Abstract Title Page

Title:
Variability in preschool teachers' interactions with children as a predictor of children's development

Author(s):
Timothy W. Curby
Laura L. Brock
Bridget K. Hamre

Abstract Body

Background / Context:

Most often, when researchers have multiple time points of data, scores are averaged to yield a more-reliable estimate of the construct being measured. As statistical methods have been developed such as growth modeling (Raudenbush & Bryk, 2002), researchers have increasingly capitalized on the multiple time points of data by modeling growth over time. One under-utilized way of looking at longitudinal data is to look not just at the mean, but also at the variability around that mean. So, from this view, if a person has multiple time points of data, not only can a mean be created but also a standard deviation. From this vantage point, it is not simply the mean that is considered a characteristic of a person that is measured, but also how much that individual varies over time. Most notably, from a traditional measurement perspective, this variability would be considered 'error' and therefore should not be considered. However, it is reasonable to believe that some people are more variable than others. For example, some people may be more emotionally labile than others even though their averages could be similar.

The idea of variability is not foreign to researchers. Researchers commonly think about the standard deviation as a measure of variability. However, even when it is reported, variability is conceived of as a between-person characteristic. In the present study, we look at variability as a within-person characteristic whereby each person has a standard deviation calculated based on the multiple time points s/he was observed. In other words, we propose that we should not just think of the quality of interactions as a characteristic of the teacher, but that variability may be a teacher characteristic as well. Thus, this poster will present analyses that differentiate those teachers who have brief, high quality interactions with those who have continuous, low quality interactions.

This view of variability has been used to a small degree in the research literature, but with important implications. Using a standard deviation as a predictor has been done in other published studies such as when Eizenman, Nesselroade, Featherman, & Rowe (1997) used variability in perceived control to predict mortality, when Butler, Hokanson, & Flynn (1994) used variability in self-esteem to predict depression, and when Kernis, Grannemann, & Mathis (1991) used variability in self-esteem as a moderator of the relation between self-esteem and depression. The proposed poster extends this work into the classroom to examine the variability in the quality of teachers' interactions with students.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this proposal is to examine whether variability in the quality of teachers' interactions (Emotional Support, Classroom Organization, Instructional Support) with students is systematically related to the children's development. In other words, we examine whether the amount that teachers vary over the course of a day is a characteristic of the teacher by seeing if there are systematic associations between teachers' variability and children' s development. If variability is simply error, we would not expect the associations to be systematically related to children's development. However, if a pattern emerges, it would support our notion that variability is itself characteristic of the classroom.

Setting:

Data for the present study were collected by the National Center for Early Development and Learning over two waves. The first wave, called the Multi-State Study of Pre-Kindergarten (MS Study), collected data from six states in the 2001-2002 academic year. The second wave of

data collection, the State-Wide Early Educations Programs (SWEEP) Study, took place in the 2003-2004 academic year. Because these two studies shared basic study designs and measures they were aggregated into one dataset for use in the present study, as is commonly done with these datasets (e.g., Burchinal, Vandergrift, Pianta, & Mashburn, 2010).

Population / Participants / Subjects:
701 classrooms were randomly selected with 694 providing observational data of classroom interactions. For the observational measure reported on herein, the MS Study trained coders conducted observations for two days in the fall and two days in the spring. For the SWEEP Study, observations were conducted once during the spring. For more information on the selection procedures and datasets, please see the NCEDL website (http://www.fpg.unc.edu/~ncedl/).

Data were collected on children in classrooms. Selected children included those who were eligible for kindergarten the next year, could follow age-appropriate directions in either English or Spanish, did not have an individual education plan, and those whose parents provided consent. When possible, two boys and two girls were randomly selected for participation. A total of 2938 children participated in the two studies (see Early et al., 2005).

Consistent with other work using these data (Mashburn et al., 2008), the present study excluded 499 children and 39 classrooms from analyses because they either did not participate in the spring assessment or were assessed in Spanish. Thus, the final sample for the present study included 2439 child participants. These children were on average 4.62 years old at the time of the fall assessment. Of the 2439 participants, 1758 had teacher-reported competence data and 1776 had teacher-reported problem behavior data available the following kindergarten year.

