

**Final Report:**

**The Effectiveness of Special Education Services for Children with Autism: A National Longitudinal Study**

**R40MC15595**

## I. Introduction

Children and youth on the autism spectrum frequently receive both health and education services, but schools remain the primary provider of services.(National Research Council (U.S.). Committee on Educational Interventions for Children with Autism., 2001). Regardless of the source of care, measuring the quality of those services remains a challenge. In both education and health care, quality is often measured with so-called process measures rather outcomes per se. Process measures generally involve some aspect of provider-client interaction believed to improve outcomes. For example, an indicator of quality of care for mental health services might involve followup outpatient visits after an inpatient admission.(Foster, 1999)

### A. Nature of the research problem

The link to outcomes is often supported by expert opinion rather than empirical evidence. For example, the American Academy of Pediatrics has produced treatment guidelines for the treatment of school-age children with attention deficit disorder.(Stein & Perrin, 2003) These guidelines indicate that “Primary care clinicians should establish a management program that recognizes ADHD as a chronic condition”. Of course, the plan is really a proxy for effective treatment. If the children receive the wrong doses of a psychostimulant, then of course, outcomes will not be improved.

In much the same way, special education uses process measures to measure and ensure treatment quality. By law, children in special education must have an “Individualized Education Plan” that sets goals for each child and a treatment plan to attain those goals. Another requirement under the special education law (the Individuals with Disabilities Education Act) is that children be educated in the least restrictive setting (i.e., a regular classroom). A series of lawsuits have identified the criteria that schools should use in picking placements for youth. These include the potential benefits of interaction with non-disabled peers.

Children and youth with autism pose special challenges for assessing the quality of schooling. Foremost among these is the heterogeneity of youth falling under the broad umbrella of “autism spectrum disorders”. This variability highlights the enormous potential for confounding. If individuals treated were completely homogenous, one source of potential confounding would be eliminated.

### B. Purpose, scope, and methods of the investigation

This project employed the latest tools of causal inference and a large dataset to determine whether a child’s participation in regular classroom setting is related to key outcomes. We focus on youth in special education and important outcomes like school completion and scores on a parent-reported measure of social and cognitive achievement. We adjust comparisons of youth who experienced different levels of inclusivity using a broad range of covariates.

## II Study Design and Methods

### A. Data

The National Longitudinal Transition Study-2 (NLTS2) is a ten-year study of youth with disabilities who were receiving special education services in public or state-supported special schools. The NLTS2 uses a nationally representative sample of youth with disabilities who were between the ages of 13 and 16 on December 1, 2000. 434 of these individuals had a primary diagnosis of autism and were included in these analyses.

The study collected data biannually in five waves from 2001 to 2009. The present study uses the Wave 2 data, collected in 2003, for characteristics of the school program and home environment and Wave 4 data, collected in 2007, for the outcome measures.

### B. Instruments used

This study uses data collected using four instruments: a parent telephone interview, a school program questionnaire, a direct assessment of the youth's academic abilities and a functional rating for youth whose disability precluded them from completing the direct assessment. The school program questionnaire was a self-administered survey that was completed by the staff member at each youth's school who was most knowledgeable about the special education program. The direct assessment used research editions of the language arts, math, science and social studies subtests of the Woodcock-Johnson III (WJ III). The functional rating was assessed using the Scales of Independent Behavior-Revised (SIB-R), which measures motor skills, social interaction, as well as personal and community living skills.

#### *Exposure: Inclusivity*

The primary exposure of interest in this analysis was the proportion of time the youth spent in a general education classroom. The school program questionnaire collected data on the courses that each student took during the 2003 school year and whether each course was taken in a general education or special education classroom. The number of courses taken in a general education classroom was divided by the total number of courses taken to calculate the proportion. The proportion of time spent in an inclusive setting was then categorized as 0%, 1-74% or 75-100% of courses taken in a general education classroom.

#### *Outcomes*

Three outcomes were assessed in this analysis using Wave 4 data: not dropping out of high school, any college attendance, and the cognitive functional scale at Wave 4. Not dropping out of high school was chosen instead of high school graduation since not all youth would be expected to graduate from high school by Wave 4.<sup>1</sup> Youth were coded as not dropping out if the parent reported that they graduated, received a certificate or GED or were still in high school at the time of Wave 4 data collection. Any college attendance was based on parent report of whether the youth attended any type of post-secondary school in the previous two years, including postsecondary classes to earn a high school degree, a two-year or four-year college or postsecondary vocational school.

