Decomposing the Variance of Child Outcomes in Multistage Sample of Head Start Children

Barbara Lepidus Carlson[1] and John Hall[2]
[1]Mathematica Policy Research, 955 Massachusetts Ave., Ste 801, Cambridge, MA 02139
[2]Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543-2393

Abstract

When designing a sample, estimates of expected precision are commonly made to help determine sample size. These calculations require specifying one or more of the following: type I error, power, population variance, design effects, finite population corrections, R^2 for covariates, and the effect size. Sometimes, an earlier study can provide some of these, but often one must rely on educated guesses. Even when other studies are available, it is not straightforward to derive the various components of variance. In this paper, we decompose the variance and design effects for several key child outcomes from two rounds of the Head Start Family and Child Experiences Survey (FACES), in the hope that they can be used to help design samples for similar multistage samples of preschool-age children. This clustered sample involves selecting Head Start programs, then centers, classrooms, and children. Working backwards from the observed total variance for these outcomes, we first factor out the design effect due to unequal weighting, and then decompose the design effect due to clustering.

Key Words: variance decomposition, design effects, Head Start

1. Motivation

Precision calculations are commonly made during the design phase of a study for key outcomes and population domains to help determine the necessary sample size to achieve a sufficiently narrow confidence interval around point estimates, or to be able to detect as statistically significant a meaningful difference between the means for two subgroups. The confidence interval calculation requires specifying the type I error rate (α), the population variance (σ^2), and the sample size. Conversely, one can specify the confidence interval, α, and σ^2 and solve the formula to determine the sample size. If the data result from a sample with a complex survey design, involving stratification, clustering, and/or weighting, the resulting design effects (defined below) should be estimated and incorporated into the variance term in the formula.[1] When calculating the minimum detectable difference (MDD) between two subgroups, one must additionally specify the desired power (π) to detect that difference as statistically significant. Note that the MDD is *not* the minimum *observed* difference that would be detected as statistically significant, but is the minimum *true underlying* difference that would result in the observed sample differences to be rejected as significantly different than zero most of the time (with

[1] Other terms could also be incorporated into the variance term, such as a finite population correction factor if appropriate and desired, a $1-R^2$ term if the estimate will be model-adjusted for covariates, or a reliability term. One can also subtract 2 times the covariance when the two groups being compared are not independent.

probability π). The following formula can be used to calculate the MDD between two subgroups:

$$mdd = \left(Z_{1-\alpha/2} + Z_{\pi}\right)\sqrt{\left(\frac{\sigma_1^2 \cdot deff_1}{n_1}\right) + \left(\frac{\sigma_2^2 \cdot deff_2}{n_2}\right)}$$

The Z values are critical values on the normal distribution.[2] Typical values of these critical values are $Z_{1-\alpha/2}$ = 1.96 for 2-sided α=.05, and Z_{π} = .84 for π=.80.

Specifying the α, π, sample sizes, confidence intervals, and minimum differences or effect sizes is straightforward. The σ^2 may be known or can be estimated from previous studies or if one is using a standardized or normalized measure, or can be calculated as p(1-p) for outcomes that are proportions or percentages. One can also specify the confidence interval or MDD in standard-deviation-sized units, setting σ^2 equal to 1 in these formulas. But incorporating the design effects into precision calculations at the design phase of a study often requires making educated guesses, especially for multi-stage clustered samples, which is often employed in surveys with in-person data collection for logistical and budgetary reasons.

Our colleagues (Peikes et al. 2011) produced a White Paper that included minimum detectable effect sizes and intraclass correlation coefficients for practice-based health studies. The motivation for this paper is to provide similar information about clustered samples of preschool children that we hope will prove useful for others designing similar studies.

2. Design Effects

A design effect is a way of quantifying the impact of complex sample design on the variance of an estimate, and is calculated as the variance of the estimate – properly accounting for the design complexities, which generally increase variances – divided by the variance one would have obtained with a simple random sample of the same size. A design effect of 1 means that the complex design had no impact on the variance, whereas a design effect of, say, 2 means that the design has doubled the variance of an estimate or has effectively halved the sample size. The design effect is often thought of as resulting from two separate factors: (1) unequal weighting and (2) clustering, and these two factors can be multiplied to obtain the overall design effect. Stratification can sometimes be employed in sampling to reduce the overall variance, but we assume for purposes of this paper that the design effect due to stratification is equal to 1.

