit is a week, we call the HLM a "Growth Trajectory Model;" when it is a year, we call it a "Gain Score Model." Both models include an estimate for an intercept that indicates the mean FFT score in the first year of the MET project across all teachers (β_{00}). This estimate, which is about the same and statistically significant for both models, indicates that average FFT score for the first year in the Gain Score Model and for the

first week in the Growth Trajectory Model was about 2.5. Table 3 also provides fixed effect estimates for the slope (β_{10}), which represents the mean change in FFT score for the Gain Score Model, and the mean rate of growth per week for the Growth Trajectory Model. Again, the values in both models are similar at about 0. Unlike the downward trend found for CLASS scores (Casabianca et al., 2015) on average, our subsample of MET project teachers shows no evidence of an overall trend in FFT scores, irrespective of the grain-size for an admissible unit of time.

The results indicate that growth trends vary significantly among teachers, and here the results do differ by modeling approach. Examination of the variance component estimates provided in Table 3 indicate statistically significant variation in mean FFT scores at the beginning of the MET project for both models, but variance in growth is only statistically significant in the Growth Trajectory Model, and for this model the negative correlation between base score and growth ($-.32$) is almost three times as that estimated for the Gain Score Model ($-.13$).

Variability in the distribution of growth across teachers by model can be further interpreted relative to the SD of $\hat{\sigma}_1$. This value is .10 for the Gain Score Model and .006 for the Growth Trajectory Model. Under the assumption that teacher-specific growth parameters are normally distributed, 95% confidence intervals around these fixed effects for each model would be about $[-.19 \text{ to } .21]$ and $[-.012 \text{ to } .012]$ respectively. The smaller magnitude for the Growth Trajectory Model needs to be interpreted relative to the time span under consideration, which can be up to 50 weeks. Hence the boundaries for growth under the Trajectory Model would be $[50 * -.012 = -.60]$ to $[50 * .012 = .60]$ which is quite a bit larger than the interval for the Gain

Score Model. Next, we turn to a comparison of the reliability of score level and score growth parameters in the two models.

Insert Table 3 about here

Reliability of Gain Score and Growth Trajectory Parameters

We begin by comparing the reliability/generalizability of status estimates (i.e., the intercept) under the Gain Score and Growth Trajectory Models. The estimated reliability for β_{00} under the Gain Score Model is $\rho(\pi_{0p}) = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_0^2+\hat{\sigma}_\varepsilon^2} = \frac{.07}{.07+.04} = .64$. The estimated reliability for β_{00} under the Growth Trajectory Model is $\rho(\pi_{0p}) = \frac{.08}{.08+.14} = .36$. The most evident explanation for the higher (though still fairly low in an absolute sense) reliability estimate for the Gain Score Model comes from the fact that the intercept for each teacher in this approach is based on the average of FFT scores across up to eight combined lessons and occasions and multiple distinct raters. In contrast, the intercept in the Growth Trajectory Model is based on just a single lesson, occasion and rater for each teacher. The difference in number of lessons over which FFT scores have been averaged is reflected in the much larger magnitude of estimated error variance for the Growth Score Model ($\hat{\sigma}_\varepsilon^2 = .14$) relative to the Gain Score Model ($\hat{\sigma}_\varepsilon^2 = .04$).

It is interesting to note that the estimate for universe score variance in teacher status scores is about the same under either approach, and this does not appear to support the hypothesis, presented earlier, that the hidden facet of occasion would introduce a positive bias in the estimation of a generalizability coefficient. Indeed, we can see this more explicitly if we compare the reliability of score levels under the Growth Trajectory Model to the reliability one

would estimate had we kept weeks as the admissible unit for time, but excluded the linear growth trend (recall equation 8). For this model, we find that $\hat{\sigma}_{\varepsilon}^2 = .15$, $\hat{\sigma}_0^2 = .07$, and $\rho(\pi_{0p}) = .32$. Though it is only by a small amount in this case, exclusion of the linear time trend and variance component appears to bias the estimated reliability of score levels downward from .36 to .32.

