JEOC REVIEW

June 19, 2018

Estimating the potential benefit of 3 to 5 star rated programs on preschool children.

Background

Compass Evaluation and Research submitted the "SUTQ Validation Study Results" for Ohio's Step Up To Quality (SUTQ) program to the Ohio Department of Jobs and Family Services in December 2016. Among the analyses in the document is a study as to whether the star rating system is predictive of Kindergarten Readiness Assessment (KRA) scores. Assuming random assignment of children to programs (there are no claims of random assignment), the study provides data that indicate the amount of additional learning that might be associated with three to five star ratings compared to one to two star ratings. There was no KRA performance data for those attending non-star rated pre-schools in the Compass study.

Research question

The validation study provides mean scores on the KRA for (1) one to two star rated programs and (2) three to five star rated programs as well as the standard deviations and the number of children (sample size) for the data. Further, the Compass study provides data for four KRA domains and the overall KRA test scores. The question of interest is whether there are significant differences between students participating in 1 or 2 star rated programs compared to those in programs rated at 3 to 5 stars and how large that difference might be.

Data being studied

Data were extracted from the report, Ohio's SUTQ Validation Study Results February 2017 (Heinemeier, S., A. D'Agostino, A. Hamilton, K. Kim, and M. Winglee. Durham: Compass) using the report's table 50 as reproduced in Table 1. The comparison of star rated programs will be based on the size of the effects the programs have on student performance measured by the comprehensive KRA test instrument.

Table 1. Data extracted from Table 50 in the validation study by Heinemeier, et al. (2017).

		2014-2015 KRA scores		2015-2016 KRA scores	
		1 -2 stars	3 - 5 stars	1 -2 stars	3 - 5 stars
	n-count	2,356	7,141	1,093	5,429
Domain					
Language and Literacy	**Mean**	264.09	266.31	263.95	266.81
	Stand. Dev.	10.7	11.6	11.4	11.9
Mathematics	**Mean**	264.84	266.70	262.87	265.68
	Stand. Dev.	12.3	13.0	12.3	12.7
Social Foundations	**Mean**	264.20	268.25	268.00	272.49
	Stand. Dev.	17.6	18.3	19.0	19.2
Physical Development	**Mean**	266.39	269.17	268.05	271.41
	Stand. Dev.	16.2	16.5	17.1	16.6
Overall Test Score	**Mean**	263.51	266.08	263.87	267.06
	Stand. Dev.	10.3	11.2	11.0	11.6

What is an effect size?

An effect size is an estimate of the magnitude of the effect of a treatment expressed, in this instance, as the number of standard deviations of superior performance of students in a 3 to 5 star rated program as compared to students in 1 to 2 star rated programs.[1]

A common way to interpret an effect size is in the context of a year's-worth of learning; how large an effect is associated with one year of schooling? Cohen (1988, p 24-27) describes an effect size of 0.2 as small, 0.5 as medium, and 0.8 as large but this classification also has critics that suggest the effect size needs to be interpreted in context. A formal study of effect sizes in grade transitions is offered by Hill, et al. (2007) and is reproduced as Table 2. [2]

Table 2. Average Annual Gain in Effect Size Measured with Nationally Administered Tests

From Hill, et al. (2007)

	Reading tests		Math tests	
Grade transition	**Mean**	**margin of error**	**Mean**	**margin of error**
Grade K-1	1.52	(+/-0.21)	1.14	(+/-0.22)
Grade 1-2	0.97	(+/-0.10)	1.03	(+/-0.11)
Grade 2-3	0.60	(+/-0.10)	0.89	(+/-0.12)
Grade 3-4	0.36	(+/-0.12)	0.52	(+/-0.11)
Grade 4-5	0.40	(+/-0.06)	0.56	(+/-0.08)
Grade 5-6	0.32	(+/-0.11)	0.41	(+/-0.06)
Grade 6-7	0.23	(+/-0.11)	0.30	(+/-0.05)
Grade 7-8	0.26	(+/-0.03)	0.32	(+/-0.03)
Grade 8-9	0.24	(+/-0.10)	0.22	(+/-0.08)
Grade 9-10	0.19	(+/-0.08)	0.25	(+/-0.05)
Grade 10-11	0.19	(+/-0.17)	0.14	(+/-0.12)
Grade 11-12	0.06	(+/-0.11)	0.01	(+/-0.11)

Hill, et al. (2007) caution their readers that interventions produce much smaller effect sizes than a grade transition. The difference between a 3 to 5 star schooling and a 1 to 2 star schooling is much like an intervention with students getting pre-school either way.

> Descriptive statistics … [a]veraged over the many different **interventions**, studies, and achievement outcomes encompassed in these … [analyses], **the mean effect sizes are in the 0.20 to 0.30 range**. Moreover, there is remarkably little variation in the means across grade levels, despite considerable variation in the interventions and outcomes represented for the different grades. (page 9, bold added)

[1] This manifestation of effect size is also commonly referred to as Cohen's d. See Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed.. Hillsdale, NJ: Erlbaum, pages 20-21.