Unequal weighting can be due to unequal sampling probabilities within stratum and/or due to weighting adjustments for nonresponse and poststratification. The design effect due to unequal weighting, DEFFw, is easy to calculate once the weights have been constructed. Two common ways to calculate DEFFw for weight w_i are:

$$DEFFw = \frac{n\sum_i w_i^2}{(\sum_i w_i)^2}$$

and

[2] For smaller sample sizes, the T distribution should be used instead.

$$DEFFw = 1 + \{cv(w_i)\}^2$$

where *cv* is the coefficient of variation.[3] However, estimating the DEFFw at the design stage is more difficult. If the DEFFw is due solely to differential sampling rates across strata, the DEFFw can be calculated if the sample and population sizes in each stratum are known. Other sources of differential sampling weights, such as probability proportional to size sampling and multiplicity adjustments, as well as weighting adjustments, will most likely have to be estimated based on experiences with other surveys with similar design characteristics.

Clustering effects occur in multi-stage samples; for example in a two-stage sample, where the first stage of selection (primary sampling unit, or PSU) might be a county and the second stage unit (SSU) a residence, or where the PSU is a school and the SSU is a student. Because the samples of residences or students are not independent – they are clustered within a sample of counties or schools, respectively – they are not providing as much information as a national random (single-stage) sample of residences or students would have. The design effect due to clustering, DEFFc, can be calculated as $1 + \rho(B-1)$, where B is the average cluster size and ρ is the intraclass correlation coefficient. The ρ is the proportion of the total variance that is due to between-PSU variation and $1- \rho$ is the proportion of the total variance that occurs within PSUs. If the SSUs (residences, students) are relatively heterogeneous within PSU (county, school), then ρ is close to 0 because virtually all of the variance occurs within PSU, leaving very little clustering effect. But if the SSUs are relatively homogeneous within PSU, ρ increases. Taken to the extreme, if all of the students within a school were identical with respect to a certain characteristic, then ρ would be equal to 1, the design effect would be equal to B, and the effective sample size would be the nominal sample size divided by B, which is equal to the number of PSUs. When there are more than two stages of sampling (for example, school district, school, classroom, and student), each stage of clustering introduces its own impact on the variance. If more than one stage of clustering will be accounted for when making estimates, the various components of the variance due to each stage of clustering should be incorporated into the variance formula, though it is quite common to specify only the PSU when estimating the variance of estimates from multi-stage (more than two-stage) samples. In fact, many software packages designed for survey data analysis only one to account for the first stage of clustering for with-replacement sample designs (or without-replacement designs with relatively small sampling fractions at the first stage).

3. The Data: Head Start FACES

3.1 The Study

In this paper, we will be using data from the Head Start Family and Child Experiences Survey, or FACES (West et al., 2011). FACES is a nationally representative sample of Head Start programs and the children and families they serve. It is a periodic, longitudinal descriptive study of program performance that provides information on classroom quality and outcomes for children across many developmental domains. FACES started in 1997, and has started new cohorts every three years, with the most last completed cohort being FACES 2009. (FACES 2014 is being launched in the fall of 2014.) In each cohort through 2009, children are followed from entry into Head Start

[3] Note that, when using the coefficient of variation generated by SAS PROC UNIVARIATE, you would first have to divide the *cv* by 100.

through the completion of the program, and then again in the spring of their kindergarten year. Mathematica has been conducting the FACES study for U.S. Administration for Children and Families starting with the 2006 cohort. The following figure shows the data collection schedule for FACES 2009. If children leave Head Start before completing the program, they are considered ineligible for followup and not considered part of the study's target population.