The situation changes when we compare the reliability of growth parameters across the two models. For the Gain Score Model, the estimated reliability for β_{10} is $\rho(\pi_{1p}) = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2+2\hat{\sigma}_{\varepsilon}^2} = \frac{.01}{.01+2(.04)} = .11$. The computation of a single reliability coefficient for the Growth Trajectory Model is more complicated since there is a unique reliability associated with each unique SST value. The mean SST for the full dataset is 1113, and the SD is 596. SSTs range from 0 to 3287. An SST of 0 results either from a measurement design with only one occasion or a design in which all of the occasions occur within the same week. The average designs included about eight occasions with an SD of about 13 weeks between occasions. For a design associated with the mean SST of 1113, the estimated reliability for β_{10} is $\rho(\pi_{1p}) = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2+\frac{\hat{\sigma}_{\varepsilon}^2}{SST_p}} = \frac{.00004}{.00004+.14/1113} = .24$. Although the reliability is low, perhaps unacceptably so, it tends to be twice as large for the Growth Trajectory Model. When teacher growth is the parameter of interest, the MET data provides an empirical example where it is better to compute a slope on the basis of eight noisy measures over a two year period than it is to compute a gain on the basis of two measures, each of which are relatively more precise.

Reliability of Growth Trajectories under Hypothetical Design Scenarios

The mean reliability for growth across all observation designs in the MET sample is 0.24, which is low. If one could fix the SST by requiring the same number of observations and

spacing between the observations over a full two year academic school year with 78 weeks[7] instead of 50, what could be predicted for the reliability of growth estimates under different design scenarios? Table 4 shows how reliability would be predicted to increase as observations increase from 4 to 10 over 78 weeks. For each hypothetical observation pattern, the first and last observations are fixed to take place 20% and 80% of the way through the span of weeks (at the 16th week and the 63rd week respectively). As the number of observations increase, they are added symmetrically within the range between weeks 16 and 39 (year 1) and 47 and 78 (year 2). As observations are increased from 4 to 10 in increments of 2, the respective reliability of the teacher growth parameter increases from .26 to .33 to .39 and tops out at .44. Note that these values are all premised on equal spacing between observations; if observations tended to be clustered together in a smaller range of weeks, the reliability estimates would be lower.

Insert Table 4 about here

The results from our analysis of MET project data suggest that given the context in which those observations were gathered, that even in a best case scenario with 10 observations over two years, only about half of the variability in growth trajectories would be attributable to "real" differences among teachers. The range of these best-case scenario values is still low, but it is perhaps worth pointing out that they are in the same ballpark as estimates of reliability and stability reported for the estimation of teacher effects in the context of student achievement using value-added models (Kane & Staiger, 2012; McCaffrey, Sass, Lockwood, & Mihaly, 2009).

[7] We get an estimate of 39 weeks per academic year by adding the calendar days from September through June and diving by 7.

That is, whether a teacher growth statistic is computed on the basis of aggregating student-level growth estimates, or from direct estimates based on growth on observation scores, the estimates will have limited reliability. But to attain an estimate of reliability close to .4 would take two years and eight instances of data collection in the context of an observation protocol, while an estimate based on value-added requires only a single data collection event annually.

Discussion

The role of occasions in the context of the design and use of teacher observation protocols is one that merits careful consideration. The scores from observation protocols are often implicitly assumed to be generalizable across the different temporal occasions of the observations. This hinges upon the exchangeability of the temporal occasions. We argue here that an ideal study of the generalizability of scores from observation protocols would attempt to disentangle the variability due to choice of occasion from the variability due to choice of lesson observed on that occasion. Yet even if this were to be done, if teaching practices change systematically over time, an analytic approach based on a conventional G-Theory variance decomposition would need to be modified to take these trends into account. In this paper we have explored, empirically, an alternate way of conceptualizing the use of teacher observation protocol scores, one that is premised not so much on a desire to generalize over measurement occasions within a school year, but on a desire to model growth that may be occurring across these occasions. We contrasted two ways this could be done, one that conceptualizes the smallest unit of time as a year, and another where the smallest unit of time is a week. In the latter case, we are building on the advice from David Rogosa and John Willett in the 1980s, who suggested

that one might be well-served to focus on the estimation of growth trajectories from many observations than from the estimation of a gain score from just two observations. What is lost in precision at each time point is made up for having more information spread over time.