[2] Hill, C. J., H. S. Bloom, A. R. Black, & M. W. Lipsey, 2007. Empirical Benchmarks for Interpreting Effect Sizes in Research. Washington: mrdc (MDRC Working Papers on Research Methodology). Instruments used in the study included: California Achievement Test - 5th edition (CAT5); Stanford Achievement Test Series - 9th edition (SAT9); TerraNova-Comprehensive Test of Basic Skills (CTBS); Gates-MacGinitie; Metropolitan Achievement Test (MAT8); TerraNova-California Achievement Tests (CAT); and Stanford Achievement Test Series: 10th Edition (SAT10). The Gates-MacGinitie does not include a mathematics component.

Results

Within each domain in the Compass report, the effect size was computed as well as the standard deviation of the effect size as shown in Table 3. For an effect size to be statistically significant, the z-score (effect size divided by the standard deviation of the effect size) should be greater than 2.

Table 3. Effect size and significance of effect sizes computed to compare 1-2 star programs to 3-5 star programs

	2014-2015 KRA scores			2015-2016 KRA scores		
	3 - 5 stars over 1 -2 stars			3 - 5 stars over 1 -2 stars		
Domain	**Effect Size**	**SD of ES**	**z-score**	**Effect Size**	**SD of ES**	**z-score**
Language and Literacy	0.195	0.024	8.2	0.242	0.033	7.3
Mathematics	0.145	0.024	6.1	0.297	0.033	8.9
Social Foundations	0.223	0.024	9.4	0.234	0.033	7.1
Physical Development	0.168	0.024	7.1	0.201	0.033	6.1
Overall Test Score	0.234	0.024	9.8	0.277	0.033	8.3

As shown in Table 3, the z-score for all ten effect size comparisons (five domains times two years of data) are highly significant lending support to the notion that 3 to 5 star programs might produce better KRA outcomes for children than 1 to 2 star rated programs. In all tested domains, students receiving the treatment at 3 to 5 star rated programs showed higher averaged KRA results than students in 1 to 2 star rated programs in each of the two years of data.

Using the expectation set by Hill, et al. (2007) that interventions produce effect sizes of 0.20 to 0.30, the data of Table 2 suggest that the difference between 1 to 2 star programs and 3 to 5 star programs are in this range although somewhat lower and more inconsistent in the 2014-2015 school year.

Analysis

Although the effect sizes associated with a 3 to 5 star program compared to a 1 to 2 star program are in the range of expectation based on reading Hill, et al. (2007), that evidence alone is not particularly informative in the context of a cost-benefit analysis. Ohio's KRA has cut scores that serve as indicators of kindergarten readiness: 263 for Language and Literacy, and 270 for the Overall Test Score. Assuming the data are normally distributed, computations suggest:

1. That while the average score (Table 1) in Language and Literacy is above the literacy cut for "on track" status, an additional 8.2% of all students might be above the cut with a 3 to 5 star program compared to a 1 to 2 star programs.
2. That while the average score (also Table 1) for the Overall Test Score is below the cut for kindergarten readiness, an additional 10.5% of all students might be above the cut with the 3 to 5 star program compared to the 1 to 2 star program.
3. The difference between the 1 to 2 star ratings and 3 to 5 star ratings are far less than one year of learning as shown in Table 2. Still, Table 4 shows the projected percentage of students that would be classed as kindergarten ready using the mean and standard deviations provided in Table 1.

Table 4. Percent of students estimated to be kindergarten ready from star rated programs, by year.

2014-2015		2015-2016	
1 - 2 stars	**3 – 5 stars**	**1 - 2 stars**	**3 – 5 stars**
26.4 %	36.3 %	28.9 %	40.0 %

For the projections of Table 4 to be meaningful, the KRA score data need to be normally distributed and the students entering the 1 to 2 star programs should be no different from those entering 3 to 5 star programs. Further, it is not known how star-rated program participants differ from other Ohio students of their age.

Caveats and notes

Perhaps the most important caveat – there is no analysis to determine whether the students in the 1 to 2 star program and the students in the 3 to 5 star program were academically, socially, and physically equivalent prior to their preschool experience. There is no control over the differences in the students entering the two programs and students were not randomly assigned to either a 1 to 2 star or a 3 to 5 star program. Nevertheless, if the pre-treated students were equals, for there to be an educationally meaningful benefit for a 3 to 5 star program over a 1 to 2 star rated program, there should be a positive difference in the effect sizes … and there is. Failure to detect such a difference would raise many questions such as whether the star ratings are relevant to the development of children. Further, it would have been better for this paper and for the validation study from which the data were drawn for the unit of analysis to have been aggregated to schools. Instead the report analyzes students as though students are independent learners while studying the dependence of their educational benefit on the star rating of the provider of that education.