Age Cohort	Fall 2009	Spring 2010	Fall 2010	Spring 2011	Fall 2011	Spring 2012
3 year olds	Start of Head Start Year 1	End of Head Start Year 1		End of Head Start Year 2		End of Kindergarten
4 year olds	Start of Head Start Year 1	End of Head Start Year 1		End of Kindergarten		

Figure 1: Data Collection Schedule for FACES 2009

FACES has a multi-stage stratified sample design. The PSUs are Head Start programs, the SSUs are centers within programs, the third stage units (TSUs) are classrooms within centers, and the fourth (and last) stage units (FSUs) are the children within classroom. The first three stages are selected with probability proportional to size, with the number of children being the size measure. In FACES 2006 and 2009, we selected 60 programs, 2 centers per program, 3 classrooms per center, and 12 newly enrolled children per classroom.[4] After accounting for eligibility and parental consent among sampled children, the baseline sample size is about 3,500 children, which means the average cluster size per PSU at baseline is about 58 children. We will use the FACES 2006 and 2009 baseline data to generate values of ρ for each stage of clustering (program, center, and classroom).

3.2 The Measures and Domains

Because ρ is specific to each measure, and can differ by population domain, we selected a set of key outcome variables from FACES (Malone et al., 2013),[5] along with a set of population domains by which these outcomes are often presented.

- Peabody Picture Vocabulary Test 4 (Standard score) {*PPVT*}
- Woodcock-Johnson {WJ} Literacy (Letter Word, Standard score)
- Woodcock-Johnson {WJ} Literacy (Spelling, Standard score)
- Woodcock-Johnson {WJ} Math (Applied Problems, Standard score)
- Pencil-tapping percentage[6] (Executive Functioning)
- Teacher Child Report – Social Skills

[4] Children returning for their second year of Head Start were excluded.

[5] Note that no percentage or proportional variables are included among these key outcomes. Had they been included, the analysis of variance approach described below may not have worked as well.

[6] The pencil tapping assessment is used to examine children's inhibitory control, working memory, and attention.

- Teacher Child Report – Behavior Problems
- Teacher Child Report – Approaches to Learning
- Body mass index

Some key population domains for this study population include:

- Race/ethnicity
- Gender
- Household language
- Family risk factors

4. Methods Used to Generate ρ

As described above, to generate the values of ρ we used baseline data from FACES 2006 and FACES 2009, which had nearly identical sample designs. We looked at the 9 key outcome variables for both the full sample and for domains based on age, race/ethnicity, gender, family risk level, and household language. While many analysis weights are constructed for the FACES study, we opted to use PRA1WT, which requires that the child had a completed child assessment, a completed parent interview, and a completed teacher child report at baseline. We used SUDAAN software,[7] using a with-replacement design and the Taylor Series approach to variance estimation.

4.1 Method 1

Our first inclination was to deconstruct the DEFF generated from estimating means in SUDAAN into its two components (DEFFw and DEFFc) and then back out the value of ρ from DEFFc, given the average cluster size. We would then use a model-based variance component analysis approach to corroborate the findings. But the deconstruction approach proved more difficult than expected when trying to back out the values for *multiple* stages and *multiple* values of ρ. Before delving further into the difficulties, first we present the formula we used for the deconstruction method. It is an expansion for a four-stage sample of a formula for three-stage sample DEFFc in Skinner at al. (1989).[8]

$$DEFFc = 1 + \rho_1(a-1)bc + \rho_2(b-1)c + \rho_3(c-1)$$

where**:**

ρ_i = ρ at sampling stage i

a = number of sampled centers (SSUs) per program (PSU);

b = number of sampled classrooms (TSUs) per center (SSU);

[7] We also used SAS PROC SURVEYMEANS and got the same design effects.

[8] Formula 2.23 on page 39. Note that there are a number of formulas in the literature to estimate DEFFc for multi-stage samples.

c = number of sampled children (FSUs) per classroom (TSU)

In this design, ρ_1 is the proportion of the total variance that is accounted for by between-program variation; ρ_2 is the proportion of the total variance that is accounted for by within-program between-center variation; and ρ_3 is the proportion of the total variance that is accounted for by within-center between-classroom variation. That leaves $1 - \rho_1 - \rho_2 - \rho_3$ as the proportion of the total variance that is accounted for by within-classroom between-child variation.