Our empirical findings using the MET project data show, not surprisingly, that the reliability of observation protocol score levels (i.e., status) are much more reliable than score growth. Score levels that were averaged across multiple lessons, occasions and raters were almost twice as reliable (.64 vs. .36) as those based on just a single combined lesson and occasion and a single rater. In this instance it does not appear that variability due to occasion introduces a bias due to an inflation in the decomposed estimate of true score variance. In contrast, when observation score growth is of interest, estimates that derive from treating a week as the smallest meaningful distinction in time were about twice as reliable (.24 vs. .11) as those that treated the school year as the smallest meaningful distinction. We found that given the variance decomposition of the MET data, a best-case scenario for designing a system of observation protocols over the course of a two year span would maximize reliability at a value somewhat below and above .40 for a total of 8 to 10 observations. These results allow for some understanding of how the choices around number of occasions as well as number of raters affects reliability of estimates of growth in teacher practices. Two key implications are 1) that priority should be given to adding observation occasions and maximizing the spacing of those occasions, and 2) if true growth variance is very small as appears to have been the case for the teacher sample in the MET project, then it will take a significant number of observations over two years before it makes much sense to consider making distinctions among teachers with respect to their growth in classroom practices.

These findings suggest it would be very difficult to defend the use of growth in observation protocol scores as a basis for high-stakes inferences or decisions about individual teachers are part of a system of teacher evaluation. We say that the estimates from this study represent a best-case scenario because among teachers and schools participating in the MET project, there were no stakes associated with classroom observations and there was no personal relationship between teachers and raters. That is, the raters had no personal investment in the outcome of the observation scores, and the raters felt no pressure that their scores would be used to make high-stakes decisions. It is well-known that the imposition of these stakes has the potential to distort the process raters use to assign scores (Campbell, 1976). The impact of this on the variance components that figure into estimates of reliability are unknown.

On the other hand, it can be argued that our results are not necessarily a best case scenario. To begin with, we know that rater drift can introduce bias into teacher observation scores (Casabianca et al., 2015), and this is not an issue we were able to examine in the empirical portion of this study. Next, we have have assumed that the FFT observation protocol scores are not systematically higher or lower by the subject of the lesson being observed (math or ELA). Second, we have ignored the impact that different cohorts of students may have when they are used as the basis for observations collected across school years. If these three factors could be either better controlled or modeled, this could possibly lead to higher estimates for the reliability of both score level and score growth[8].

Even if our results do represent a best case scenario, there might still be some promise to the use of observation protocols to model teacher growth over time, provided that one is focusing

[8] We thank an anonymous reviewer for bringing these last two points to our attention.

on a longer term time horizon and primarily interested in comparing group averages, or only making distinctions among teachers in the tails on the growth distribution. For example, a longitudinal trajectory of observation scores could be useful in studying the efficacy of professional development programs or interventions for early career teachers over their first 3 to 5 years on the job. Ultimately however, as Bell et al., 2012 remind us, such uses will hinge upon more than issues of generalizability and reliability, but will depend upon establishing a more fully persuasive and comprehensive validity argument.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. w23478). National Bureau of Economic Research.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher*, *44*(2), 77-86.

Bill and Melinda Gates Foundation. (2013). Measures of Effective Teaching: 1 - Study Information. ICPSR34771-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2013-09-23. http://doi.org/10.3886/ICPSR34771.v2

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Blazar, D., & Kraft, M. A. (2015). *Teacher and teaching effects on students' academic behaviors and mindsets*. Mathematica Policy Research.