Research Design:
Secondary analysis

Data Collection and Analysis:
Measures used in Data Collection
The *Classroom Assessment Scoring System* (CLASS; La Paro, Pianta, Hamre, & Stuhlman, 2002; Pianta, La Paro, & Hamre, 2008) was used to provide a measure of the quality of interactions that teachers offer students. Classrooms were generally observed for several cycles on an observation day (*Range* = 1 - 7, *Mean* = 6.57). An observation day was defined for full day programs as being until nap and, for half-day programs, until students left for the day. Each observation cycle consisted of a 20-minute observation followed by a 10 minute rating period. During each rating cycle, nine dimensions of quality in teachers' interactions with children were coded. Each dimension is scored on a Likert scale from 1 *low* to 7 *high*. Three domains are subsequently formed from nine dimensions (Hamre & Pianta, 2007): Emotional Support, Classroom Organization, and Instructional Support. *Emotional Support* is the composite of four observed dimensions (α = .84): Positive Climate, Negative Climate, Teacher Sensitivity, and Over Control. *Classroom Organization* was composited from three observed dimensions (α = .82): Behavior Management, Productivity, and Instructional Learning Formats. *Instructional Support* was measured by two dimensions (α = .77): Concept Development and Quality of Feedback.

Training & Reliability. All raters attended a two-day training provided by the developers of the CLASS instrument. To be deemed reliable, raters had to be within 1 scale point of the

master coded score on 80% of the dimensions across five twenty-minute video segments. All raters met or exceeded this criterion of reliability. At spring, raters' reliability was again tested by dual coding in a classroom with a master coder. Raters' mean kappa was .73, with 93% of ratings within one scale point of the master coder.

Variability. Average within-day Variability was calculated for each CLASS domain (Emotional Support, Classroom Organization, Instructional Support) in a multistep process. The goal was to produce a variable in standard deviation units that could be entered as a predictor. Computationally, this was accomplished by first computing the variance for *each day* of observation for Emotional Support, Classroom Organization, and Instructional Support. Next, the average within-day variance was calculated by taking the mean of the within-day variances. The square root of the variances was calculated to convert them into standard deviations. Thus, analyses were conducted with standard deviations of Emotional Support, Classroom Organization, and Instructional Support as predictors.

Child Demographic Information

Demographic data were collected from the family using a short questionnaire. These data included the child's gender and ethnicity (White, Hispanic, Black, multiracial/other), information about family income (converted into a dichotomous variable to indicate whether or not the family was below 150% of the poverty line for the family size), and years of maternal education.

Academic Outcomes

The *Peabody Picture Vocabulary Test—3rd edition* (PPVT-III; Dunn & Dunn, 1997) is a widely-used measure designed to capture children's receptive vocabulary. It was administered in the fall ($\alpha = .95$) and spring ($\alpha = .95$) of their pre-k year. During the assessment, an examiner provides children with a set of four pictures and says a word corresponding to one of the pictures. Children are asked to point to the picture that best represents the word spoken by the examiner.

The Oral & Written Language Scale (OWLS; Carrow-Woolfolk, 1995) is designed to assess expressive language of individuals age 3 - 21 and was administered in the fall ($\alpha = .90$) and spring ($\alpha = .90$) of children's pre-k year. During the assessment, an examiner read a verbal stimulus aloud while the child looked at a card containing at least one picture. Children orally responded by answering a question, completing a sentence, or generating a new sentence.

Two subtests were used from the *Woodcock-Johnson III Tests of Achievement*: Rhyming and Applied Problems (Woodcock, McGrew, & Mather, 2001). During the Rhyming subtest, children are presented with a word and are asked to name a word that rhymes. The Rhyming scale has a range of 0-17, and is not standardized. The Applied Problems subtest was also administered in the fall ($\alpha = .81$) and spring ($\alpha = .82$) of the study years. In the subtest, children are provided with a number of orally-administered mathematics problems on quantity, simple addition and subtraction, and concepts of time and money.