If one disregards the "-1" terms in the multistage DEFFc formula, the coefficient for ρ_1 is equal to *abc*, or the total number of children per PSU; the coefficient for ρ_2 is equal to *bc*, or the total number of children per SSU, and the coefficient for ρ_3 is *c*, the total number of children per TSU.

In an attempt to generate ρ_1, ρ_2, and ρ_3 by deconstructing the overall DEFF of estimated means from SUDAAN, we did the following. First, we calculated DEFFw using the actual weights for the sample in each estimate. Next, to get what was needed to estimate ρ_1, we calculated the average cluster size b (children per program). We then calculated the overall DEFF specifying the actual PSU (programs) and sampling strata. We then divided each overall DEFF by DEFFw to get DEFFc. Then, using the simpler two-stage formula for DEFFc $(1 + \rho (b-1))$, we solved for ρ, and used this as our estimate of ρ_1, the between-program variance proportion. Note that this formula assumes no further clustering after the program level while, in reality, the DEFF generated by SUDAAN is based on data that has further clustering.

Then, to get what was needed to estimate ρ_3, we repeated the process, but tricked SUDAAN into treating the SSUs (centers) as strata and the TSUs (classrooms) as PSUs. We took this new overall DEFF, and again divided by DEFFw to get DEFFc. After factoring out the average cluster size per classroom from this DEFFc, we again solved for ρ, and used this as our estimate of ρ_3 - the within-center between-classroom variance proportion.

Having estimated ρ_1 and ρ_3, and we can then use the expanded formula to solve for ρ_2, the middle stage of clustering – the percent of total variance account for by within-program between-center variation.

4.2 Method 2

We had originally planned to use a model-based variance component analysis to corroborate what we found with Method 1. The problem would be finding a ready-to-use software procedure that provided variance components associated with each stage of sampling while properly accounting for the unequal weighting. The regression procedures for survey data in SUDAAN and SAS properly accounted for the weighting, but did not provide variance components. We then tried several standard regression or analysis of variance procedures in SAS (such as the GLM, GENMOD, GLMMOD, MIXED, VARCOMP, ANOVA, and NESTED procedures),[9] with the hope of specifying not only the clusters (programs, centers, and classes) but also the weight as an independent variable. However, none of these methods met our needs. Some procedures required balanced designs, which this study did not have. The procedures that provided variance components did not allow for continuous independent variables such as weights,

[9] We also tried using a weighted Hierarchical Linear Modeling framework and got somewhat similar, but not the same, variance components as obtained in the other methods.

while those that did allow for continuous variables did not provide variance components.[10,11] To move beyond these obstacles, we opted to run procedures without including weights, and chose PROC NESTED, which provided for each key outcome the proportion of total variance associated with each level of clustering. To compare with the DEFFc from Method 1, we used the Method 2 variance components resulting from the PROC NESTED as ρ's and plugged them into the expanded Skinner formula for DEFFc .

5. Results

Tables 1 through 4 on the following pages show the estimated values of ρ (expressed as a percentage) at the program, center, and classroom levels for each of the nine key outcome measures in fall 2006 and fall 2009.[12] In each table, we have indicated particularly large values of ρ (greater than or equal to 5 percent of the total variance) using red font.

Table 1 includes the full sample of children, and shows the estimated values using both calculation methods: (1) decomposing the design effect (top half), and (2) model-based variance component analysis (bottom half), as well as the estimated design effect due to clustering. There are several ways to compare values of ρ(program), ρ(center), ρ(classroom), and within-classroom variation in Table 1:

- Across the nine variables
- Between years (2006 vs. 2009)
- Between the two methods of estimating ρ

While many of the same patterns exist in the estimates resulting from the two methods, the numbers can be quite different. Even more troubling is the fact that some of the values of ρ in Method 1 are negative (highlighted in blue font). According to Kish (1965, p. 163), the value of ρ can be slightly negative (-1/[b-1], where b is the average cluster size), but this is rare. But many of the values for ρ(center) and ρ(class) from our Method 1 are large negative numbers.