Brabeck, M. (2014). Assessing and evaluating teacher preparation programs. American Physiological Association Task Force Report.

Braun, H., Chudowsky, N., & Koenig, J. (2010). Getting value out of value-added. *Social Sciences. Washington, DC: National Academies Press*.

Brennan, R. L. (2001). Multifacet Universes of Generalization and D Study Designs. In *Generalizability Theory* (pp. 95-139). Springer, New York, NY.

Campbell, D. T. (1976). Assessing the impact of planned social change. The Public Affairs Center, Dartmouth College, Hanover New Hampshire, USA.

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. *Policy and Practice Brief. MET Project. Bill & Melinda Gates Foundation*.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783. https://doi.org/10.1177/0013164413486987

Charalambous, C. Y., Kyriakides, E., Tsangaridou, N., & Kyriakides, L. (2017). Exploring the reliability of generic and content-specific instructional aspects in physical education lessons. *School Effectiveness and School Improvement*, *28*(4), 555-577.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633-79.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Cronbach, L., Gleser, G.C., Harinder Nanda, A.N., and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley & Sons, Inc.

Danielson, C. (2011). *The framework for teaching: Evaluation instrument*. Princeton, NJ: Danielson Group.

Danielson Group. (2013). Charlotte Danielson. Retrieved from https://www.danielsongroup.org/charlotte-danielson/

Echevarria, J., Vogt, M., & Short, D. (2008). Making content comprehensible for English learners: The SIOP model.

Everson, K. C. (2017). Value-added modeling and educational accountability: Are we answering the real questions?. *Review of Educational Research*, *87*(1), 35-70.

Goe, L., Bell, C., & Little, O. (2008). Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. *National Comprehensive Center for Teacher Quality*.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, *44*(2), 87-95.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589-612.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education. 119*(3), 445-470.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267-71.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education finance and policy*, *4*(4), 319-350.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, *26*(4), 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. http://doi.org/10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Hoffman, J. V., Sailors, M., Duffy, G. R., & Beretvas, S. N. (2004). The effective elementary classroom literacy environment: Examining the validity of the TEX-IN3 observation system. *Journal of Literacy Research*, *36*(3), 303-334.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Lord, F. M. (1956). The measurement of growth. *ETS Research Bulletin Series*, *1956*(1), i-22.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores: Some la- tent trait models and their use in inferring an examinee's ability*. Reading, MA: Addison-

Wesley.

Marder, M., Walkington, C., Abraham, L., Allen, K., Arora, P., Daniels, M., & Walker, M. (2010). The UTeach Observation Protocol (UTOP) training guide (adapted for video observation ratings). *Austin, TX: UTeach Natural Sciences, University of Texas Austin*.

Marzano, R. J. (2007). *The art and science of teaching: A comprehensive framework for effective instruction*. ASCD.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400-422.

McCaffrey, D. F., Han, B., & Lockwood, J. R. (2009). Turning student test scores into teacher compensation systems. In M. Springer (Ed.), *Performance incentives: Their growing impact on American K-12 education* (pp. 113-147). Washington, DC: Brookings Institution Press, 2009.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability. Monograph*. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138.

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.

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., La Paro, K. M., & Hamre, B. K. (2006). Classroom assessment scoring system: Manual k-3 version. *Charlotesville, VA: Center for Advanced Study of Teaching and Learning, University of Virginia*.

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & Paro, K. M. (2007). Classroom Assessment Scoring System (CLASS), secondary manual. Charlottesville: University of Virginia Center for Advanced Study of Teaching and Learning.

Piburn, M., & Sawada, D. (2000). Reformed Teaching Observation Protocol (RTOP) Reference Manual. Technical Report.