---

<sup>1</sup> Wave 5 data are not yet available.

The functional cognitive scale at Wave 4 was the same as the scale used for Wave 2.

### C. Statistical Design and Methods

In an observational study, the link between an exposure of interest and an outcome represents an association. Moving from that association to drawing causal inference depends on a key assumption. Researchers commonly assume “ignorability” or the absence of unobserved confounding. Under this assumption, an individual at level of association represents what one would observe for other children were they to experience that exposure. This assumption involves unobserved differences over and above any covariates used to adjust comparisons of individuals at different levels of exposure.

In the case at hand, suppose all children spent all of their time in regular classrooms or none. Ignorability means that children who are always in regular classrooms can be used to represent the experiences of other children (those never in regular classrooms) had they spent all of their time in regular classrooms. Ignorability would mean that comparisons of children matched on observed characteristics at different levels of inclusivity would reveal the effect of inclusivity (and not just the association). Ignorability essentially assumes that inclusivity or other exposure is as if randomly assigned among sub-groups of participants sharing the same set of observed characteristics.

Is this assumption plausible? It is impossible to fully test this assumption empirically, but at least a necessary condition for plausibility is that one select the correct covariates and omits incorrect ones. “Correct” in this sense means potential confounders—variables that influence both inclusivity and the outcomes of interest. Other covariates need to be omitted. (These include “colliders”. These are variables potentially on the causal path from the exposure to the outcome that are themselves subject to unobserved confounding. Such variables will induce a spurious relationship between the exposure and the unobserved determinants of the colliders.)

Analyses grounded in ignorability generally involve comparisons of outcomes across levels of exposure adjusted for the covariates selected. For example, an analyst might regress the outcome on exposure and the covariates selected. A second condition for causal inference in this case—even if ignorability is correct—is that the mechanics of regression (or other method) work correctly. By this we mean that the adjustment mechanism fully “balance” the distribution of the covariates across levels of exposure. In a regression context, achieving balance involves specifying the functional form of the regression model correctly.

An alternative methodology for adjusting comparisons across levels of exposure for covariates involves propensity scores. These are the predicted probability of exposure, and they represent a convenient summary of the covariates. The propensity score can be used to calculate adjusted between-group means in a variety of ways, such as matching. Propensity score-based methods assume ignorability but have advantages over regression, such as producing estimates of the effect of the exposure with a clear interpretation and checking covariate balance.

Propensity scores can be used in analyses that take various forms. We use inverse probability of treatment weights. Unlike matching, for example, this methodology easily generalizes beyond two levels of exposure. The weights are calculated as one

over the probability of the exposure actually received.(Foster, 2003; Hirano & Imbens, 2004; Imbens, 2000) These weights can be incorporated in the analyses like survey weights—they represent pseudo-populations where the covariates and exposure are no longer related.

As discussed above, we model inclusivity as an ordered category. In order to generate predicted levels of inclusivity, we used a multinomial logit model.(Greene, 2008)<sup>2</sup>

### *Handling of missing data*

Table 1 demonstrates that the data suffer from fairly extensive missing data. In a multi-variate analysis that that considered here, limiting the analysis to the complete-cases would dramatically reduce the overall sample size. With that in mind, our analyses involved multiply imputed data; the data were imputed under the missing at random assumption. This assumption means that individuals who lack data can be represented by the experiences of those with the same value of the covariates who actually provided data.(Little & Rubin, 1987)

In analyzing missing data, one conducts separate analyses of the imputations (in our case 5) and combines the estimates using Rubin's rules.(Little & Rubin, 1987) the standard errors of the resulting estimates reflect the uncertainty in each imputation-specific estimate as well as variation across imputations in the estimates. The latter captures the uncertainty stemming from the fact that the data are missing.