As a measure of emergent literacy children's ability to recognize letters of the alphabet was determined. During this assessment, children were asked to identify a mix of capital and lowercase letters. The highest possible score is 26.

Social Outcomes

The Teacher-Child Rating Scale (TCRS; Hightower et al., 1986) is a 38 item, teacher-report measure of children's social competence and problem behaviors. On each item, teachers assessed how well a phrase describes a child using a scale from 1 *Not at all* to 5 *Very well*. The social competence scale comprises 20 items measuring teachers' perceptions of children's assertiveness, peer social skills, task orientation, and frustration tolerance. Examples of social

competence items include: "participates in class discussions," "completes work," and "well-liked by classmates." The Problem Behaviors scale comprises 18 items measuring teachers' perceptions of children's conduct, internalizing, and learning problems. Examples of problem behavior items include ''disruptive in class,'' ''anxious,'' and ''difficulty following directions.''

Data Analysis

Multiple imputation was conducted such that missing data was imputed for all missing cells (except social outcomes) using hot-deck sampling procedures available in R statistical software. Five datasets were then analyzed using Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 2002) features available in HLM 6.0 software with the results aggregated across the five imputed datasets for presentation. This analysis plan was used such that the nested nature of the data was respected, whereby multiple students were nested within the same pre-k classrooms. This was true for all outcomes, including the kindergarten ratings of social competence and problem behaviors because classroom predictors (mean levels and Variability of emotional, organizational, and instructional supports) were all based on preschool teachers. As a first step, unconditional models were analyzed that only accounts for the nesting of the data (children within classrooms), in order to estimate the amount of variance at the child and classroom levels. Then, models were run that included both child-level and classroom-level predictors. For the academic outcomes, children's fall score was entered as a predictor so that spring scores represent gains over the pre-k year.

Findings / Results:

Table 1 summarizes results across the five academic outcomes and two social outcomes. Of interest in the present study are the variability predictors, which are at the classroom level. A clear pattern emerged for variability in Emotional Support. For five of seven outcomes (three academic and both social outcomes), less variability in Emotional Support was related to better outcomes for children. Specifically, variability in Emotional Support was related to WJ Rhyming ($b = -.91$, $t = -2.05$, $p < .05$, $\delta = .05$), WJ Applied Problems ($b = -3.42$, $t = -2.23$, $p < .05$, $\delta = .06$), and Letter Naming ($b = -2.49$, $t = -2.29$, $p < .05$, $\delta = .06$). For the other two academic outcomes PPVT and OWLS, the coefficients were in the same direction, but failed to meet conventional levels of significance. Variability in Emotional Support was also associated with kindergarten Social Competence ($b = -.35$, $t = -2.70$, $p < .01$, $\delta = .09$) and kindergarten Problem Behaviors ($b = .20$, $t = 2.02$, $p < .05$, $\delta = -.07$). There was only one other significant relationship between variability in any CLASS domain and a child outcome: a main effect for variability in Instructional Support on Letter Naming ($b = 1.41$, $t = 2.16$, $p < .05$, $\delta = -.05$). This relationship indicated that more variability in Instructional Support was associated with improved performance on the letter-naming task in spring.

Conclusions:

In terms of measurement, the main finding of this study is that variability can be an important predictor of child outcomes. This supports the notion that although error is confounded with variability, and variability is inherently less reliable than a mean (Estabrook, Grimm, & Bowles, 2006), variability can still be used as a predictor if there are strong underlying associations.

Notably, viewing variability as a characteristic is at odds with classical measurement theory. However, given that the associations are both consistent with theory and form an empirical patter, we have come to the conclusion that variability in CLASS domains is itself at least in part a function of the teacher. This view of variability is more consistent with

generalizability theory which divides up variance into different sources. There are many potential sources of variance including both main effects such as the variance due to the teacher (shared variance across time) and interactions (teacher X time) whereby some teachers are particularly high or low during particular times of the day (Curby et al, in press). Our study suggests that the interaction term person X time may hold important predictive variance.