Praetorius, A., Pauli, K., Reusser, C., Rakoczy, K., K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175-214.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

Sanders, W.L., Saxton, Am>m, & Horn, S.P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In J.

Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve?. *Economics of Education Review, 64*, 50-74.

Willett, J. B. (1989). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.

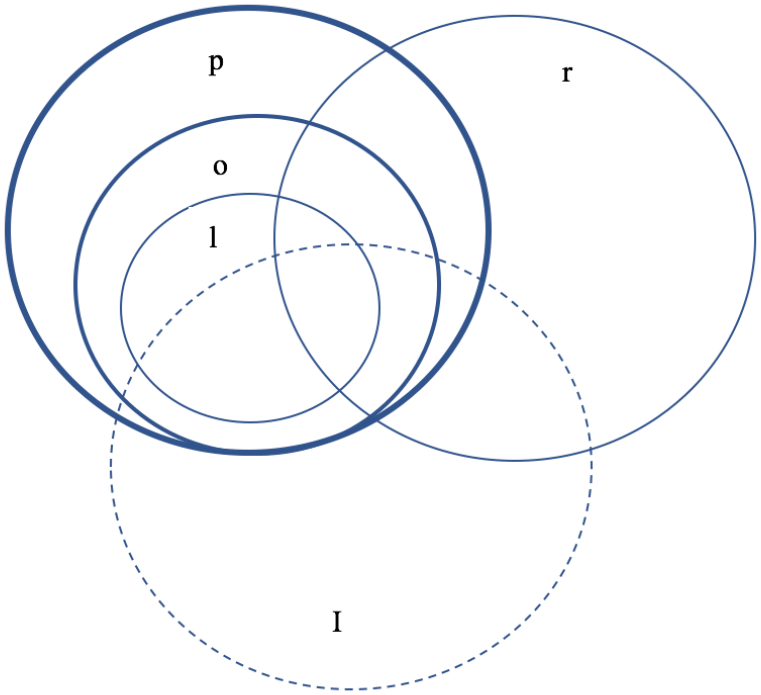


Figure 1. Venn Diagram for the Ideal l:o:p x r x I Design of an Observation Protocol