### *Choice of covariates*

The parent interview provided data on explanatory variables including the severity of the youth's disability, the level of family support for education and demographic information. Four measures of severity were included: the number of domains affected by the disability, a functional cognitive scale, a social skills scale and whether the youth was able to be evaluated using the direct assessment. The number of domains affected ranged from 0-7 and included vision, hearing, expressive communication, receptive language, bidirectional communication, use of arms, hands, legs and feet, and general health. The functional cognitive scale ranged from 4-16 and measured the youth's ability to perform daily activities such as reading signs, telling time or using the telephone. The social skills scale ranged from 0-18 and measured the youth's ability to interact with family and friends. A scale ranging from 0-12 was used to measure family support for education and included the frequency with which the parent attended school meetings, school or class events or volunteered at the school. Demographic characteristics included the youth's race and the parent's level of education.

---

<sup>2</sup> We might have employed an ordered logit. Such a model would have been more parsimonious; it involves one regression coefficient for each covariate (half as many as the multinomial logit).

## IV. Detailed Findings

### Descriptive Statistics

Table 1 describes the sample. The table reports the covariates (section A); the level of inclusivity experienced (section B); and outcomes of interest (section C). One can see that the vast majority of Autistic youth are male. Roughly six in ten participated in direct assessments as part of the study. The average child had his or her condition identified early (age 2).

The table also provides information on the exposure, inclusivity. One can see that nearly half (45%) spent no time in the regular classroom. 17% spent three-quarters or more of their school day in regular classroom settings. Presumably, this variation reflects the child's characteristics as well as "supply side" factors, such as the range of special education services offered in the school where the child attended school.

### Generating propensity scores

Table 2 presents the results of our multinomial logit. One can see that there are two coefficient estimates for each covariate. These represent the log-odds of that choice relative to the reference category, 0% inclusivity. In general, the covariates generally do not predict inclusivity. The handful that correspond to significant coefficients could reflect chance findings given the large number of coefficient estimates.

In typical analyses involving propensity scores, the next, key step would be to check covariate balance. Given the weak relationship between the covariates and exposure (inclusivity), that step is unnecessary. For better or worse, there is no confounding relationship between the exposure and the covariates to be removed.

### Outcome analyses

Table 3 presents the raw and adjusted levels of the three outcomes across the levels of the inclusivity variable. Looking across the three outcomes, one can see that the effect of adjusting for the covariates narrows the variation across levels of exposure. For not dropping out of high school, those spending no time in inclusive settings are least likely to continue in high school—the gap between those youth and those spending all of their time in inclusive settings is 17 percentage points. Adjusting for the covariates narrows this gap to 6 points.

For the second outcome, college attendance, the gap between the highest and lowest categories is enormous—57 percentage points. Those with moderate inclusivity fall in between as one would expect. Adjusting for the covariates closes this gap to 14 points. This estimate is sizable but the null hypothesis of no effect cannot be rejected because of the high imprecision associated with the estimate.

For the third outcome, the score on the functional cognitive outcome, the between-group difference is largely unchanged by adjusting for the covariates (actually rising from 5.58 to .03). These effects are tiny.

## V. Discussion

In general, our analyses suggest that inclusivity does not improve child outcomes. This conclusion depends on the plausibility of the ignorability assumption and by extension

the covariates employed. As discussed above, these covariates were selected by experts for use in the NLTS and generally have good psychometric properties. Still, one interpretation of these findings is as an assessment of criterion validity. If these measures cannot predict features of special education involvement (like inclusivity), then perhaps these measures are not as strong as the psychometrics suggest.

Of course, another possibility is that placement in special education is essentially randomly assigned. This possibility is alarming. Perhaps youth find their way into placement based on school and district characteristics like funding unrelated to their own needs and goals. We know from other data that schools, districts and states differ enormously in their funding and eligibility requirements for special education. We have not included these variables in our analyses, and they may be potential confounders. It is not obvious, however, what direction these characteristics would bias our calculations. One possibility is that better funded settings are more creative and have more tailored placements, mixing special services with inclusive settings. In that case, these analyses would exaggerate the benefits of inclusivity.

Given that inclusivity *should* reflect youth characteristics, observational analyses of this type are challenging. In fact, we are assuming in essence that the schools do not meet their obligation to these youth. If they did, confounding by both observed and unobserved variables would be so severe that the analyses would be impossible. One alternative would be to replace the ignorability assumption with another methodology, such as instrumental variables estimation. Such analyses would involve finding a variable that influences placement but that is not a direct determinant of outcomes. This area is one for future research.