If we focus on the second method, we see that some values of ρ are relatively stable between 2006 and 2009, but some differ quite a bit. As one would expect we see big differences in clustering effects at various stages across the nine variables. Some variables (such as the PPVT[13]) have high clustering effects at the program and center levels; others (such as the teacher-reported social skills and behavior problems) have high clustering effects as the classroom level; and still others (such as the three WJ measures) have little clustering effects anywhere, with nearly all of the variation occurring within classrooms.

[10] Most provided mean square errors, but did not provide the expected mean square coefficients needed to convert those to variance components. Calculating these coefficients proved difficult due to the unbalanced design, though we did figure out how to derive approximate coefficients based on average cluster sizes.

[11] We also tried running a series of models, introducing one clustering variable, then two, then three to discern the increase in R^2 from each additional variable. But this did not yield satisfactory results.

[12] Note that two of the nine measures – Pencil-Tapping and Learning Approaches – were not collected in 2006.

[13] The PPVT measures general aptitude for vocabulary, which may be more a product of shared experience in communities rather than classroom learning.

Tables 2 through 4 show estimated values of ρ by key subgroups (race/ethnicity, gender, and home language) using only the model-based variance component method. Other patterns emerge when looking at children by important domains. In Table 2, we see different patterns for White Non-Hispanic children, Black Non-Hispanic children, and Hispanic children. (Other racial/ethnic groups were excluded.) For example, Hispanic children appear to have much higher clustering effects at the program level for the PPVT (vocabulary) than do the other children, and they also have higher clustering effects at the center level for the spelling, applied problems, social skills, and behavior problems.

Girls appear to have a higher clustering effect at the center level than do boys for the PPVT vocabulary measure (Table 3), while boys appear to have a higher clustering effect at the classroom level for this measure. Boys have a higher clustering effect at the center level than do girls for the spelling measure.

Even starker differences emerge when we look at children by whether the primary language spoken to the child at home is English or not (Table 4). For nearly all the measures, children who speak a language other than English at home have much higher clustering effects at the center level than do children who speak English. The English-at-home children have minimal clustering effects for nearly all the variables, aside from the teacher-reported social skills, behavior problems, and learning approaches, which are highly clustered at the classroom level. We also looked at children separately by whether their family has no or only one family economic risk factor (single parenthood, low maternal education, household poverty) vs. two or all three of these. The pattern of clustering effects did not seem to differ much between children in these two types of family situations and so the table containing these values is not presented in this paper.

As discussed earlier in the paper, for Method 1, we calculated the overall DEFF using SUDAAN, specifying only the PSU (program) and stratum, then factored out the DEFFw to get an estimate of DEFFc. For Method 2, we calculated DEFFc using the values of ρ from the variance component analysis and the expanded formula for DEFFc from Skinner. In general, we found that the two methods generated different values, with some estimated DEFFc values from Method 1 being lower than those from Method 2, and others higher – though the *average* ratio of Method 2 to Method 1 DEFFc was 1.02. Each method has possible factors that could render an inaccurate estimate (with Method 1 not fully accounting for later stages of clustering beyond the PSU, and Method 2 not fully accounting for the design effect due to weighting). Because we do not know which one is more accurate in which scenarios, we do not present the DEFFc values in these tables.