It is hard measure variability directly (i.e., have an observational protocol that measures variability), thus using statistical techniques that generate an estimate of variability is an easy, low-cost way to incorporate a measure of variability into a research study. The measure used in the present study, the CLASS (Pianta et al., 2008), is both widely used and recommends that all classrooms be observed at least four times during an observation day. Thus, any study that has used the CLASS to date could add variability as a predictor.

Ultimately, this study raises many more questions than it answers. For example, there are other ways to conceptualize variability (and consistency). Because of the scaling differences, when is it more appropriate to use variance as the metric instead of standard deviation units? Is it possible that interventions are affecting the variability in the construct of interest? The present study would suggest that interventions that promote consistent emotionally supportive interactions would have benefits for students even if there were no affect on the mean level of emotional support. What are predictors of variability? Do certain temperamental characteristics (e.g., adaptability) predispose children to do better in classrooms with less variability? Would decreasing the observation window (for example reducing it from a 20 minute observation to a 5 minute observation) give a better estimate of the variability, thereby producing a stronger statistical effect for it? Future research can begin to address these questions.

In terms of the classroom quality, this study highlights the practical implications of research that includes variability as a predictor. Findings from this study reveal that preschoolers in an emotionally variable classroom environment had worse academic and behavioral outcomes, relative to children in more emotionally consistent classrooms. As evidenced by the fact that the mean level of Emotional Support was a non-significant predictor but variability was, variability in Emotional Support may be more salient than mean levels when predicting academic and social outcomes. In past research, mean levels of emotional support were related to social outcomes—including studies that have used the same datasets (e.g., Mashburn et al., 2008). However, with variability in Emotional Support included as a predictor, there were no main effects for mean levels of Emotional Support. This finding is important because, at least when it comes to children's social development, the variability of Emotional Support may be more important than mean levels. In other words, it may be better to have a teacher with consistently mediocre emotionally supportive interactions than one who is more variable, but on average, offers higher levels of emotional supportive interactions. This finding does not negate the notion that teachers who engage in high quality Emotional Support represent the ideal scenario for children's learning. Mean levels of Emotional Support also play an important role in classroom processes; for example the level of teachers' Emotional Support has been found to be an important moderator, whereby at-risk children disproportionately benefited from an emotionally supportive teacher (Hamre & Pianta, 2005). The present study illustrates that only looking at mean levels of Emotional Support may ignore valuable information that is readily available when looking at variability around that mean.

Appendices

Appendix A. References

Burchinal, M., Vandergrift, N., Pianta, R., Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25,* 166-176.

Butler, A. C., Hokanson, J. E., Flynn, H. A. (1994). A comparison of self-esteem lability and low trait self-esteem as vulnerability factors for depression. *Journal of Personality and Social Psychology, 66,* 166-177.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales*. Circle Pines, MN: American Guidance Service.

Curby, T.W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B.K., & Pianta, R.C. (in press). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal.*

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test-revised. .* Circle Pines, MN: American Guidance Service.

Early, D., Barbarin, O., Bryant, D., Burchinal, M., Chang, F., Clifford, R., et al., (2005). *Pre-K in eleven states: NCEDL's Multi-State Study of Pre-K & Study of State-Wide Early Education Programs (SWEEP)*. Retrieved on June 30, 2010 from the FPG Child Development Institute: http://www.fpg.unc.edu/~ncedl/pdfs/SWEEP_MS_summary_final.pdf

Eizenman, D. R., Nesselroade, J. R., Featherman, D. L., & Rowe, J. W. (1997). Intraindividual variability in perceived control in an older sample: The MacArthur Successful Aging studies. *Psychology and Aging, 12,* 489-502.