Another possibility is simply that inclusivity is a poor indicator of quality of treatment and services. In that case, researchers need to continue to work to identify process measures of quality of care that indeed predict outcomes. Schools might also focus on standardized testing, but requiring children and youth with autism to take tests like other children may be of limited value. Appropriate outcome assessment would reflect the challenges these children and youth face. This task is no doubt complicated by the heterogeneity of children and youth labeled as “autistic”. A good measure of quality would predict treatment outcomes for different children and youth and would be easy enough to measure that care could be monitored by state and federal education authorities. Such measures could be written into IEPs as a planning tool.

An alternative to observational studies like this one would be to limit research to studies where participation could be randomly assigned. Few studies in the autism literature employ randomization (Stein & Perrin, 2003), and it is far from clear that randomization can solve the problem of measuring quality of care. Schools could hardly randomly assign youth to different levels of inclusivity—they obviously want (and are legally required) to act in the child’s best interest. Efficacy studies could document links between outcomes and interventions to improve putative measures of quality of care. For example the Multimodal Treatment for Attention Deficit Disorder study included a medication management arm that included high-quality care. The improvement of children in that arm would suggest that effective management and careful titration of medication improves quality of care (including outcomes). However, such “proof” shifts the

focus to an equally formidable challenge—translating efficacious treatment into care in the real-world.

It seems clear that development of quality indicators for special education in general and autism in particular will have to rely on observational methods like those employed here. Observational studies would appear to offer the only real hope of providing samples large enough to differentiate among children and youth with autism. Progress in developing such measures may be facilitated by improved administrative databases. Such studies offer the potential sample sizes required for these types of analyses as well as the possibility of examining how well the quality measures work across a range of communities and schools.

## **VI. List of products**

Foster, EM. Pearson, E. (Under review) Is Inclusivity an Indicator of Quality of Care for Autistic Children in Special Education? Pediatrics.

\

**Foster, EM, Pearson, E. Does transition planning improve outcomes for individuals exiting Special Education?"**



## References

- Foster, E. M. (1999). Do aftercare services reduce inpatient psychiatric readmissions. *Health Services Research, 34*(3), 715-736.
- Foster, E. M. (2003). Propensity score matching: an illustrative analysis of dose response. *Medical Care, 41*(10), 1183-1192.
- Greene, W. H. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, N.J.: Prentice Hall.
- Hirano, K., & Imbens, G. W. (2004). The Propensity Score with Continuous Treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* (pp. 73-84).
- Imbens, G. W. (2000). The Role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706-710.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*: John Wiley & Sons.
- National Research Council (U.S.). Committee on Educational Interventions for Children with Autism. (2001). *Educating children with autism*. Washington, DC: National Academy Press.
- Stein, M. T., & Perrin, J. M. (2003). Diagnosis and treatment of ADHD in school-age children in primary care settings: a synopsis of the AAP Practice Guidelines. American Academy of Pediatrics. *Pediatr Rev, 24*(3), 92-98.

<b>Table 1. Descriptive Statistics</b>				
		<b>N</b>	<b>Mean</b>	<b>SD<sup>1</sup></b>
<b>Covariates</b>				
<b>Gender</b>				
	Male	328	0.83	
	Female	67	0.17	
<b>Race</b>				
	White	250	0.63	
	African-American	73	0.19	
	Other	71	0.18	
<b>Family Income</b>				
	<\$25,000	78	0.22	
	\$25,000-\$50,000	84	0.24	
	>\$50,000	188	0.54	
<b>Parent Education</b>				
	High school graduate or less	113	0.29	
	Some college/Associates degree	125	0.32	
	Bachelors degree or higher	154	0.39	
<b>Received disciplinary action during the current school year</b>				
		69	0.15	
<b>Grades</b>				
	Mostly A's and B's	131	0.59	
	Mostly B's and C's	67	0.30	
	Mostly C's and below	23	0.10	
<b>Number of times youth has changed schools</b>		375	0.26	0.50
<b>Number of days youth was absent in a month</b>		423	1.12	2.41
<b>Functional cognitive skills (Range 4-16)</b>		364	10.54	3.97
<b>Number of domains influenced by disability (Range 0-7)</b>		395	2.82	1.10
<b>Social skill scale (Range 0-18)</b>		392	8.73	3.53
<b>Means of Assessment</b>				
	Alternate	141	0.39	
	Direct	220	0.61	