Table 1: Design Effects and Values of ρ for the Full FACES Sample

Variable	FACES	ρ (program) x 100	ρ (center) x 100	ρ (class) x 100	Within-Classroom
Method 1. Decompose DEFFc					
Vocabulary (PPVT)	2006	9.70	13.79	0.01	76.50
	2009	9.79	10.39	7.83	71.98
Letter-Word (WJ)	2006	3.02	5.26	-2.35	94.06
	2009	0.64	1.69	-2.87	100.54
Spelling (WJ)	2006	9.17	14.08	-2.47	79.22
	2009	1.53	0.25	5.72	92.50
Appl. Problems (WJ)	2006	4.97	7.43	-0.88	88.48
	2009	1.68	2.22	-0.17	96.27
Pencil-Tapping	2009	0.55	-5.90	19.66	85.70
Social Skills	2006	1.27	-10.43	28.71	80.45
	2009	3.83	-2.97	26.13	73.02
Behavior Problems	2006	2.30	-4.93	19.26	83.37
	2009	4.52	0.33	18.26	76.89
Learning approaches	2009	6.04	-1.52	30.82	64.66
Body Mass Index	2006	1.46	3.61	-3.63	98.56
	2009	1.44	-0.31	7.19	91.68
Method 2. Model-Based Variance Components					
Vocabulary (PPVT)	2006	14.98	6.42	4.67	73.93
	2009	12.47	8.98	4.88	76.67
Letter-Word (WJ)	2006	0.61	2.91	1.66	94.82
	2009	1.63	3.09	0.00	95.28
Spelling (WJ)	2006	3.96	3.87	3.83	88.34
	2009	0.53	3.17	1.23	95.07
Appl. Problems (WJ)	2006	2.72	2.10	4.38	90.80
	2009	1.70	1.39	1.13	95.78
Pencil-Tapping	2009	0.00	3.86	1.63	94.51
Social Skills	2006	0.00	1.22	24.06	74.72
	2009	0.73	4.37	20.57	74.33
Behavior Problems	2006	0.00	5.22	17.63	77.15
	2009	2.94	4.95	13.92	78.19
Learning approaches	2009	1.03	5.01	22.48	71.48
Body Mass Index	2006	0.00	3.00	0.00	97.00
	2009	0.00	2.41	1.12	96.47

Table 2: Values of ρ by Race/Ethnicity (Method 2)

Variable	FACES	ρ (program) x 100	ρ (center) x 100	ρ (class) x 100	Within-Classroom
White Non-Hispanic Children					
Vocabulary (PPVT)	2006	1.11	9.57	0.00	89.32
	2009	4.64	0.69	6.35	88.32
Letter-Word (WJ)	2006	1.14	12.19	0.00	86.67
	2009	0.37	4.97	0.14	94.52
Spelling (WJ)	2006	1.67	0.16	8.48	89.69
	2009	3.32	0.00	10.06	86.62
Appl. Problems (WJ)	2006	0.00	7.24	0.00	92.76
	2009	1.65	0.00	5.63	92.72
Pencil-Tapping	2009	1.30	13.27	0.00	85.43
Social Skills	2006	0.79	0.00	27.11	72.10
	2009	3.52	0.05	19.80	76.63
Behavior Problems	2006	0.00	8.09	13.22	78.69
	2009	0.00	6.14	14.06	79.80
Learning approaches	2009	0.00	0.00	21.89	78.11
Body Mass Index	2006	0.00	2.57	10.09	87.34
	2009	2.32	0.00	11.25	86.43
Black Non-Hispanic Children					
Vocabulary (PPVT)	2006	5.04	4.69	0.00	90.27
	2009	2.21	0.43	9.74	87.62
Letter-Word (WJ)	2006	0.00	6.82	0.00	93.18
	2009	2.36	2.37	0.00	95.27
Spelling (WJ)	2006	10.11	0.82	2.68	86.39
	2009	1.60	1.33	0.00	97.07
Appl. Problems (WJ)	2006	7.51	1.57	3.76	87.16
	2009	3.62	0.00	1.14	95.24
Pencil-Tapping	2009	3.46	0.00	0.00	96.54
Social Skills	2006	1.79	3.64	24.89	69.68
	2009	0.00	1.28	27.37	71.35
Behavior Problems	2006	1.10	6.85	20.60	71.45
	2009	4.78	0.75	20.79	73.68
Learning approaches	2009	0.33	0.00	30.76	68.91
Body Mass Index	2006	0.00	6.10	0.00	93.90
	2009	0.00	4.95	1.46	93.59
Hispanic Children					
Vocabulary (PPVT)	2006	11.75	7.21	0.08	80.96
	2009	10.45	3.72	7.24	78.59
Letter-Word (WJ)	2006	0.05	0.68	4.70	94.57
	2009	3.36	3.97	0.00	92.67
Spelling (WJ)	2006	1.14	5.86	8.22	84.78
	2009	1.87	5.41	4.38	88.34
Appl. Problems (WJ)	2006	0.00	9.61	3.30	87.09
	2009	0.00	10.30	0.00	89.70
Pencil-Tapping	2009	1.49	0.00	8.80	89.71
Social Skills	2006	0.00	8.95	17.13	73.92
	2009	0.02	6.77	19.70	73.51