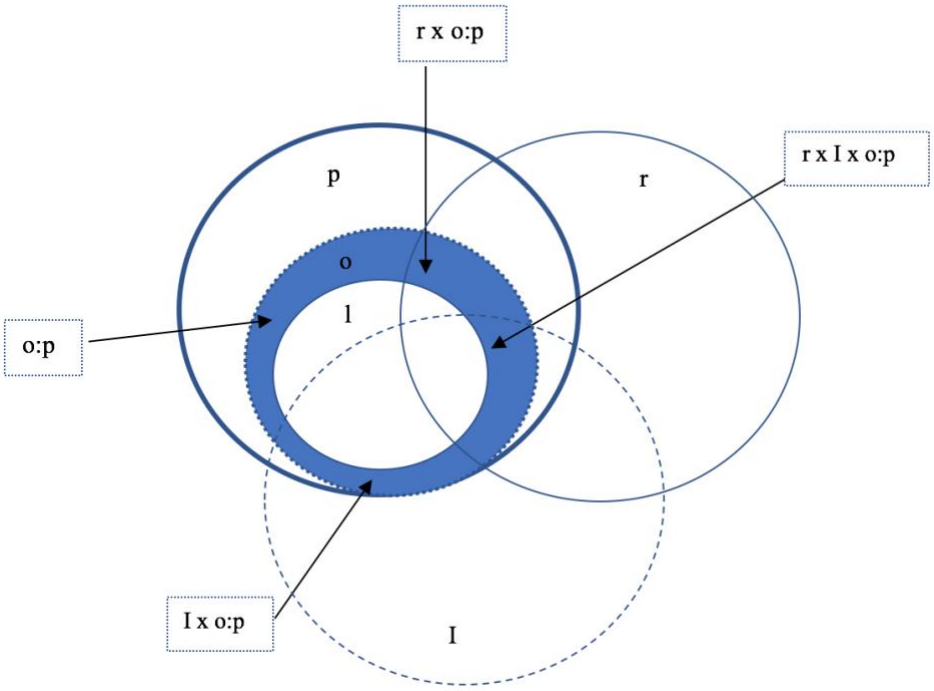
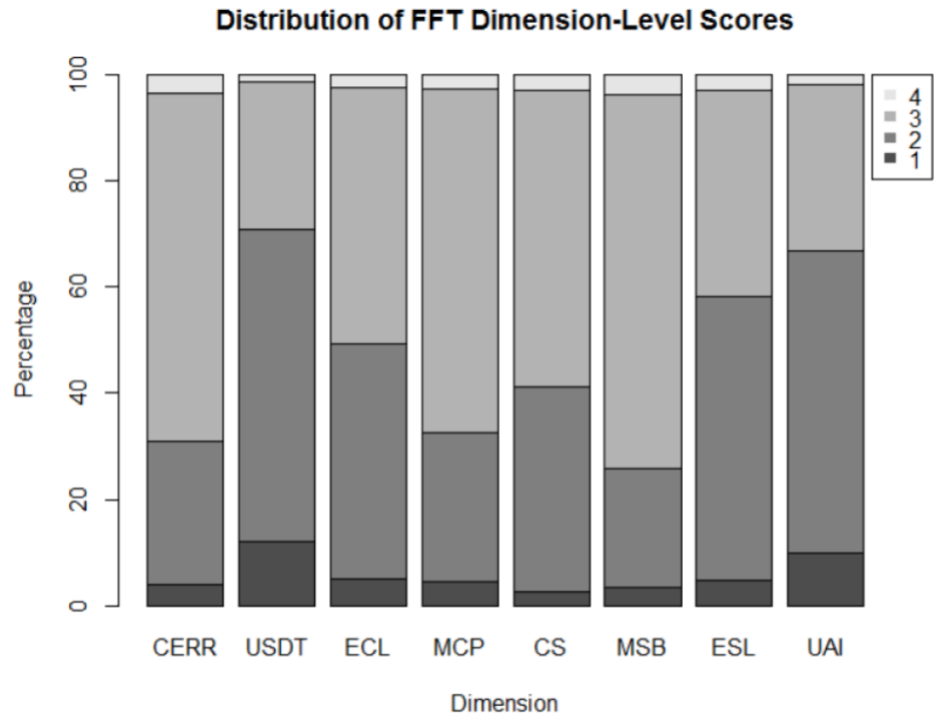


Figure 2. The Hidden Occasion Facets in the Realized l:p x r x I Design



Creating an environment of respect and rapport (CERR)
Establishing a culture for learning (ECL)
Managing classroom procedures (MCP)
Managing student behavior (MSB)
Communicating with students (CS)
Using questioning and discussion techniques (USDT)
Engaging students in learning (ESL)
Using assessment in instruction (UAI)