<b>Table 1. Descriptive Statistics (continued)</b>			
Rating of persistence at tasks			
Never	66	0.17	
Sometimes	180	0.47	
Very often	140	0.36	
Age disability diagnosed	382	2.02	1.81
Family support for education at home	358	6.21	2.04
Family support for education at school	383	1.62	0.96
<b>Exposure: Inclusivity</b>			
0%	215	0.45	
1-74%	185	0.38	
75-100%	82	0.17	
<b>Outcomes</b>			
Graduated received certificate/GED or still in high school	364	0.86	
Attended some college since leaving high school	261	0.40	
Functional cognitive skills (Range 4-16) - Wave 4	321	10.39	4.13
Notes:			
1) Standard deviations (SD) reported for continuous variables only.			

<b>Table 2. Multinomial Logit Model Predicting Level of Inclusivity</b>						
	1%-75%			>75%		
	Beta	SE	P-value	Beta	SE	P-value
Gender (Male=1; Female=0)	-0.12	0.42	0.77	-0.25	0.60	0.67
Race (white: reference category)						
African-American	0.46	0.38	0.22	0.57	0.91	0.53
Other	-0.37	0.35	0.28	-0.25	0.59	0.68
Family income (<\$25,000: reference category)						
\$25,000-\$50,000	0.14	0.43	0.75	0.53	0.66	0.42
>\$50,000	0.30	0.41	0.47	0.63	0.71	0.38
Parental Education (High school graduate or less: reference category)						
Some college/Associates degree	-0.48	0.38	0.20	-0.35	0.76	0.64
Bachelors degree or higher	-0.36	0.38	0.35	0.27	0.58	0.64
Received disciplinary action during the current school year	0.64	0.38	0.09	0.66	0.47	0.16
Grades						
Mostly B's and C's	-0.25	0.29	0.39	0.09	0.45	0.84
Mostly C's and below	0.00	0.50	1.00	-0.51	0.65	0.44
Number of times youth has changed schools	0.29	0.28	0.30	-0.19	0.45	0.67
Number of days youth was absent in a month	-0.06	0.07	0.33	0.06	0.08	0.42
Functional cognitive skills (Range 4-16)	-0.06	0.22	0.80	-0.20	0.41	0.62
Number of domains influenced by disability	-0.02	0.12	0.84	-0.08	0.15	0.58
Social skill scale (Range 0-18)	0.04	0.04	0.41	0.03	0.06	0.63
Direct Assessment	1.13	0.30	0.00	1.52	0.71	0.03
Rating of persistence at tasks	0.02	0.22	0.93	0.02	0.27	0.95
Age disability diagnosed	0.03	0.08	0.74	-0.01	0.11	0.92
Family support for education at home	0.26	0.07	0.00	0.25	0.13	0.06
Family support for education at school	-0.08	0.15	0.61	0.10	0.25	0.69
Intercept	-2.35	1.20	0.05	-5.50	2.23	0.01
Note: 0% is the reference category						

<b>Table 3. Impact of inclusivity on Key Outcomes</b>							
Percentage of time spent in an inclusive setting		Not Dropping Out		Some College		Functional Cognitive	
		Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
0% (Reference category)							
1-74%	Effect	0.06	0.01	0.23	0.00	2.59	-0.33
	SE	-0.04	-0.04	-0.07	-0.10	-0.39	-1.00
	p-value	0.08	0.88	<.01	0.98	<.01	0.75
75-100%	Effect	0.17	0.08	0.57	0.14	5.58	0.03
	SE	-0.07	-0.07	-0.08	-0.16	-0.51	-1.95
	p-value	0.01	0.24	<.01	0.40	<.01	0.99
p-value for joint significance		0.05	0.68	<.01	0.67	<.01	0.95
Obs		434.00		434.00		434.00	
Note: the "effects" for the two dichotomous outcomes are the so-called marginal effects, that is the adjusted mean difference in the predicted probability of the outcome. For the third, the effect is the type regression coefficient.							