Variable	FACES	ρ (program) x 100	ρ (center) x 100	ρ (class) x 100	Within-Classroom
Behavior Problems	2006	0.00	5.93	17.60	76.47
	2009	4.06	5.77	13.06	77.11
Learning approaches	2009	2.19	7.64	22.13	68.04
Body Mass Index	2006	0.00	4.16	0.00	95.84
	2009	0.72	1.79	0.00	97.49

Table 3: Values of ρ by Gender (Method 2)

Variable	FACES	ρ (program) x 100	ρ (center) x 100	ρ (class) x 100	Within-Classroom
Female Children					
Vocabulary (PPVT)	2006	14.14	15.17	1.17	69.52
	2009	11.72	12.32	2.41	73.55
Letter-Word (WJ)	2006	0.00	8.51	0.00	91.49
	2009	4.11	0.00	0.00	95.89
Spelling (WJ)	2006	1.36	4.18	6.92	87.54
	2009	2.64	1.50	2.11	93.75
Applied Problems (WJ)	2006	0.00	4.27	3.02	92.71
	2009	2.37	0.00	4.65	92.98
Pencil-Tapping	2009	0.72	2.07	3.85	93.36
Social Skills	2006	0.00	3.20	25.64	71.16
	2009	0.00	5.61	22.97	71.42
Behavior Problems	2006	0.00	11.66	18.19	70.15
	2009	4.38	2.34	21.08	72.20
Learning approaches	2009	1.73	6.20	25.16	66.91
Body Mass Index	2006	0.00	2.76	0.00	97.24
	2009	0.00	1.36	6.83	91.81
Male Children					
Vocabulary (PPVT)	2006	13.45	2.14	6.54	77.87
	2009	13.83	5.83	5.23	75.11
Letter-Word (WJ)	2006	1.55	1.09	0.00	97.36
	2009	3.04	3.06	1.09	92.81
Spelling (WJ)	2006	4.64	7.91	0.81	86.64
	2009	0.06	5.40	1.86	92.68
Applied Problems (WJ)	2006	6.06	2.61	0.26	91.07
	2009	0.00	4.35	0.00	95.65
Pencil-Tapping	2009	0.00	3.10	0.00	96.90
Social Skills	2006	0.00	1.38	25.14	73.48
	2009	3.04	0.61	23.87	72.48
Behavior Problems	2006	0.30	3.56	19.65	76.49
	2009	2.86	4.40	12.96	79.78
Learning approaches	2009	2.41	0.31	25.17	72.11
Body Mass Index	2006	1.51	1.11	0.00	97.38
	2009	0.00	5.65	0.00	94.35

Table 4: Values of ρ by Primary Language Spoken to Child in Household (Method 2)