Figure 3. Distribution of FFT Dimension-Level scores

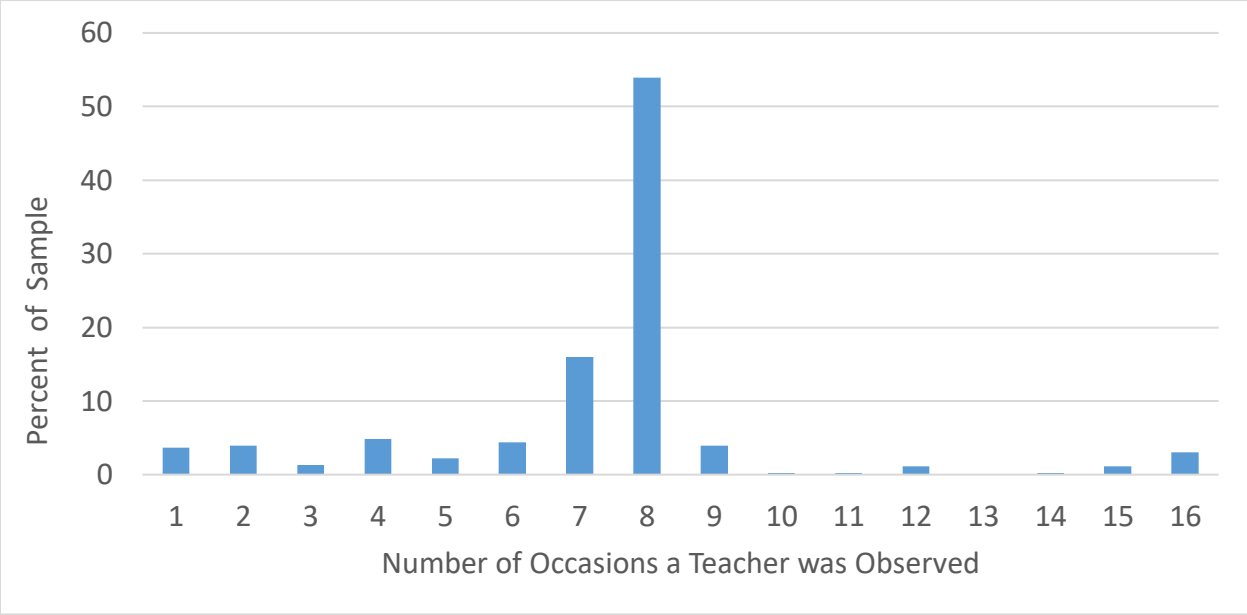


Figure 4. Frequency Distribution of Observation Occasions Across Teachers (N=458)

Table 1. FFT Domains and Dimensions

Domains	Dimensions
Classroom Environment	• Creating an Environment of Respect and Rapport • Establishing a Culture for Learning • Managing Classroom Procedures • Managing Student Behavior
Instruction	• Communicating with Students • Using Questioning and Discussion Techniques • Engaging Students in Learning • Using Assessment in Instruction

Table 2. MET Project Teacher Demographics

	Dates Unavailable	Experience Unavailable	Analytic Sample
Male	17	16	22
White	53	53	65
Black	36	36	25
Masters +	27	33	23
Elementary (Grades 4-5)	40	47	26
Secondary (Grades 6-9)	60	53	74
Novice	6	--	16
N	616	495	458

Note: Values in rows 1-7 are percentages, values in last row are total count.

Table 3. Comparisons of HLM Results

		Gain Score Model		Growth Trajectory Model	
Fixed Effects		Coefficient	SE	Coefficient	SE
Intercept					
	β_{00}	2.48	0.02	2.49	0.02
Slope					
	β_{10}	0.01	0.02	0.0001	0.0006
Random Effects		Variance Component	p-value	Variance Component	p-value
Level 1: e_{ti}		0.04	--	0.14	--
Level 2					
Intercept: r_{0i}		0.07	<0.001	0.08	<0.001
Slope: r_{1i}		0.01	0.08	0.00004	<0.001
Corr (β_{00}, β_{10})		-0.13		-0.32	
N		441*		458	

*17 teachers only provided multiple observation scores in one of the two years. These teachers would be included in the growth trajectory model but excluded from the gain score model because their scores would be collapsed to a single observation.

Table 4. Reliability of Growth Parameters Generalizaed to Different Designs over full 2 Years (78 Weeks)

Observations	Weeks	*SST*	Reliability of Growth
4	16, 32, 47, 63	1217	.258
6	16, 24, 32, 47, 55, 63	1697.5	.327
8	16, 20, 24, 32, 47, 51, 55, 63	2202	.386
10	16, 20, 24, 28, 32, 47, 51, 55, 59, 63	2722.5	.438

Note: *SST* stands for "sum of squared time" and is computed as $\sum_{t=1}^{T}\left(X_{tp} - \bar{X}_p\right)^2$. Reliability computed using Equation 6.