Estabrook, R., Grimm, K.J. & Bowles, R.P. (2006, November). *Reliability of individual standard deviations.* Presented at the Gerontological Society of America Conference in Dallas, TX. Downloaded May 14, 2009 from http://people.virginia.edu/~cre2w/.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first grade classroom make a difference for children at risk of school failure? *Child Development, 76,* 949-967.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. Snow (Eds.), *School readiness and the transition to kindergarten* (pp. 49-84). Baltimore: Brookes.

Hamre, B.K., Pianta, R.C., Mashburn, A., & Downer, J. (2007). *Building and validating a theoretical model of classroom effects in over 4,000 early childhood and elementary classrooms*. Retrieved on June 30, 2010 from the Foundation for Child Development: http://www.fcd-us.org/resources/resources_show.htm?doc_id=507559

Hightower, A. D., Work, W. C., Cowen, E. L., Lotyczewski, B. S., Spinell, A. P., Guare, J. C., et al. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review, 15*, 393-409.

Kernis, M. H., Grannemann, B. D., & Mathis, L. C. (1991). Stability of self-esteem as a moderator of the relation between level of self-esteem and depression. *Journal of Personality and Social Psychology, 61,* 80-84.

La Paro, K. M., Pianta, R. C., Hamre, B., & Stuhlman, M. (2002). *Classroom Assessment*

Scoring System (CLASS) Pre-K Version. Charlottesville, VA: University of Virginia.

Mashburn, A.J., Pianta, R., Hamre, B.K., Downer, J.T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development, 79*, 732-749.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) Pre-K Version*. Baltimore: Brookes Publishing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III: Tests of Achievement*. Itasca. ILL: Riverside Publishing.

Appendix B. Tables and Figures

Table 1
Results of Hierarchical Linear Modeling analyses

	PPVT	OWLS	WJ RHYMING	WJ APPLIED PROBLEMS	LETTER NAMING	K COMPETENCE	K PROBLEM BEHAVIORS
Classroom Variance	69.15	40.78	3.54	36.29	21.35	.08	.03
Child Variance	135.60	128.10	12.60	129.13	67.61	.52	.34
Total Variance	204.75	168.88	16.14	165.42	88.96	.59	.37
ICC	.34	.24	.22	.22	.24	.13	.09
p	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Fixed Effects	*Coef.*	*Coef.*	*Coef.*	*Coef.*	*Coef.*	*Coef.*	*Coef.*
Intercept	100.20***	96.68***	4.38***	101.61***	14.54***	3.75***	1.38
Emotional Support Mean[a]	.48	0.14	0.04	0.28	0.26	-0.01	0 .02
Classroom Organization Mean[a]	-.22	-0.19	-0.13	0.18	-0.27	0.05	-0.04
Instructional Support Mean[a]	.88*	1.66***	0.35**	0.74*	-0.47	0.00	0 .02
Emotional Support Variability[a]	-1.06	-2.55	**-0.91***	**-3.42***	**-2.49***	**-0.35***	**0.20***
Classroom Organization Variability[a]	-0.46	0.95	0.21	1.67	-0.09	0.15	-0.06
Instructional Support Variability[a]	-0.27	-1.56	-0.20	-0.56	**1.41***	-0.07	-0.06
Gender (Male = 1)	-0.55	-1.50***	-0.30*	-1.34***	-1.23***	-0.31***	0.30***
Ethnicity: Hispanic (v. White)	-7.07***	-4.28***	-0.88***	-1.51*	-0.07	0.13**	-0.13**
Ethnicity: Black (v. White)	-4.55***	-1.64**	-0.53**	-3.15***	1.00*	0.01	0.00
Ethnicity: Multiracial/Other (v. White)	-2.60***	-2.24***	0.47*	-0.44	0.52	0.07	-0.05
Poor	-1.98***	-1.67***	-0.48**	-1.58**	-0.74*	-0.16***	0.14***
Maternal Education (years)[a]	0.62***	0.54***	0.21***	0.69***	0.35***	0.03***	-0.03***
Fall Score[a]	0.52***	0.56***	0.64***	0.47***	0.66***	*n.a.*	*n.a.*

[a] variable was centered for analysis

* p < .05, ** p < .01, *** p < .001