Variable	FACES	ρ (program) x 100	ρ (center) x 100	ρ (class) x 100	Within-Classroom
English					
Vocabulary (PPVT)	2006	7.37	2.75	3.12	86.76
	2009	4.45	3.45	7.34	84.76
Letter-Word (WJ)	2006	0.58	2.68	2.56	94.18
	2009	2.52	3.08	0.00	94.40
Spelling (WJ)	2006	4.21	4.23	2.09	89.47
	2009	0.68	3.00	0.84	95.48
Appl. Problems (WJ)	2006	2.40	2.55	4.65	90.40
	2009	2.75	1.40	0.43	95.42
Pencil-Tapping	2009	1.30	0.00	3.38	95.32
Social Skills	2006	0.00	0.48	20.67	78.85
	2009	0.44	2.20	22.20	75.16
Behavior Problems	2006	0.00	4.37	17.32	78.31
	2009	2.76	3.42	15.72	78.10
Learning approaches	2009	0.00	2.70	26.11	71.19
Body Mass Index	2006	0.36	0.64	4.00	95.00
	2009	0.42	0.79	3.19	95.60
Language Other Than English					
Vocabulary (PPVT)	2006	4.87	3.33	4.25	87.55
	2009	7.80	2.53	2.88	86.79
Letter-Word (WJ)	2006	7.02	3.83	2.23	86.92
	2009	0.69	8.35	5.92	85.04
Spelling (WJ)	2006	1.19	6.67	13.47	78.67
	2009	2.64	5.61	2.61	89.14
Appl. Problems (WJ)	2006	1.78	0.00	15.94	82.28
	2009	1.51	6.39	6.60	85.50
Pencil-Tapping	2009	2.16	6.98	0.00	90.86
Social Skills	2006	0.00	9.44	27.32	63.24
	2009	0.43	8.06	18.25	73.26
Behavior Problems	2006	0.00	7.92	21.85	70.23
	2009	0.93	10.31	8.22	80.54
Learning approaches	2009	2.29	8.40	19.23	70.08
Body Mass Index	2006	0.00	5.33	0.00	94.67
	2009	0.00	3.94	0.00	96.06

6. Conclusion

This paper started out being about the results – providing values of ρ that can be used by others designing multi-stage studies of young children in center or classroom settings. But in the process, the paper became as much about the methodology as the results. We thought we would be able to generate values of ρ relatively easily, having access to a fair amount of data and design parameters for a multi-stage study. Presumably this part of the paper – the "how" – is also useful for others.

We learned that one cannot completely disaggregate the design effect into its components (Method 1) when there is more than one stage of clustering. While it is easy enough to

factor out the DEFF due to weighting from the overall DEFF for the mean to get the DEFF due to clustering, that DEFF is based on specifying only the primary sampling units, but the clustering effects (between vs. within PSUs) are based on observations that have further clustering beyond the PSUs (namely, centers and classrooms). It becomes difficult to disentangle the center and classroom contributions to the variance. Our results using this methodology were suspect, when compared to those of Method 2, but most importantly because of the negative values for ρ(center) and even for ρ(classroom).

Method 2, which involved decomposing the variance using a model-based approach, was originally supposed to corroborate the results from Method 1, but ended up as our definitive method. While this approach did not allow for us to incorporate the impact of weighting on the total variance, it did allow us to specify all stages of clustering. Another down side to Method 2 would be the fact that analysis of variance techniques tend to break down with an unbalanced design (for example, when the number of children per class is not exactly equal across classes). Note that neither method accounted for the impact of stratification on the variance. And both methods are subject to the fact that they are based on sample data, whereas we are trying to estimate population variance components.

With respect to the results themselves, we see large differences across variables and domains, and even between 2006 and 2009 cohorts. When designing a study, then, it will require setting priorities and/or striking a balance across the different key measures and subgroups. In other words, if a key outcome variable has a high clustering effect at the program level, then sampling more programs would make sense (to the extent the budget allows); whereas a key outcome with a high clustering effect at the classroom level would mean sampling more classrooms within centers to minimize variance.

References

Kish L (1965). *Survey Sampling*. Wiley & Sons, New York

Malone L, Carlson BL, Aikens N, Moiduddin E, Klein AK, West J, Kelly A, Meagher C, Bloomenthal A, Hulsey L, and Rall K (2013). *Head Start Family and Child Experiences Survey: 2009 User's Manual*. Report submitted to the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. Washington, DC: Mathematica Policy Research.

Peikes D, Dale S, Lundquist E, Genevro J, Meyers D (2011). *Building the evidence base for the medical home: what sample and sample size do studies need?* White Paper (Prepared by Mathematica Policy Research under Contract No. HHSA290200900019I TO2). AHRQ Publication No. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality.

Skinner CJ, Holt D, and Smith TMF (1989). *Analysis of Complex Surveys*. Wiley & Sons, New York

West J, Tarullo L, Aikens N, Malone L, and Carlson BL (2011). *FACES 2009 Study Design*. OPRE report 2011-9. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.